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.

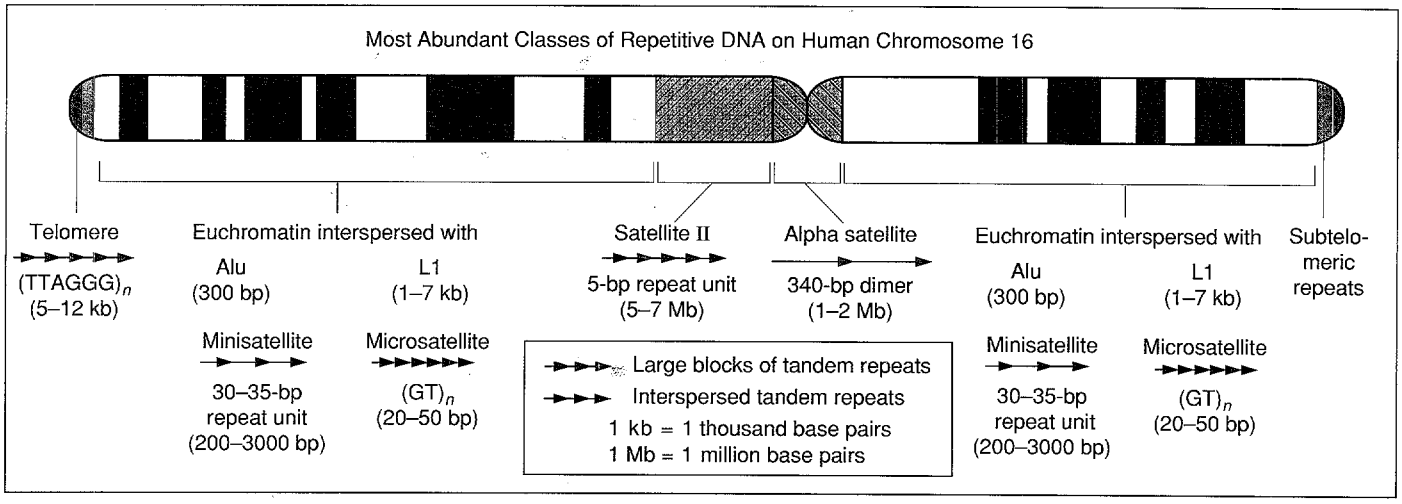
Telomere Repeat: The tandemly repeating unit TTAGGG located at the very ends of the linear DNA molecules in human and vertebrate chromosomes. The telomere repeat $(TTAGGG)_n$ extends for 5000 to 12,000 base pairs and has a structure different from that of normal DNA. A special enzyme called telomerase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