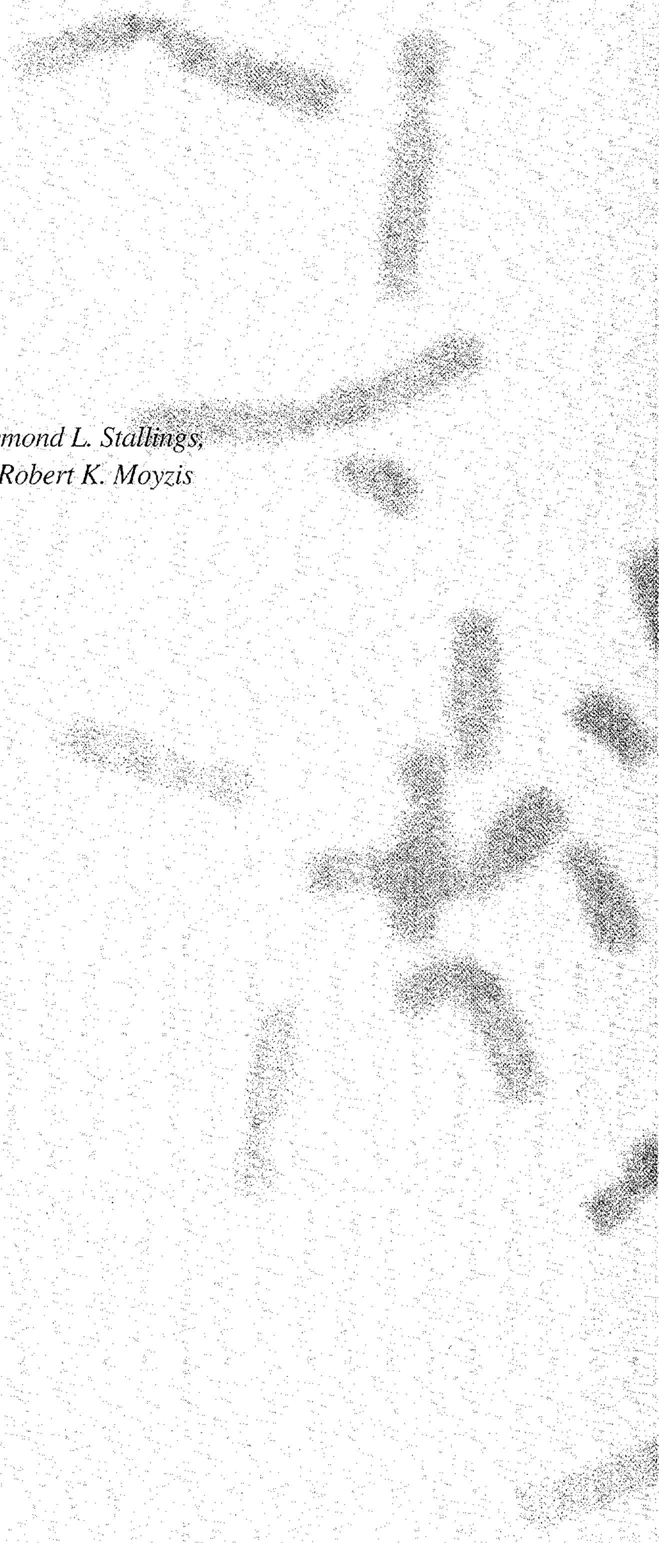
The Mapping of Chromosome 16

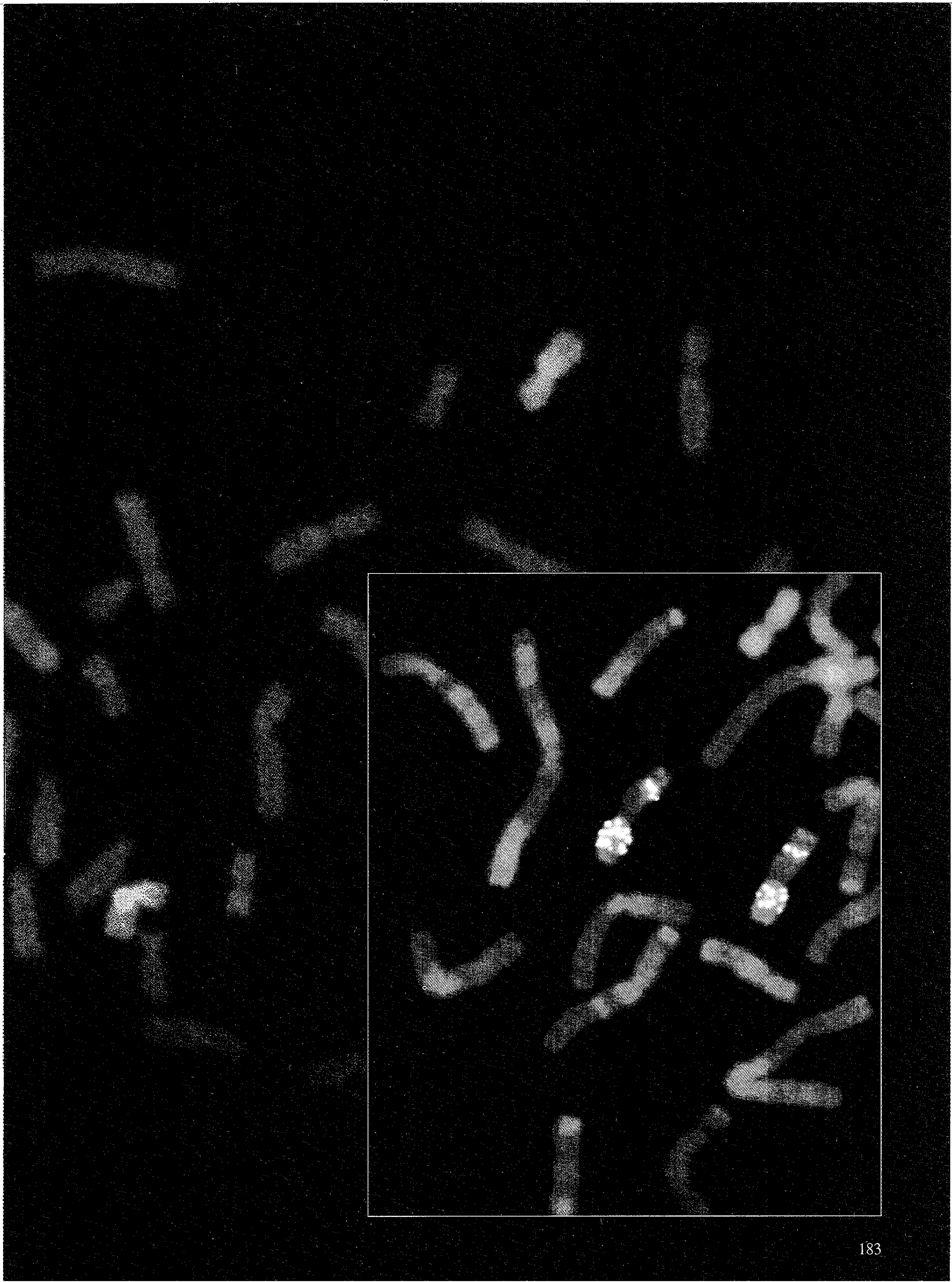
Norman A. Doggett, Raymond L. Stallings, Carl E. Hildebrand, and Robert K. Moyzis



Human chromosome 16 is the main focus of the mapping efforts at Los Alamos. The large photomicrograph on these opening pages illustrates the starting point for those mapping efforts, the evaluation of our chromosome-16-specific library of cloned fragments. Among the 23 pairs of human chromosomes, one pair, chromosome 16, is identified by fluorescence in-situ hybridization. Thousands of yellow fluorescent probes derived from the clone library have hybridized to both copies of chromosome 16. The high density and uniform coverage of the fluorescent signals were a strong indication that we could use the library to construct a map of overlapping cloned fragments spanning the entire length of the chromosome.

The image inset into the photomicrograph illustrates another aspect of our mapping project: the discovery of a new class of repetitive sequences that are specific to chromosome 16. Probes for one of those repetitive sequences are shown hybridized to several regions on both arms of chromosome 16. This sequence, now thought to facilitate the chromosomal rearrangements associated with a type of acute nonlymphocytic leukemia, is being used to help isolate the genes that are disrupted by the rearrangements. Details are presented in "What's Different about Chromosome 16?"

Here we present the story of our efforts to construct a physical map for an entire human chromosome, the progress in integrating that map with the corresponding genetic-linkage map, and the current applications of the map to the isolation of disease genes.



Setting the Stage

Both the molecular and the physical technology for constructing physical maps of complex genomes have developed at a blistering pace over the past five years, due largely to the initiation of the Human Genome Project. These technologies include the cloning of very large DNA fragments, electrophoretic separation of million-base-sized DNA fragments, and sequence-based mapping using the polymerase chain reaction (PCR) to identify unique sequences along the genome. The latter provides a language for interrelating various types of genome maps. The significance of these developments is discussed in Part II of "Mapping the Genome."

In 1988, when our laboratory initiated the physical mapping of chromosome 16, the cloning of very large DNA fragments in yeast artificial chromosomes (YACs) was just beginning in a handful of laboratories and only one library of YAC clones containing all the DNA in the human genome had been constructed worldwide. The total human-genomic YAC library was constructed at Washington University, where the technique of YAC cloning had originally been developed. The polymerase chain reaction had not yet become a standard tool of molecular biology, and the use of sequence-tagged sites (STSs) as unique DNA landmarks for physical mapping had not yet been conceived (see "The Polymerase Chain Reaction and Sequence-tagged Sites" in "Mapping the Genome"). Thus, in 1988 the most modern tools for large-scale physical mapping of human chromosomes were still waiting in the wings. On the other hand, a number of mapping techniques had been developed and were being applied to the genomes of some of the favorite organisms of molecular biologists.

Cassandra Smith and Charles Cantor had used pulsed-field gel electrophoresis to order the very large restriction fragments produced by cutting the *E. coli* genome with two rare-cutting restriction enzymes. The resulting long-range restriction map of *E. coli* demonstrated that pulsed-field gel electrophoresis is a way to study the long-range order of landmarks on the DNA of human chromosomes. Contig maps, or physical maps of ordered, overlapping cloned fragments, were near completion for the genomes of *E. coli* (about 5 million base pairs) and the yeast *S. cerevisiae* (about 13 million base pairs). Those maps were constructed using lambda-phage clones, which carry an average DNA insert size of 20,000 base pairs. Work had also begun on mapping the genome of the nematode (100 million base pairs) using cosmid clones. Cosmids carry the much longer average insert size of 35,000 base pairs.

The haploid human genome, which includes one copy of each human chromosome, has 3 billion base pairs and is therefore about 250 times the size of the yeast genome and 30 times the size of the nematode genome. When plans for the Human Genome Project were being discussed in the late 1980s, it was natural to consider dividing the human genome by chromosome and mapping one chromosome at a time.

Ongoing work at Los Alamos on human DNA and on adapting flow-sorting technology to separating individual human chromosomes set the stage for the Laboratory to play a key role in the Human Genome Project. In particular, as part of the National Gene Library Project, a group led by Larry Deaven had constructed twenty-four libraries, or unordered collections of lambda-phage clones, each containing DNA from one of the twenty-four human chromosomes (see "Libraries from Flow-sorted Chromosomes"). Those chromosome-specific libraries were designed as a source of probes to find polymorphic DNA markers for constructing genetic-linkage maps (see "Modern Linkage Mapping") and as a source of clones for rapid isolation of genes using cDNAs, or coding-region probes, to pick out the appropriate clones from the libraries. Deaven and his group were also constructing larger-insert chromosome-specific libraries using cosmid vectors. The large DNA inserts were prepared by partially digesting sorted chromosomes with restriction enzymes, thereby creating overlapping fragments. The cloned fragments would therefore be useful in constructing physical maps of ordered, overlapping clones covering extended regions of human chromosomes. Among the first chromosome-specific cosmid libraries to be constructed at Los Alamos was one for human chromosome 16.

Human chromosomes range in size from 50 million base pairs for chromosome 21 to 263 million base pairs for chromosome 1. Chromosome 16, which is about 100 million base pairs in length, was chosen as our primary target for large-scale physical mapping. We selected chromosome 16 for a number of technical reasons including: (1) the availability of a hybrid-cell line containing a single copy of chromosome 16 in a mouse-chromosome background, which permitted accurate sorting of human chromosome 16 from the mouse chromosomes and thus the construction of a high-purity chromosome 16-specific library of cosmid clones for use in map construction; (2) identification of a chromosome 16-specific satellite repetitive-sequence probe permitting accurate purity assessments of sorted chromosomes; and (3) the availability,

[***Opening pages: large photomicrograph courtesy of Evelyn Campbell; inset image courtesy of David Ward, Yale University School of Medicine.***]

Table 1. Disease Genes Localized to Human Chromosome 16

Location	Symbol	Cloned	Disease
16p13.3	HBA	Yes	Thalassemia
16p13.3	PKD1	No	Autosomal dominant polycystic kidney disease
16p13.3	MEF	No	Familial Mediterranean fever
16p13.3	RTS	No	Rubinstein-Taybi syndrome
16p12	CLN3	No	Batten's disease (juvenile-onset neuronal ceroid lipofuscionosis)
16q12	PHKB	No	Glycogen-storage disease, type VIIIb
16q13	CETP	Yes	Elevated high-density lipoprotein (HDL), (CETP deficiency)
16q22.1	LCAT	Yes	Corneal opacities, anemia, proteinuria with unesterified hypercholesterolemia (Norum disease)
16q22.1	TAT	Yes	Richner-Hanhort syndrome, oculocutaneous tyrosinemia II (TAT deficiency)
16q22.1	ALDOA	Yes	Hemolytic anemia (ALDOA deficiency)
16q24.3	APRT	Yes	Urolithiasis, 2,5 dihydroxyadenine (APRT) deficiency
16q24	CYBA	No	Autosomal chronic granulomatous disease
16q	CTM	No	Marner's cataract
16q	CMH2	No	Familial hypertrophic cardiomyopathy

through collaboration, of a panel of a large number of hybrid-cell lines containing portions of chromosome 16. This hybrid-cell panel enables probes from chromosome 16 to be localized into intervals along the chromosome having an average length of 1.6 million base pairs.

Chromosome 16 is also interesting to the biomedical community. It contains gene loci for several human diseases of both clinical and economic importance, including polycystic kidney disease, a class of hemoglobin disorders, and several types of cancer (including leukemia and breast cancer). Table 1 lists disease genes that have been localized to chromosome 16 through genetic-linkage analysis. A physical map of overlapping clones for chromosome 16 would facilitate rapid isolation of those genes not yet cloned.

It takes about 2500 cosmid clones laid end to end to represent all the DNA in chromosome 16 once, and so our chromosome 16-specific library of 25,000 cosmid clones represented a tenfold coverage of the chromosome. In 1988, with funds from the Department of Energy, we took on the physical mapping of chromosome 16.

Developing a Mapping Strategy

Our initial strategy for constructing an ordered-clone, or contig, map for chromosome 16 was to fingerprint cosmid clones chosen at random, determine the overlaps between pairs of clones from the similarities between fingerprints, and assemble the clone pairs into contigs, or islands of overlapping clones. This basic clone-to-fingerprint-to-contig strategy, which is described in "Physical Mapping—A One-Dimensional Jigsaw Puzzle" in "Mapping the Genome", had been applied successfully to the mapping of the *E. coli*, yeast, and nematode genomes. However, those maps of less complex genomes had taken many years of work. In addition, the human genome contains many classes of repetitive sequences that tend to complicate the process of building contigs. When faced with the mapping of human chromosome 16, which is about ten times larger than the yeast genome, we needed to develop a strategy that would increase the speed of contig building while retaining the required accuracy.

Lander and Waterman's 1988 analysis of random-clone fingerprinting suggested the key to increased mapping efficiency. That paper showed that the size of the smallest detectable clone overlap was an important parameter in determining the rate at which contigs would increase in length and therefore the rate at which contig maps would near completion. In particular, the calculated rate of progress increases significantly if the detectable clone overlap is reduced from 50 percent to 25 percent of the clone lengths.

In the mapping efforts for yeast and *E. coli*, the overlap between two clones was detected by preparing a restriction-fragment fingerprint of each clone and identifying restriction-fragment lengths that were common to the two fingerprints. With this method, two clones have to overlap by at least 50 percent in order for one to declare with a high degree of certainty that the two clones do indeed overlap. (See "Physical Mapping—A One-Dimensional Jigsaw Puzzle" for a description of restriction-fragment fingerprinting.) Clearly, increasing the information content in each clone fingerprint would make smaller overlaps detectable.

The Repetitive-Sequence Fingerprint

The unique feature of our initial mapping strategy was what we call the repetitive-sequence fingerprint. Repetitive sequences compose 25 to 35 percent of the human genome. The box at right shows the most abundant classes of repetitive sequences and the approximate locations of those sequences on human chromosome 16.

Various Classes of Human Repetitive DNA Sequences

Described below are the most abundant classes of repetitive DNA on human chromosomes. The figure shows the locations of these classes on chromosome 16. Numbers in parentheses indicate the size of continuous stretches of each repetitive DNA class.

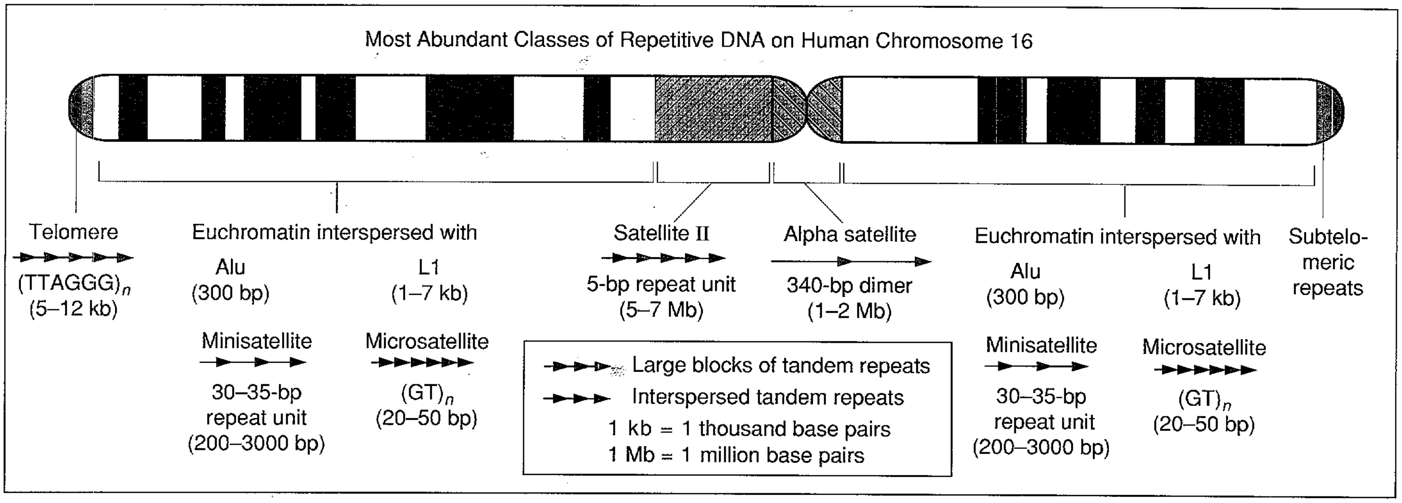
Telomere Repeat: The tandemly repeating unit TTAGGG located at the very ends of the linear DNA molecules in human and vertebrate chromosomes. The telomere repeat (TTAGGG)$_n$ extends for 5000 to 12,000 base pairs and has a structure different from that of normal DNA. A special enzyme called telomerase replicates the ends of the chromosomes in an unusual fashion that prevents the chromosome from shortening during replication.

Subtelomeric repeats: Classes of repetitive sequences that are interspersed in the last 500,000 bases of nonrepetitive DNA located adjacent to the telomere. Some sequences are chromosome specific and others seem to be present near the ends of all human chromosomes.

Microsatellite repeats: A variety of simple di-, tri-, tetra-, and penta-nucleotide tandem repeats that are dispersed in the euchromatic arms of most chromosomes. The dinucleotide repeat (GT)$_n$ is the most common of these dispersed repeats, occurring on average every 30,000 bases in the human genome, for a total copy number of 100,000. The GT repeats range in size from about 20 to 60 base pairs and appear in most eukaryotic genomes.

Minisatellite repeats: A class of dispersed tandem repeats in which the repeating unit is 30 to 35 base pairs in length and has a variable sequence but contains a core sequence 10 to 15 base pairs in length. Minisatellite repeats range in size from 200 base pairs up to several thousand base pairs, have lower copy numbers than microsatellite repeats, and tend to occur in greater numbers toward the telomeric ends of chromosomes.

Alu repeats: The most abundant interspersed repeat in the human genome. The Alu sequence is 300 base pairs long and occurs on average once every 3300 base pairs in the human genome, for a total copy number of 1 million. Alus are more abundant in the light bands than in the dark bands of giemsa-stained metaphase chromosomes. They occur throughout the primate family and are homologous to and thought to be descended from a small, abundant RNA gene that codes for the 300-nucleotide-long RNA molecule known as 7SL. The 7SL RNA combines with six proteins to form a protein-RNA complex that recognizes the signal sequences of newly synthesized proteins and aids in their translocation through the membranes of the endoplasmic recticulum (where they are formed) to their ultimate destination in the cell.



L1 repeats: A long interspersed repeat whose sequence is 1000 to 7000 base pairs long. L1s have a common sequence at the 3′ end but are variably shortened at the 5′ end and thus have a large range of sizes. They occur on average every 28,000 base pairs in the human genome, for a total copy number of about 100,000, and are more abundant in Giemsa-stained dark bands. L1 repeats are also found in most other mammalian species. Full-length L1s (3.5 percent of the total) are a divergent group of class II retrotransposons—"jumping genes" that can move around the genome and are thought to be remnants of retroviruses. [Class II retrotransposons have at least one protein-coding gene and contain a poly A tail (or series of As at the 3′ end) as do messenger RNAs.] Recently, a full-length, functional L1 was discovered. It was found to code for a functional reverse transcriptase—an enzyme essential to the process by which the L1s are copied and re-inserted into the genome.

Alpha satellite DNA: A family of related repeats that occur as long tandem arrays at the centromeric region of all human chromosomes. The repeat unit is about 340 base pairs and is a dimer, that is, it consists of two subunits, each about 170 base pairs long. Alpha satellite DNA occurs on both sides of the centromeric constriction and extends over a region 1000 to 5000 base pairs long. Alpha satellite DNA in other primates is similar to that in humans.

Satellite I, II, and III repeats: Three classical human satellite DNAs, which can be isolated from the bulk of genomic DNA by centrifugation in buoyant density gradients because their densities differ from the densities of other DNA sequences. Satellite I is rich in As and Ts and is composed of alternating arrays of a 17- and 25-base-pair repeating unit. Satellites II and III are both derived from the simple five-base repeating unit ATTCC. Satellite II is more highly diverged from the basic repeating unit than Satellite III. Satellites I, II and III occur as long tandem arrays in the heterochromatic regions of chromosomes 1, 9, 16, 17, and Y and the satellite regions on the short (p) arms of chromosomes 13, 14, 15, 21, and 22.

Cot1 DNA: The fraction of repetitive DNA that is separable from other genomic DNA because of its faster re-annealing, or renaturation, kinetics. *Cot* 1 DNA contains sequences that have copy numbers of 10,000 or greater. ■

Our work on the distribution of repetitive sequences had shown that the tandem-repeat sequence $(GT)_n$, where n is typically between 15 and 30, was scattered randomly across most regions of the human genome with an average spacing of 30,000 base pairs. The in-situ hybridization in Figure 1 shows that $(GT)_n$ is scattered throughout the arms of human chromosomes but is noticeably absent from the regions around the centromere. (The centromeric regions consists of large blocks of tandem-repeat sequences known as satellite DNA. Gene sequences are absent from these regions. Regions containing large blocks of tandem repeats are known as heterochromatin, and regions devoid of large tandem repeat blocks are known as euchromatin.)

We reasoned that the sequence $(GT)_n$ would appear, on average, about once in each cosmid clone containing a human DNA insert of 35,000 base pairs from the euchromatic arms of chromosome 16. Therefore, we could enrich the information content of the usual restriction-fragment fingerprint of each clone by determining, through hybridization of a radio-labeled $(GT)_{25}$ probe, which restriction fragments in each fingerprint contain the $(GT)_n$ sequence. As we will illustrate below, this information allowed us to detect overlaps between cosmid clones that were as small as 10 percent of their lengths.

To reduce the initial complexity of the mapping, we preselected from our chromosome 16-specific library of clones (through hybridization) those clones that were positive for the $(GT)_n$ sequence and negative for satellite DNA. In other words, we chose to build contigs around those sites in chromosome 16 that contain $(GT)_n$. Since those sites are widely scattered across the chromosome, we expected those contigs to cover the chromosome in a fairly uniform way except for the centromeric region, which can be mapped using an alternative approach. We identified about 3000 $(GT)_n$-positive clones from our library and made a repetitive-sequence fingerprint for each one.

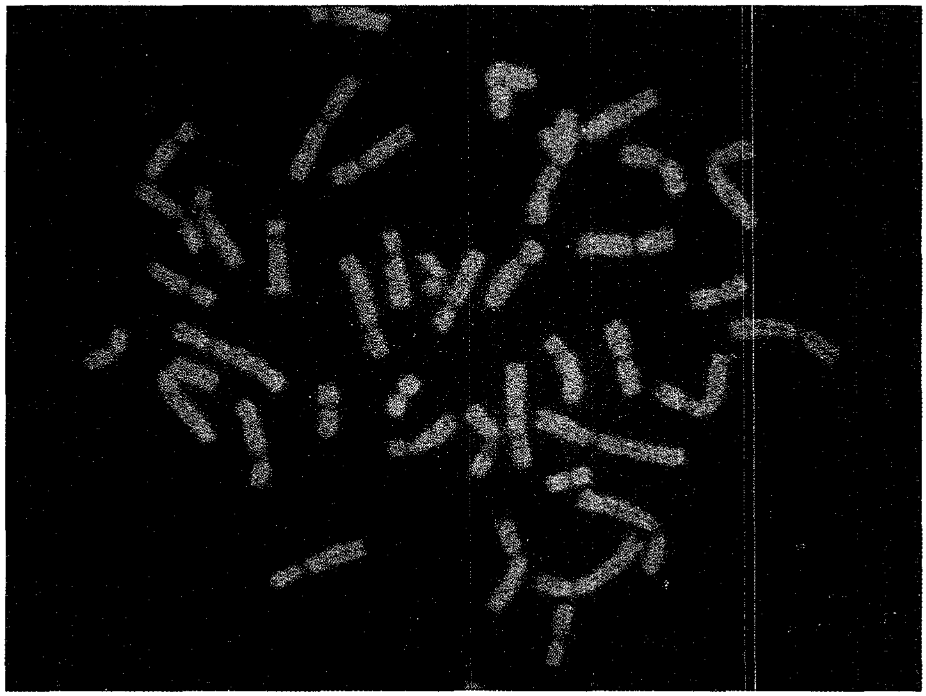


Figure 1. GT Hybridization on Human Chromosomes

The photomicrograph shows in-situ hybridization of human chromosomes using biotin-labeled $(AC)_{25}$ as a probe (yellow). $(AC)_{25}$ hybridizes to sites of the microsatellite repeat $(GT)_n$. Those sites are underrepresented at the centromeric regions of some chromosomes and at the distal half of Yq. However, $(GT)_n$ appears to be uniformly distributed on all euchromatic regions of the human genome.

The repetitive-sequence fingerprint was made by digesting each cosmid clone with restriction enzymes, sizing the resulting restriction fragments, and determining which of those fragments contain $(GT)_n$ as well as another type of repetitive DNA known as *Cot*1, which is also scattered throughout the arms of the chromosome (see box). *Cot*1 is the most abundant fraction of repeated DNA in the human genome, consisting predominantly of Alu and L1 repeated sequences.

The first step in fingerprinting was to isolate many copies of the DNA insert in each cosmid clone, divide those copies into three batches, and digest each batch with the restriction enzymes *Eco*RI, *Hin*dIII, and a mixture of both *Eco*RI and *Hin*dIII, respectively. The restriction fragments from each of the three digests were separated in parallel along three lanes of an agarose gel by electrophoresis. DNA fragments having known lengths were separated on adjacent lanes to determine the fragment lengths from each restriction-enzyme digest. The fragments in the gel were stained with ethidium bromide (a fluorescent dye that binds to DNA) and the gel was photographed under ultraviolet light to produce an image

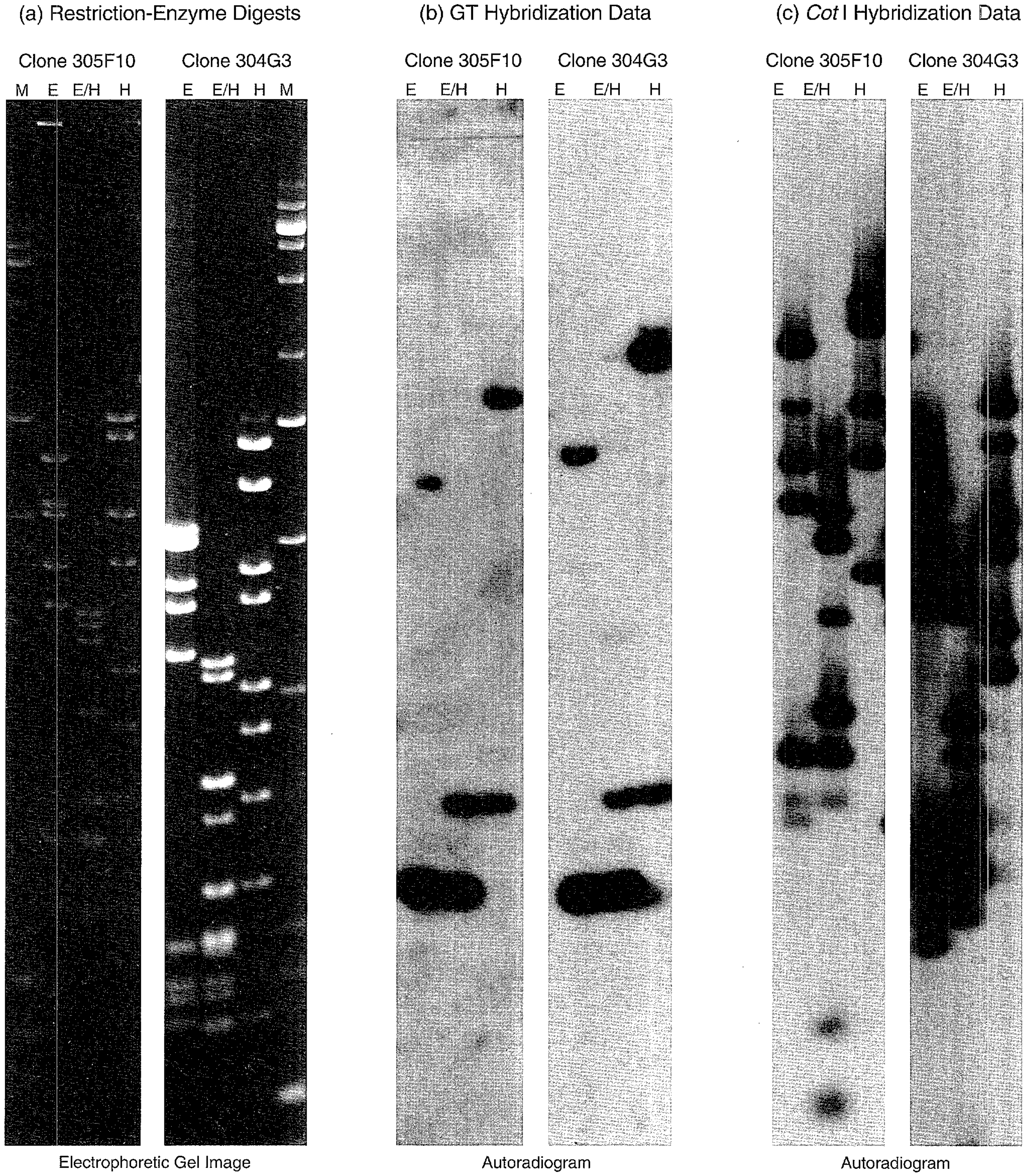


Figure 2. Repetitive Sequence Fingerprints of Two Overlapping Cosmid Clones

The repetitive-sequence fingerprint of a clone has three parts. The figure shows a comparison of those parts for two clones that have a high likelihood of overlap based on the similarities between their fingerprints. (a) Fluorescent images of DNA fragments separated by agarose gel electrophoresis. The three gel lanes for each clone contain the restriction fragments produced by completely digesting that clone with the restriction enzymes *Eco*RI (E), *Eco*RI and *Hind*III (E/H), and *Hind*III (H), respectively. The marker lanes (M) contain standard fragments of known lengths, which are used to calibrate the restriction-fragment lengths. (b) Autoradiographic images of the gels in (a) after hybridization with the GT probe. (c) Autoradiographic images of the gels in (a) after hybridization with the *Cot*I probe. Clone 305F10 and clone 304G3 have identical GT-hybridization patterns, a strong indication of overlap.

showing the distinct bands of DNA fragments in the gel, each band made up of many copies of a particular restriction fragment. This gel image was then digitized with a CCD camera, the DNA fragments were assigned sizes according to their positions on the gel relative to the known fragment lengths using a commercial software package. These sizes were the stored in our mapping database. Figure 2 shows the gel images for two clones that were determined to overlap one another based on their complete repetitive-sequence fingerprints.

The second step in fingerprinting was to determine which restriction fragments contained $(GT)_n$ and *Cot*1 repetitive DNA. We accomplished this step using standard hybridization techniques. (See "Hybridization Techniques" in "Understanding Inheritance.") Specifically, DNA from each gel was transferred to two different nylon or nitrocellulose membranes using the blotting procedure developed by Edwin Southern in 1975. This blotting procedure preserves the relative positions that the DNA fragments have on the gel. Once the fragments are immobilized on the two membranes, radio-labeled copies of the $(GT)_n$ sequence are used as hybridization probes on one membrane and radio-labeled copies of the *Cot*1 sequences are used as probes on the second membrane. The bands of fragments that contain those sequences and therefore bind, or hybridize, to the radioactive probes can be visualized by exposing an x-ray film to the membrane, a process known as autoradiography. Alongside the gel images shown in Figure 2 are the corresponding autoradiographs, or blot images, produced by the $(GT)_n$ hybridization and *Cot*1 hybridization. Together, the gel image and the two blot images for each clone constitute the repetitive-sequence fingerprint of that clone.

The fingerprint data are scored by first noting the lengths of the restriction fragments on the gel image. Then the gel image and the two blot images for each clone are aligned to determine the hybridization score of each band of restriction fragments. To help us accomplish this task for thousands of clones in an efficient manner, Mike Cannon of the Computer Division at Los Alamos developed a computer program called SCORE. This program takes the fragment lengths determined from the gel image and creates a schematic of the gel image. The blot image is then scanned, and its image size is adjusted to match the schematic of the gel image. Each band is then scored for the presence or absence of a positive hybridization signal from the $(GT)_n$ probe and for the degree of hybridization of the *Cot*1 probe. *Cot*1 creates a low, medium, or high hybridization signal depending on whether the restriction fragment contains short, intermediate, or long stretches of *Cot*1 sequences. (Operation of the SCORE program is illustrated in "SCORE: A Program for Computer-assisted Scoring of Southern Blots" in "Computation and the Human Genome Project.")

Determining the Likelihood That Two Clones Overlap

Once the clones have been fingerprinted and the fingerprint data scored and entered into the database, the next step is to determine from the similarities between fingerprints which pairs of clones overlap one another. The problem of determining clone overlap from such fingerprint data is probabilistic, as explained in "Physical Mapping—A One-Dimensional Jigsaw Puzzle." We have two types of information, the sizes of the restriction fragments and the hybridization scores for each fragment. The two questions we need to answer are: Given that the fingerprints of two clones share certain restriction-fragment lengths and hybridization scores, first, what is the probability that they overlap? and second, what is the extent of that overlap?

The first question was addressed by David Torney, a member of the Theoretical Biology and Biophysics Group at Los Alamos. He and his collaborator David Balding developed a complete statistical analysis of the problem, taking into account the known statistical properties of the restriction-fragment lengths, experimental errors in restriction-fragment lengths, hybridization errors, and the expected distribution of the repetitive sequences. They also developed a simplified computer algorithm based on their complete theoretical analysis and on extensive analysis of the actual fingerprint data generated at Los Alamos. That algorithm determines the likelihood that two cosmid clones overlap given the repetitive-sequence fingerprints of those clones.

Figure 3 illustrates how the information content in the repetitive-sequence fingerprint allows the detection of small overlaps. In particular, when $(GT)_n$ is present in the overlap region of two clones, the similarities between the repetitive-sequence fingerprints of those clones yield a nearly unambiguous signature of overlap, even if the region of overlap is small. In the example shown, clones A and B have only a 10 percent overlap, but the overlap region contains the single $(GT)_n$ sequence present on those clones along with two cutting sites for *Eco*RI and one cutting site for *Hin*dIII. Consequently the GT hybridization patterns on the blot images of the two clones are identical within experimental errors and contain one GT-positive band for each restriction-enzyme digest. The likelihood that two

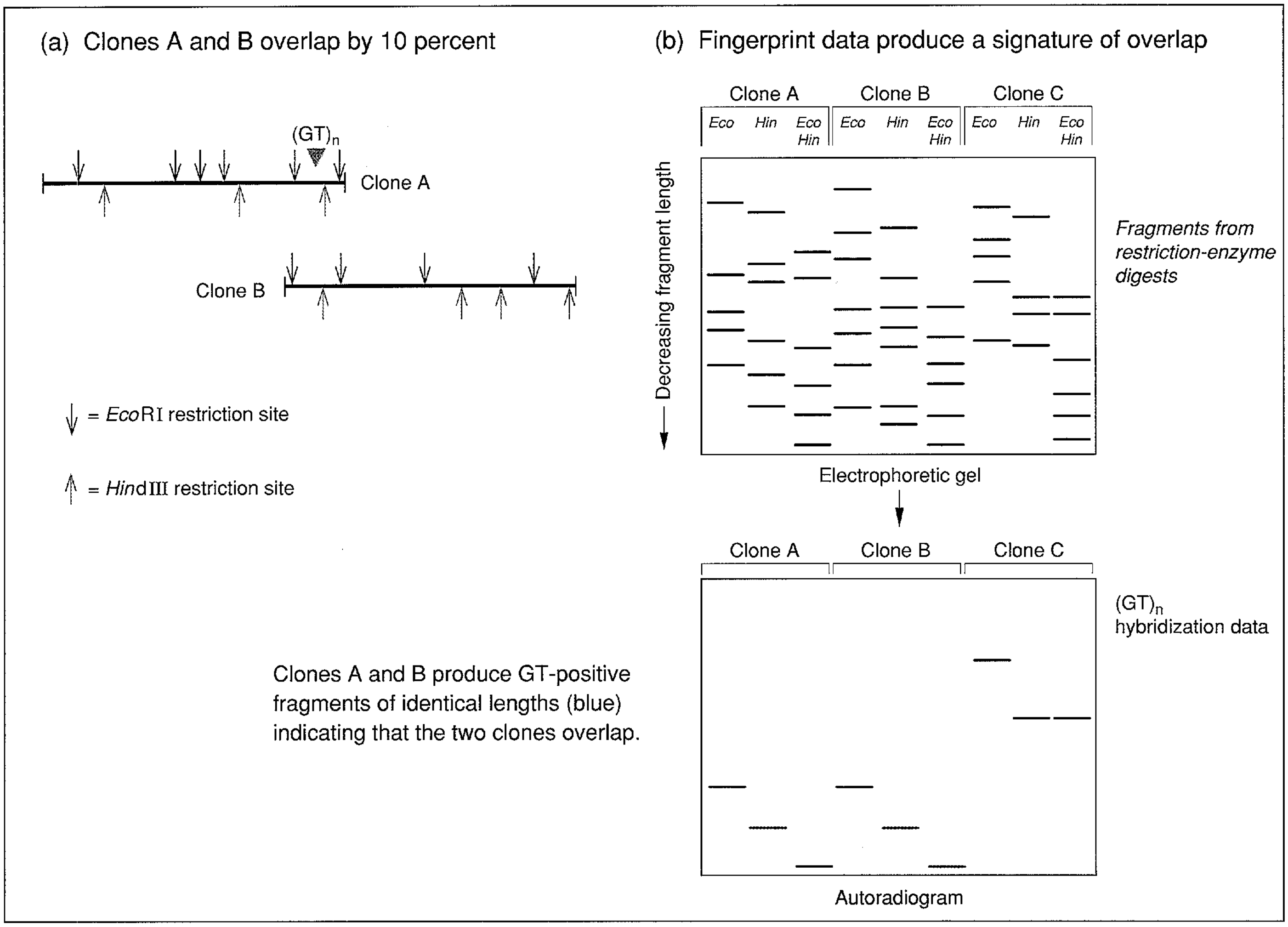

(a) Clones A and B overlap by 10 percent
(GT)n
Clone A
Clone B
= EcoRI restriction site
= HindIII restriction site
Clones A and B produce GT-positive fragments of identical lengths (blue) indicating that the two clones overlap.
(b) Fingerprint data produce a signature of overlap
Clone A
Clone B
Clone C
Eco
Hin
Eco Hin
Decreasing fragment length
Fragments from restriction-enzyme digests
Electrophoretic gel
(GT)n hybridization data
Autoradiogram

Figure 3. Detection of Small Clone Overlaps Using Repetitive-Sequence Fingerprints

Shown in (a) is a diagram of two clones, A and B, that overlap by 10 percent of their lengths. Arrows indicate restriction (cutting) sites for the restriction enzymes *Eco*RI and *Hin*dIII. Clones A and B each contain a single $(GT)_n$ site, which happens to occur in the short overlapping region. Shown in (b) is a diagram of the restriction-fragment fingerprints and corresponding $(GT)_{25}$ hybridization data produced from clones A and B as well as a third clone C. The identical $(GT)_n$ hybridization pattern from clones A and B is sufficient information to infer that the two clones have a very high likelihood of overlap.

such identical patterns would arise from non-overlapping clones is extremely low. In general, if two cosmid clones from our chromosome-specific library produce the same GT-hybridization pattern, they have an extremely high probability of overlapping, even if they share only one GT-positive region.

The detailed computer algorithms used to estimate the probability of clone overlap from the fingerprint data will not be presented here. Suffice it to say those algorithms are based on Bayes' theorem for conditional probabilities and use parameters for estimating errors in restriction-fragment sizes and hybridization results that were determined through detailed statistical analysis of the experimental conditions. The computer algorithms were used to examine all possible pairs of fingerprinted clones and determine the probability of overlap for each clone pair.

Assembling the Contig Map

As illustrated in "Physical Mapping—A One-Dimensional Jigsaw Puzzle," restriction-fragment fingerprint

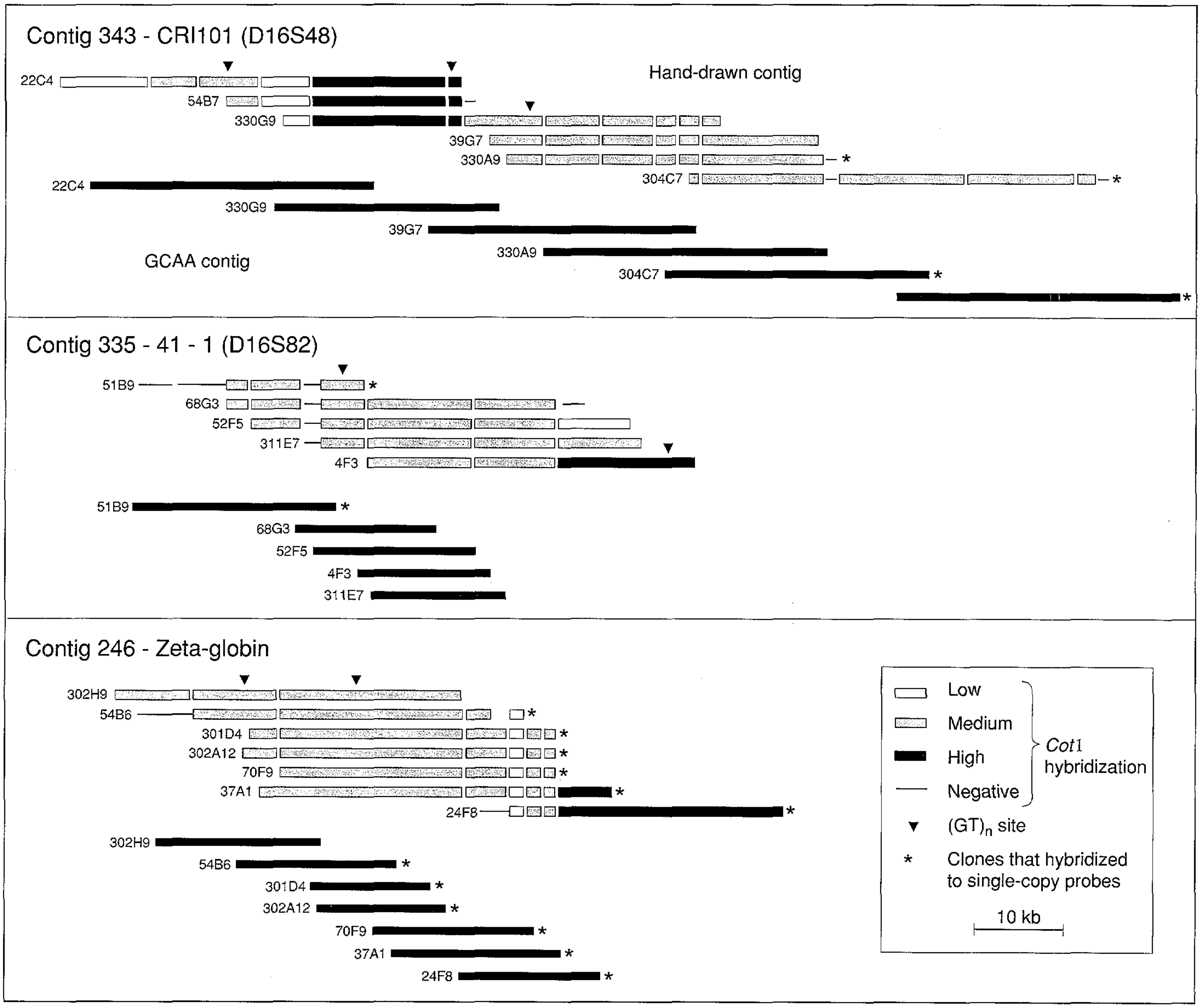


Figure 4. Comparison of Hand-drawn and Computer-generated Cosmid Contigs from Chromosome 16.
Groups of overlapping clones are arranged into contigs showing the linear arrangement and extents of clone overlap deduced from repetitive-sequence fingerprint data. The hand-drawn representations show which restriction fragments were positive for GT and *Cot* 1 hybridization probes and provides a partial ordering of the restriction fragments. The corresponding GCAA-generated contig shows the extent of overlap between clones and the contig length. Additions to GCAA are planned that will enable the algorithm to generate contigs similar to the hand-drawn contigs. As shown, the GCAA contigs sometimes differ in length from the hand-drawn contigs.

data can be used to assemble islands of contiguous, overlapping clones showing the position of each clone relative to the others and the extent of overlap between each pair of overlapping clones.

Initially we assembled contigs by sorting the output of the pairwise overlaps into sets of multiply overlapping clones. More recently Jim Fickett and Michael Cinkosky of the Laboratory's Theoretical Biology and Biophysics Group developed a "genetic algorithm" for contig assembly called GCAA, which has sped up this process considerably. The algorithm is based on optimization theory. Figure 4 compares hand-drawn cosmid contigs for chromosome 16 with versions generated by the genetic algorithm. The hand-drawn contigs are sometimes more accurate, but each one takes many hours to construct. In contrast, the computer algorithm can handle data from thousands of clones and construct hundreds of contigs automatically in a short time. It also allows manual changes to be made through interactive software. The genetic algorithm has been invaluable to our mapping efforts, as has the whole suite of informatics tools developed at Los Alamos for managing, analyzing, utilizing, and sharing mapping data. Some of those tools are described in

"Computation and the Genome Project."

About 3145 GT-positive cosmid clones and an additional 800 GT-negative cosmid clones were fingerprinted and then assembled into contigs in the manner described above. The clones formed 576 contigs with an average size of 100,000 base pairs and containing, on average, four or five clones. The largest cosmid contig spanned approximately 300,000 base pairs. These contigs cover about 58 million base pairs, or 58 percent of chromosome 16. There were also 1171 singletons (single fingerprinted clones not contained within a contig). Experiments discussed below suggest that the singletons cover 26 percent of the chromosome. Together the 4000 fingerprinted clones cover about 84 percent of chromosome 16.

If the minimum detectable overlap between clones is 50 percent of the clone lengths, the equations of Lander and Waterman suggest that one would have to fingerprint about 16,000 clones of an average length of 35,000 base pairs to reach an average contig size of 100,000 base pairs for a chromosome the length of chromosome 16. We reached an average contig size of 100,000 base pairs after fingerprinting only 4000 clones. That reduction was due to two factors. First, the repetitive-sequence fingerprints enabled the detection of clone overlaps composing between 10 and 25 percent of the clone lengths depending on the positions of the $(GT)_n$ sites. In fact, the average length of each detected overlap region was 20 percent of the clone lengths. Second, we did not fingerprint clones at random but rather preselected clones containing $(GT)_n$. By focusing our mapping efforts around regions of $(GT)_n$ sites, we effectively reduced the size of the region that was being mapped during the initial phases of mapping. These two factors resulted in the rapid construction of relatively large cosmid contigs.

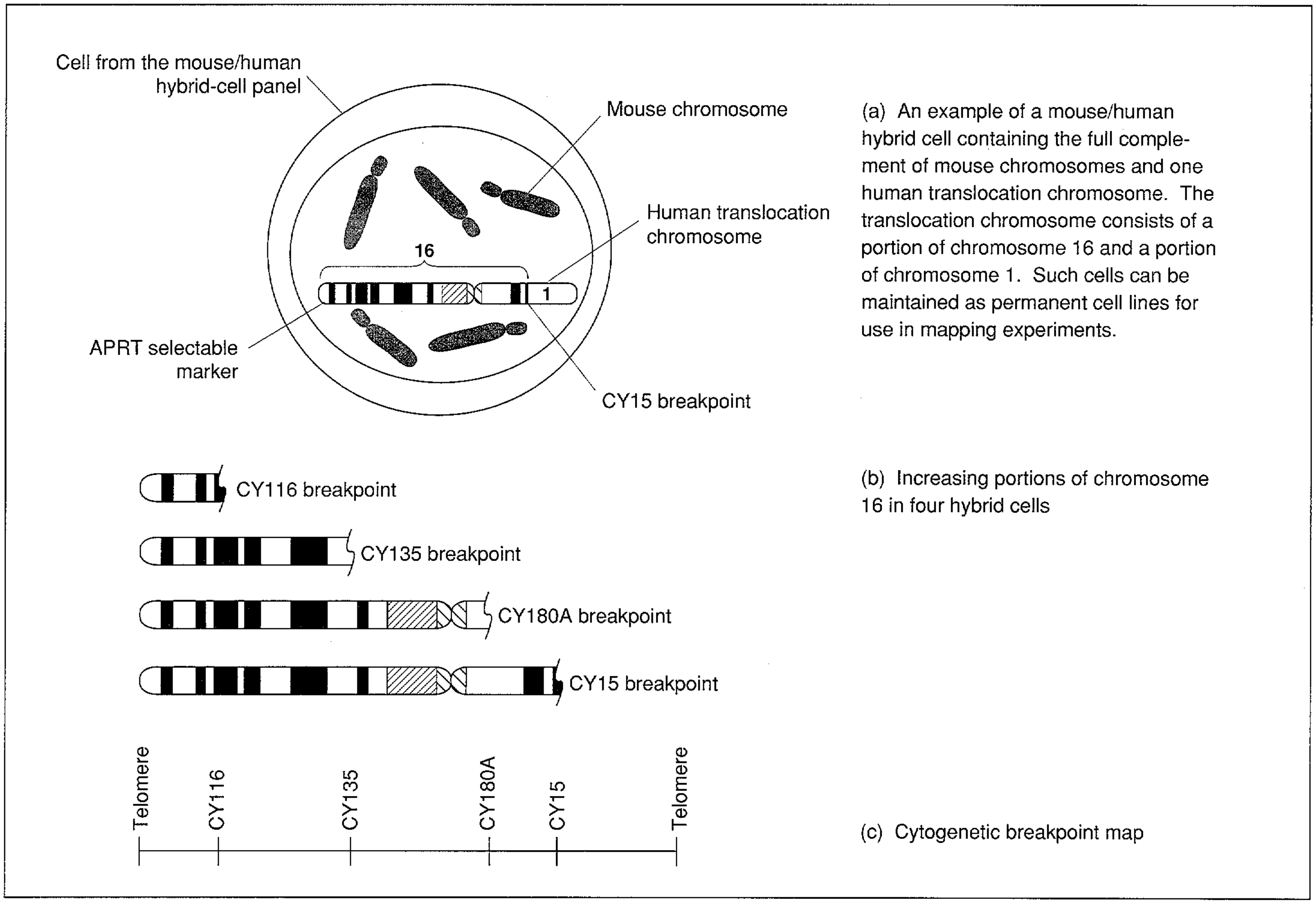
Several other features are distinctive about our cosmid-fingerprinting approach. By sizing the restriction fragments from each clone, we know the extent of overlap between clones in a contig, and therefore we can estimate the length of each contig. In contrast, mapping methods that determine clone overlap from hybridization-based or STS data alone cannot determine the extent of the overlap or the length of the contigs without further analysis. Restriction-fragment lengths also provide us with information to generate partially ordered restriction maps for each contig. Finally, as a result of the GT and *Cot*1 hybridizations, we know which fragments contain GT repeats and which fragments contain *Cot*1 DNA. A GT repeat at a given site in the genome varies in length among the population and therefore provides a source of polymorphic markers for genetic-linkage mapping. Our contig map thus provides the positions of fragments containing those potential markers. The *Cot*1 hybridization is useful because fragments that do not hybridize to the *Cot*1 probe are free of the most abundant classes of repetitive DNA and are therefore likely to contain single-copy sequences, which may be candidates for genes. Finally, as the map is further developed and the repetitive-sequence distribution more accurately determined, it may reveal new insights into genome organization and the molecular evolution of mammalian chromosomes.

Evaluation of the Cosmid Contig Map

After constructing the 576 cosmid contigs, we first wanted to ascertain their distribution on chromosome 16. David Callen and Grant Sutherland in Australia located 140 of our cosmid contigs on their panel of mouse/human hybrid cells. The 50 different hybrid cells in their panel contain, in addition to the full complement of mouse chromosomes, increasingly longer portions of human chromosome 16, starting from the far end of the long arm of the chromosome (see Figure 5). In effect, the panel divides the chromosome into bins, or intervals, 1.6 million base pairs in length. They found the 140 cosmid contigs to be distributed evenly over the intervals defined by the hybrid-cell panel.

Second, to evaluate the accuracy of the contigs, we picked 19 pairs of clones from 11 different contigs and checked whether each pair that had been assigned to the same contig hybridized to the same large restriction fragment and therefore came from the same region of chromosome 16. The DNA for these experiments was isolated from a mouse/human hybrid-cell line containing human chromosome 16 only. Eight rare-cutting restriction enzymes were used to make eight different complete digests of the DNA, and the resulting large restriction fragments were separated in parallel by pulsed-field gel electrophoresis. The fragments were then blotted onto filters, and each filter was probed with one clone from each pair. This analysis confirmed that the two members of each of the 19 clone pairs came from the same region of the genome.

A second check on contig accuracy involved hybridization of 43 single-copy probes (probes containing sequences that appear only once in the human genome) to membranes containing a gridded array of our 4000 fingerprinted clones. The single-copy probes were graciously provided by a large number of collaborators and associates. Ideally, if a single-copy probe hybridizes to more than one clone, those clones should be contained within a single contig and



(a) An example of a mouse/human hybrid cell containing the full complement of mouse chromosomes and one human translocation chromosome. The translocation chromosome consists of a portion of chromosome 16 and a portion of chromosome 1. Such cells can be maintained as permanent cell lines for use in mapping experiments.

(b) Increasing portions of chromosome 16 in four hybrid cells

(c) Cytogenetic breakpoint map

Figure 5. Hybrid-Cell Panel and the Cytogenetic Breakpoint Map for Chromosome 16

A panel of 50 different mouse/human hybrid cells, each containing an increasingly longer portion of chromosome 16 starting from the tip of the long arm of the chromosome, is a convenient tool for constructing a low-resolution physical map of the chromosome. The hybrid cells are formed by fusing mouse cells with human cells and growing them in a medium in which only those cells containing a particular gene (APRT) can survive. Thus APRT is called a selectable marker. It is near the end of the long, or q, arm of chromosome 16. During the fusion process and subsequent growth, human chromosomes that lack the selectable marker are lost, resulting in a mouse/human hybrid containing a single human chromosome 16. The 50 different hybrids were derived from a collection of patients' cells that had each undergone translocations (breakage and rejoining) of chromosome 16 with another human chromosome. (a) The type of hybrid cell produced by the fusion process and selectively grown for inclusion in the panel is shown. The hybrid cell contains the full complement of mouse chromosomes and one chromosome produced by a translocation between human chromosomes 16 and 1. Because this chromosome includes the portion of the q arm of chromosomes 16 containing APRT, it survived the fusion and selective growth process. (b) Increasing portions of chromosome 16 contained in some of the hybrid cells of the panel are shown. The panel contains 50 hybrid cells and, in effect, divides the chromosome into intervals with an average length of 1.6 million bases. Each portion ends at a so-called breakpoint of the chromosome, a natural site of chromosomal translocation. (c) A cytogenic map of chromosome 16 indicating the locations of the breakpoints in (b). The complete cytogenetic breakpoint map derived from the hybrid cell panel contains 50 breakpoints separated by intervals with an average length of 1.6 million base pairs. A human DNA probe or clone from chromosome 16 can be localized to a region between two breakpoints by showing that it hybridizes to the DNA from all hybrid cells containing that region and *does not* hybridize to the DNA from the hybrid cell in which that region is absent.

should overlap one another because they contain the same unique sequence. Our analysis showed no unequivocal false-positive overlaps in our contigs, and it also enabled us to detect overlaps between some singleton clones and our existing contigs.

The hybridizations of single-copy probes to the gridded arrays of fingerprinted clones also allowed us to estimate how much of chromosome 16 is covered by our fingerprinted clones. Out of 43 probes, 25 hybridized to clones within contigs, 11 hybridized to singletons, and 7 did not hybridize to any of the fingerprinted clones. These results suggest that our cosmid contigs cover 58 percent of chromosome 16, and the singleton cosmids cover 26 percent of the chromosome for a total coverage of 84 percent.

Our goal was to construct a map composed of at most 100 contigs, each having an average size of about a million base pairs. Having already achieved substantial coverage, we were at a point where continued random fingerprinting of cosmid clones was no longer the most efficient way to achieve this goal. At that point the likelihood of fingerprinting a new clone that was not yet represented in contigs was diminishing, while the likelihood that the new clone would fall within pre-existing contigs was increasing. The gaps between cosmid contigs could be closed by a directed approach called chromosome walking (see Figure 9 in "DNA Libraries") but to "walk" from one cosmid clone to the next would be a very slow and labor-intensive process.

Fortunately, by that time YAC technology had matured. In 1991 Mary Kay McCormick at Los Alamos successfully constructed chromosome 21-specific YAC libraries from flow-sorted chromosomes using a modified cloning technique. Eric Green and Maynard Olson at Washington University, in collaboration with Bob Moyzis and coworkers at Los Alamos, had developed a substantial number of STS markers for chromosome 7 from our chromosome 7-specific library of M13 clones (a library of cloned single-stranded DNA fragments for sequencing). They thereby demonstrated the feasibility of generating large numbers of STS markers for use in physical mapping.

Green and Olson had already used STS-content mapping to construct a contig of YAC clones covering the region surrounding the cystic-fibrosis gene. In particular, they had developed a set of STS markers from pre-existing genetic-linkage markers, which had been used to find the gene, and from cDNAs for sequences within the cystic-fibrosis gene. Then they used those STSs to screen a YAC library made from total-genomic human DNA and pick out the YAC clones containing each marker. Two YACs that contain the same STS marker must overlap because each STS is a unique sequence that has been shown to appear only once on the genome. Thus, based on the STSs contained in each YAC, they were able to construct a contig of overlapping YAC clones spanning about 1.5 million base pairs and containing the cystic-fibrosis gene.

These advances made it feasible for us to consider closing the gaps in our cosmid contig map with YAC clones from chromosome 16. We decided that the most efficient strategy would be to work with a chromosome 16-specific YAC library.

Improving YAC Cloning Techniques

YACs are cloning vectors that replicate as chromosomes in yeast cells and can accommodate human DNA inserts as large as 1 million base pairs. These large inserts are extremely useful for attaining long-range continuity in contig maps, and therefore the use of YAC clones in large-scale mapping of the human genome was becoming widely adopted by 1990.

From our point of view, however, prior to McCormick's work at Los Alamos on improving YAC cloning techniques, YAC cloning had some serious drawbacks. First, large amounts of human DNA were required to construct YAC clone libraries because the efficiency of transforming yeast cells by the addition of a YAC clone was relatively low. Consequently, creating a chromosome 16-specific library of YAC clones from the small DNA samples obtained by sorting chromosomes would be difficult if not impossible.

Second, we knew that 30 to 50 percent of the clones in most YAC libraries were chimeric, that is, they contained DNA from two or more nonadjacent regions of the genome. Such clones can be produced when more than one YAC or partial YAC recombinant molecule enters a yeast cell, and, during the transforming process, the human DNA inserts in these recombinant molecules recombine with each other to produce a YAC containing two different human inserts instead of only one. Chimeras are also produced when two DNA fragments are accidentally ligated prior to their ligation with the vector arms of the yeast artificial chromosome.

Chimeric YACs can be identified during the construction of contig maps, but when a large percentage of clones in a YAC clone library are chimeric, the difficulty of map construction increases considerably and the process is error-prone.

These two major difficulties were overcome in 1991 when McCormick succeeded in constructing a chromosome 21-specific YAC library from sorted chromosomes. Not only was she able to work with small amounts of DNA but also only a few percent of the resulting clones are chimeric. The modified cloning techniques she developed to accomplish this technical tour de force are described in "Libraries from Flow-sorted Chromosomes." Following this breakthrough, McCormick applied the new YAC-cloning techniques to the construction of a chromosome 16-specific YAC library for specific use in our mapping effort.

Closing Gaps in the Contig Map with YACs

The YAC library for chromosome 16 contains about 550 clones, and the clones contain inserts with an average size of 215,000 base pairs. Assuming that our 576 cosmid contigs are randomly distributed over chromosome 16, we estimate that the average gap between cosmid contigs is 65,000 base pairs. Thus each gap should be closed with a single YAC clone. Figure 6 outlines our procedure for incorporating YAC clones into the cosmid contig map. We first develop STS markers from the end clones of our cosmid contigs. We then use PCR-based screening to pick out YAC clones that contain each STS and therefore overlap with the cosmid contig from which the STS was derived. Details of this work are presented in "The Polymerase Chain Reaction and Sequence-Tagged Sites" in "Mapping the Genome," and the design of the pooling scheme used to screen the YAC library is described in an accompanying sidebar "YAC Library Pooling Scheme for PCR-based Screening."

Figure 7 presents the results of screening the library for one STS. To date, we have made 89 STS markers from end clones of cosmid contigs and have incorporated 30 YAC clones into the contig map by showing that they contain STSs derived from those end clones.

Figure 6. YAC Closure of Gaps in the Cosmid Contig Map

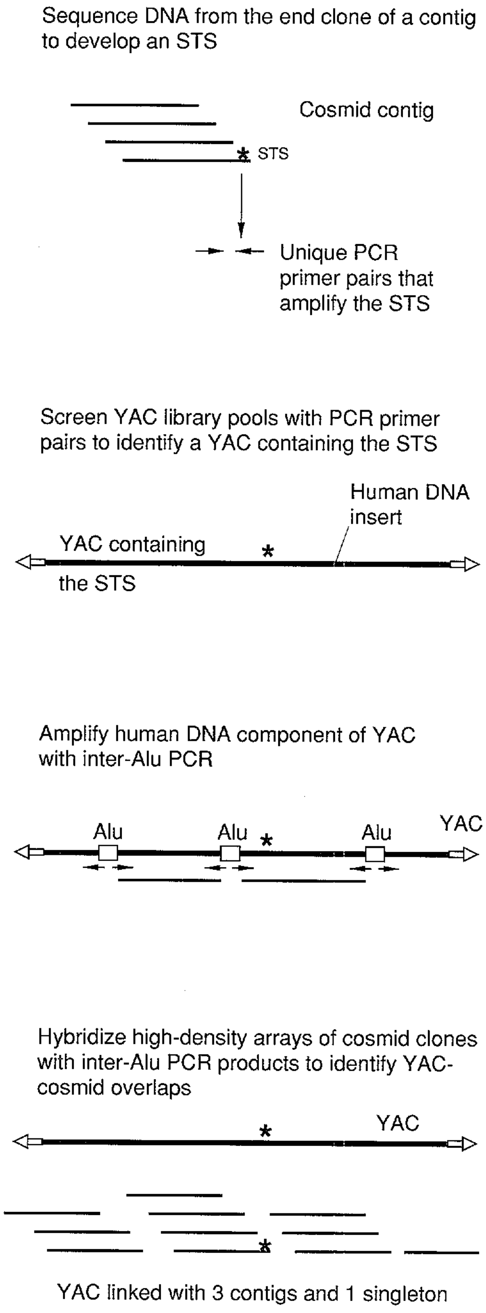
Both STS markers and YAC inter-Alu PCR products are being used to identify overlaps between chromosome 16 YAC clones and our cosmid contigs. The procedure is outlined below.

(a) Sequence-tagged sties (STSs) are generated from the end clones of cosmid contigs. This involves sequencing about 300 base pairs from the end clone, identifying a pair of candidate primer sequences, synthesizing the primers, and checking that the two primers, when used in the polymerase chain reaction, will amplify a single region of the genome. If so, the amplified region is an STS.

(b) YAC clones containing the STS are identified by PCR-based screening of pools of YAC clones from our chromosome 16-specific YAC library. A YAC containing the STS must overlap the cosmid clone from which the STS was derived. Figure 8 illustrates the steps in the screening process.

(c) To identify all cosmid clones that overlap with a YAC, inter-Alu PCR products are generated from each YAC and labeled for use as a hybridization probe. (Note that the inter-Alu products represent only a portion of the human insert in the YAC clones.)

(d) The probe is then hybridized to membranes containing high-density arrays of fingerprinted cosmid clones. Cosmid clones that yield positive hybridization signals must overlap the YAC. A single YAC often overlaps several cosmid contigs, as shown in the figure. However, the hybridization data do not determine the relative positions of the cosmid contigs.


Sequence DNA from the end clone of a contig to develop an STS
Cosmid contig
STS
Unique PCR primer pairs that amplify the STS
Screen YAC library pools with PCR primer pairs to identify a YAC containing the STS
Human DNA insert
YAC containing the STS
Amplify human DNA component of YAC with inter-Alu PCR
Alu
Alu
Alu
YAC
Hybridize high-density arrays of cosmid clones with inter-Alu PCR products to identify YAC-cosmid overlaps
YAC
YAC linked with 3 contigs and 1 singleton

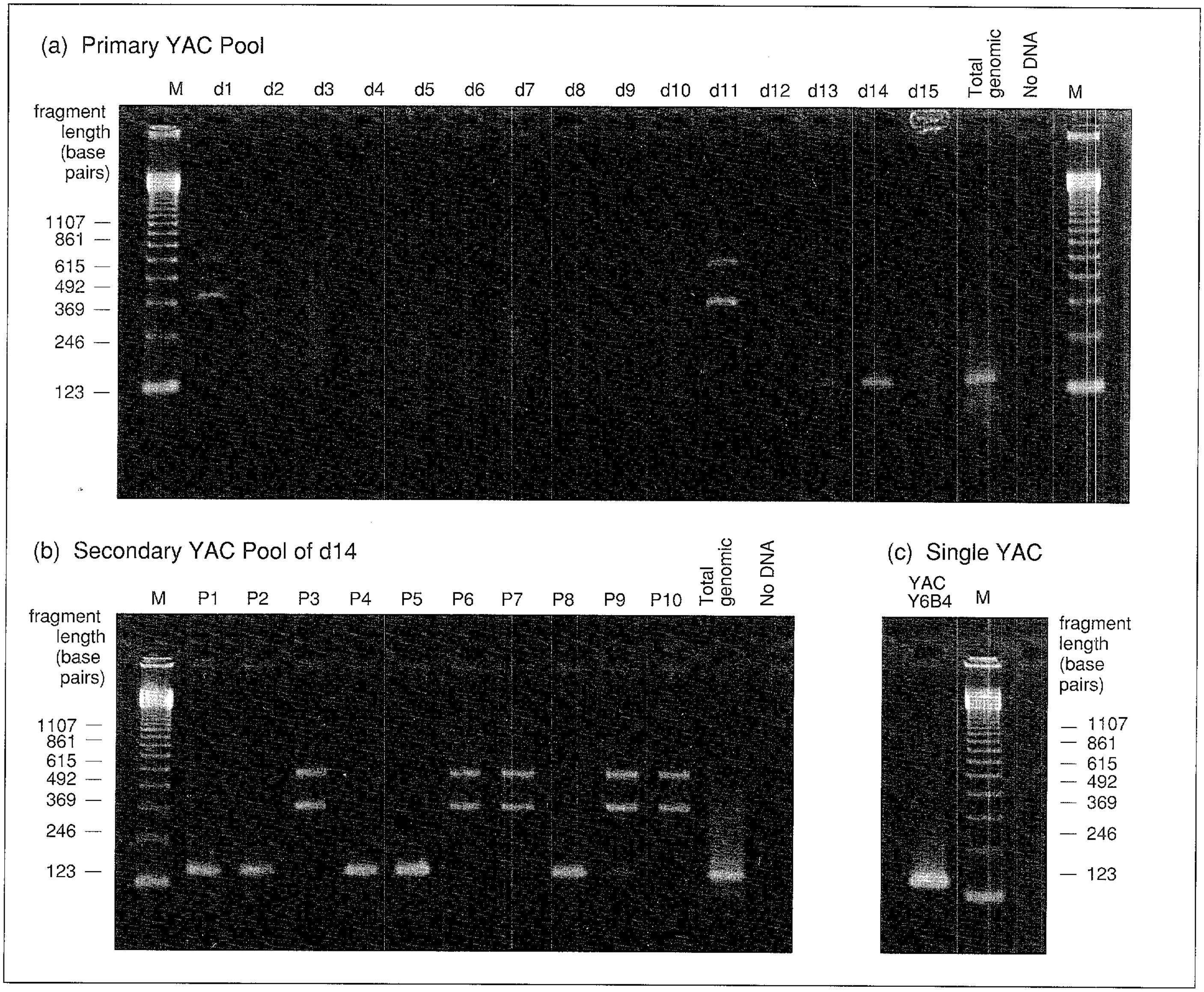


Figure 7. PCR-based Screening of YAC Library Pools for Clones Containing an STS

Our library of 540 YACs was divided into 15 sets of 36 YACs each. These 15 sets are called the primary pools, or detectors, and are numbered d1 through d15. The 36 YACs in each primary pool are then divided into 10 secondary pools (p1 through p10) according to David Torney's design for the 1-detector (see "YAC Library Pooling Scheme for PCR-based Screening" in "Mapping the Genome"). Each of the 36 YACs occur in 5 pools of the 1-detector. (a) An electrophoretic gel in which the PCR products produced by screening the primary pools for STS 25H11 have been separated by length. The lane third from the right, marked "total genomic DNA," contains the STS 25H11, which was amplified from total-genomic human DNA. In this experiment only detector 14 produced a PCR product that has the same length as STS 25H11. Multiple bands at different lengths in lanes d1 and d11 indicate PCR amplification of regions other than STS 25H11 and can therefore be ignored. (b) To determine which YAC was responsible for the positive signal from primary pool d14, we screen the 10 secondary pools composing the 1-detector for d14. Five of these pools, p1, p2, p4, p5, and p8, were identified as positive for STS 25H11. YAC clone Y6B4 was the only YAC that occurred in each of these five pools. (Multiple bands in p3, p6, p7, p9, and p10 were again the result of spurious PCR amplification and did not match the length of STS 25H11.) (c) Finally, the PCR was run on YAC Y6B4 only. The results confirm that this YAC contains STS 25H11. This pooling strategy allows error correction of false negatives in the secondary pools. If less than five positives were identified, this would increase the number of likely candidate YACs that could then be individually checked to find the correct YAC. In other pooling strategies, false negatives lead to the loss of YAC candidates.

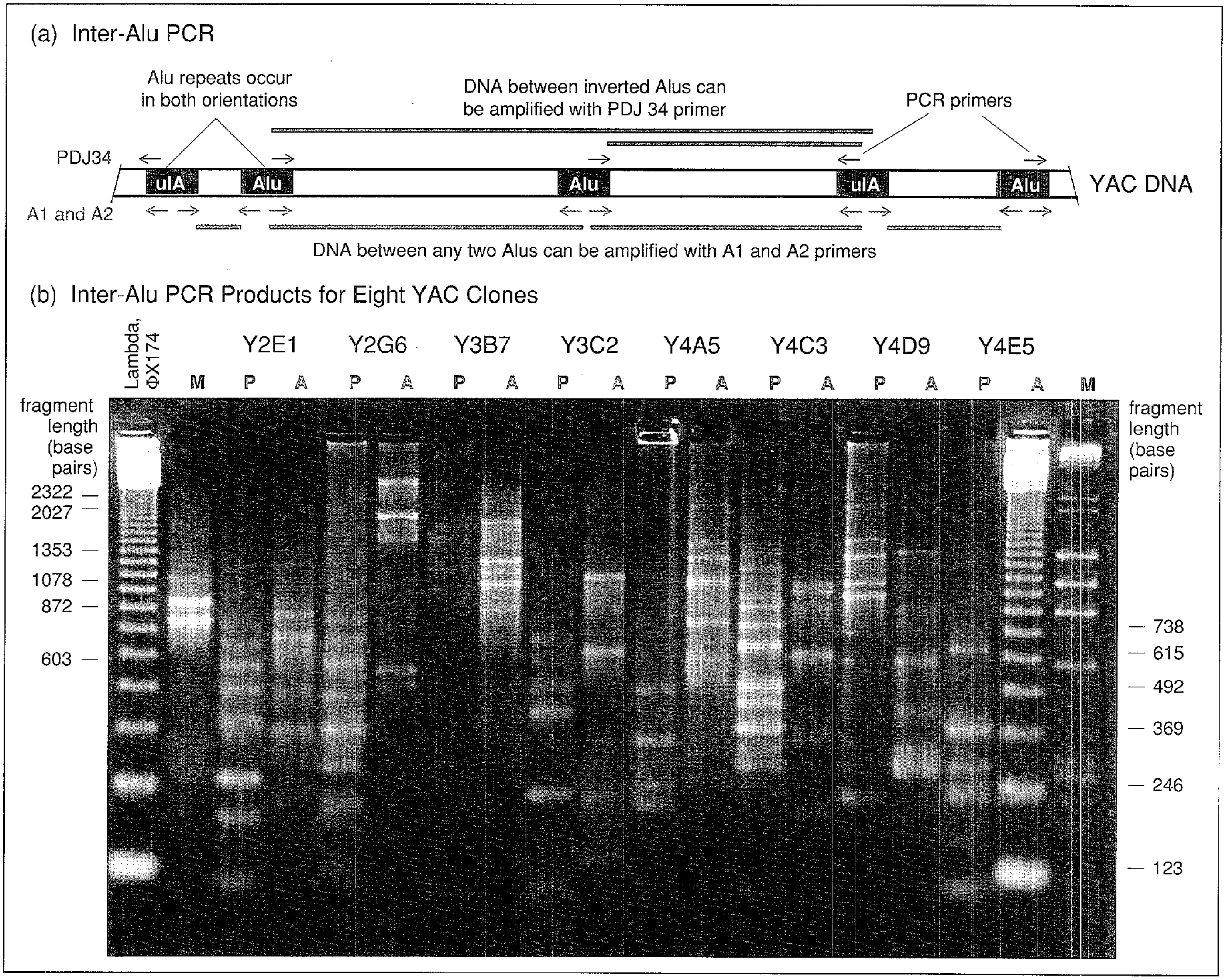

(a) Inter-Alu PCR
Alu repeats occur in both orientations
DNA between inverted Alus can be amplified with PDJ 34 primer
PCR primers
PDJ34
uIA
Alu
Alu
uIA
Alu
YAC DNA
A1 and A2
DNA between any two Alus can be amplified with A1 and A2 primers
(b) Inter-Alu PCR Products for Eight YAC Clones
Lambda, ΦX174
Y2E1
Y2G6
Y3B7
Y3C2
Y4A5
Y4C3
Y4D9
Y4E5
M P A P A P A P A P A P A P A P A M
fragment length (base pairs)
2322
2027
1353
1078
872
603
fragment length (base pairs)
738
615
492
369
246
123

Figure 8. Inter-Alu PCR Amplification of DNA from YAC Clones

(a) Primers whose sequences match the ends of the Alu repetitive sequence can be used in the polymerase chain reaction to amplify the DNA occurring between of Alu sequences in the human DNA insert of a YAC clone. Alu sequences are 300 base pairs long, occur on average at intervals of 3300 base pairs in the human genome and are absent from the yeast genome. As shown in the figure, Alu sequences can be oriented in opposite directions along the DNA in the genome. The figure shows two sets of Alu primers. Those marked PDJ34 match only one end of the Alu sequence and therefore can amplify DNA between Alu sequences of opposite orientation. Primers A1 and A2 match either end of the Alu sequence and therefore can amplify DNA between any two Alu sequences. The polymerase chain reaction can be used to amplify regions up to several thousand base pairs in length. (b) Agarose gel containing inter-Alu PCR products of YAC clones. Alu primers PDJ34 (from Pieter de Jong, LLNL) or A1 and A2 (from Michael Scicillano, M.D., Anderson Hospital) were used in the PCR to amplify human DNA from eight different YAC clones and the amplified products were separated by electrophoresis on eight lanes of the gel shown in the figure. The first two and last lanes contain fragments of known lengths and are used to calibrate the lengths of the PCR products. Inter-Alu PCR products range in size from 100 base pairs to greater than 2500 base pairs. Each of the YACs shown yielded from 5 to 15 such PCR products.