Physical Mapping

a one-dimensional jigsaw puzzle

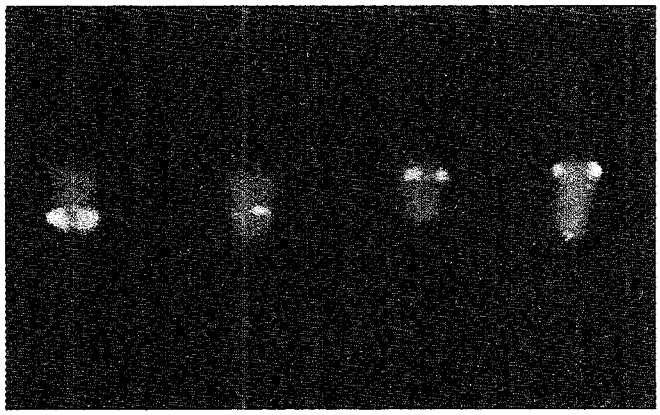
The human genome consists of forty-six double-stranded DNA molecules. Each molecule is made up, on average, of 130 million base pairs strung in a linear order between two sugar-phosphate backbones, and each is wound around proteins to form a chromosome. In order to study genes and other interesting regions of the genome at the molecular level, standard practice is to isolate the DNA and break up the long molecules into many fragments. We then make many identical copies of each fragment by cloning and pick out the clones of interest. Almost all methods for analyzing DNA at the molecular level require many copies of the fragment of interest. Therefore, cloning is essential for procedures such as finding the positions of restriction-enzyme cutting sites, determining the sequence of nucleotide bases in a particular DNA fragment, and identifying polymorphic DNA markers. However, in fragmenting the DNA molecules prior to cloning, we lose all information about the physical locations of fragments along the genome itself.

Problem: How do we find the chromosomal positions of known genes, polymorphic markers, and other cloned portions of the human genome?

Low-Resolution Physical Mapping by In-Situ Hybridization

In contrast to a linkage map, which specifies statistical distances between variable DNA markers and genes in terms of recombination fractions (see "Classical Linkage Mapping"), a physical map specifies physical distances between landmarks on the DNA molecule of each chromosome.

In-Situ Hybridization on Human Chromosome 21



Four DNA probes labeled with a fluorescent dye produce positive hybridization signals at four locations along chromosome 21. Because metaphase chromosomes are made up of two nearly identical sister chromatids, each probe produces a pair of signals.

One standard low-resolution method for finding the physical position of a cloned fragment is in-situ hybridization on metaphase chromosomes. We first find a segment within the cloned region whose base sequence occurs nowhere else in the genome. We then synthesize many copies of a single strand of that unique segment and label each copy with a fluorescent tag to make it useful as a DNA probe. A solution containing the DNA probe is then applied to a spread of chromosomes that have been arrested at metaphase and fixed to a microscope slide. (Metaphase is the phase of cell division during which chromosomes have condensed to form the wormlike shapes easily visible under a light microscope.) Under appropriate conditions the probe binds, or hybridizes, only to the chromosomal DNA with a base sequence exactly complementary to that of the probe (see "Hybridization" in "Understanding Inheritance"). The position on a metaphase chromosome where the probe has hybridized is imaged with a fluorescence microscope as a bright spot. Because DNA molecules are wound very tightly during metaphase, the resolution achieved with in-situ hybridization is low, about 3 million base pairs. In other words, the hybridization signals from two probes less than 3 million base pairs apart will overlap one another and cannot be resolved into two distinct spots. In-situ hybridization using

four cloned inserts as probes produced the bright spots on the metaphase chromosomes in the micrograph shown on the page opposite.

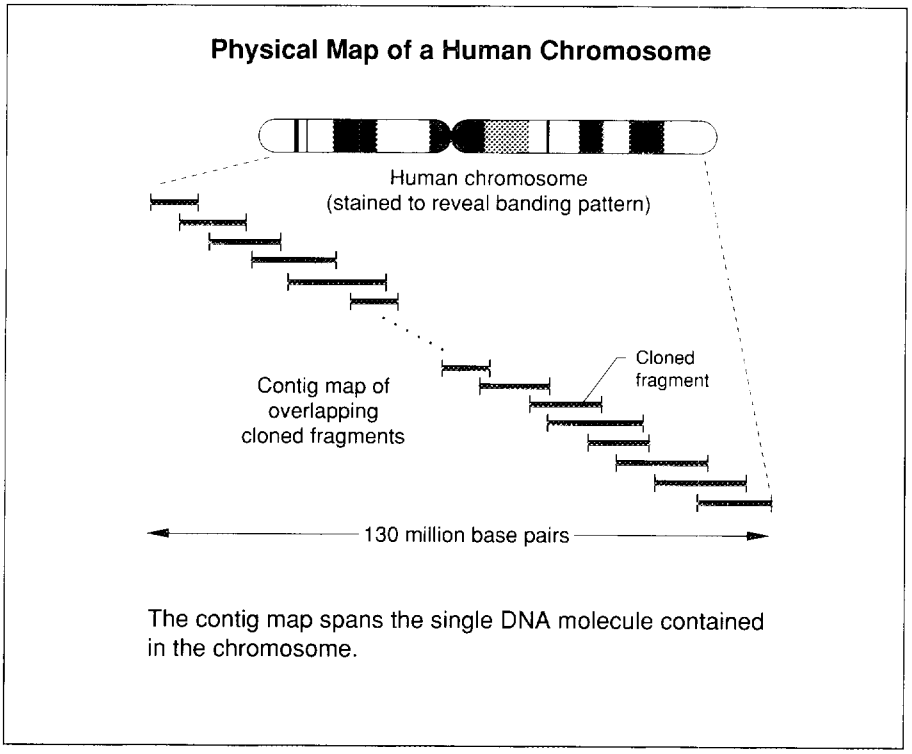
High-Resolution Physical Mapping by Construction of Contig Maps of Overlapping Clones

To determine the positions of genomic landmarks with much greater resolution, we can replace the chromosomes themselves with twenty-four contig maps, one for each of our twenty-two homologous chromosome pairs and one for each of our two sex chromosomes. A contig map is a set of contiguous overlapping cloned fragments that have been positioned relative to one another. In a complete contig map for a human chromosome, the cloned fragments would include all the DNA present in the chromosome and follow the same order found on the DNA molecule of the chromosome. As in any physical map, distances are measured in base pairs.

Using these contig maps, we can localize any cloned fragment or other DNA probe, again by hybridization, to a much smaller portion of the genome, namely to one of the cloned fragments in one of the maps. Moreover, we can determine the position of any DNA probe relative to all other landmarks that have been similarly localized. Once contig maps are constructed, the entire genome will be available as cloned fragments, and we will be able to use these clones to analyze any region down to the level of its base sequence.

Example: The figure at right is a schematic of a contig map for one chromosome. Right now, the top priority of the Human Genome Project is to construct a contig map for each of the twenty-four different chromosomes in the human genome. Those maps, when integrated with the corresponding genetic-linkage maps, will provide a means of finding the segments of DNA that contain disease genes (see "Modern Linkage Mapping"). The clones that make up the map also provide the material needed to sequence the human genome.

Many different strategies are being developed to make contig maps of human chromosomes. (Details of the Los Alamos effort to map a human chromosome are presented in "The Mapping of Chromosome 16.") Here we introduce the basic principles of contig-map construction.



Question: How do we obtain the clones that compose the contig maps?

Answer: We prepare a collection, or library, of cloned human DNA fragments in a manner such that (1) essentially all parts of the genome are probably present in the library and (2) the human DNA fragments in the clones overlap one another. Overlaps among the cloned fragments are essential because they allow us to reconstruct the order in which the fragments appear along the genome.

Example: The figure illustrates the steps in preparing a library of cloned DNA fragments. We start by isolating the DNA from many human cells. Then we break up the DNA into a large set of overlapping fragments by partial digestion of the DNA with a restriction enzyme. A restriction enzyme digests a DNA molecule by recognizing and cleaving the molecule at every occurrence of a particular short sequence usually four to eight base pairs long. Such a site is called a restriction site and is marked on the figure by a dot. Since complete digestion would yield nonoverlapping fragments (every copy of a particular DNA molecule would be cleaved at the same places), we interrupt the digestion process before it reaches completion, thereby leaving many restriction sites intact at random locations along each molecule. (In the figure, cleavage is indicated by a vertical line through the restriction site.) Such partial digestion ensures that each resulting fragment will overlap other fragments in the set.

Next, each of these fragments is joined to a cloning vector to form a recombinant DNA molecule. A cloning vector is a small DNA molecule that, after entering a host organism (such as yeast or bacteria), is replicated by the cellular machinery of the host organism. The cloning vector shown here is a small circular DNA molecule that has been engineered to include a single cutting site for the restriction enzyme chosen to digest the sample of human DNA. Copies of the cloning vectors are cut at that site and are mixed with the human DNA fragments, and the enzyme DNA ligase is added to the mixture. The "sticky ends" of a cloning vector (which are formed by restriction-enzyme cleavage) bind to the "sticky ends" of a human DNA fragment, and the ligase catalyzes the chemical union of the sugar-phosphate backbones of the two DNAs into a recombinant DNA molecule. We then expose a population of the host organism to the recombinant DNA molecules, and, if we are lucky, each recombinant DNA molecule enters a host organism and is there replicated as the host replicates. Each host colony containing clones of a particular fragment is individually plucked and stored in a well of a 96-well microtiter dish where the cells can be grown up again and again. This library of clones provides a renewable supply of all the fragments that have survived the cloning process.

To create a contig map of a single human chromosome, many groups are starting with a chromosome-specific library of cloned fragments constructed by starting with many copies of a particular chromosome. Chromosome-specific libraries are being made by the National Laboratory Gene Library Project at Los Alamos and Livermore and are available to research groups throughout the world (see "Libraries from Flow-sorted Chromosomes").

The cloned fragments in a DNA library are "anonymous"; that is, we know nothing about them except their approximate length, which is determined by the length of the DNA insert that can be successfully incorporated into the cloning vector we have chosen. Until recently cosmids were the cloning vectors most often used for map construction. Cosmids reproduce in the bacterial host *E. coli*, and they accept DNA inserts ranging from about 25,000 to 45,000 base pairs in length. Therefore about 4000 cosmid clones could accommodate all the DNA in an average human chromosome. However, to achieve the overlaps among cloned fragments required in the construction of a contig map and to better assure that all the chromosomal DNA is represented in the clone library, the usual practice is to construct a library with up to ten times that number of cosmid clones.

Question: How do we position the cloned DNA fragments along the DNA molecules in the genome?

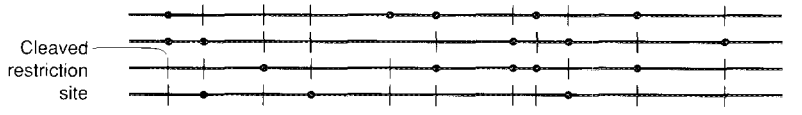
Answer: Positioning cloned DNA fragments is analogous to solving a one-dimensional jigsaw puzzle, but rather than looking for interlocking pieces, we look for detectable overlaps between clones, that is, for clones that have a unique stretch of human DNA in common. Because the number of pieces in the puzzle is so large, we need a rapid method for detecting overlaps between pairs of clones. If we could sequence each clone, we could identify overlaps unambiguously, provided the overlapping region is not a sequence that repeats elsewhere in the genome. However, given the current state of sequencing technology, that approach is totally impractical.

A practical and successful probabilistic method for detecting overlaps is to make a "fingerprint" of each clone (more precisely, of the human DNA insert within each clone) and compare the

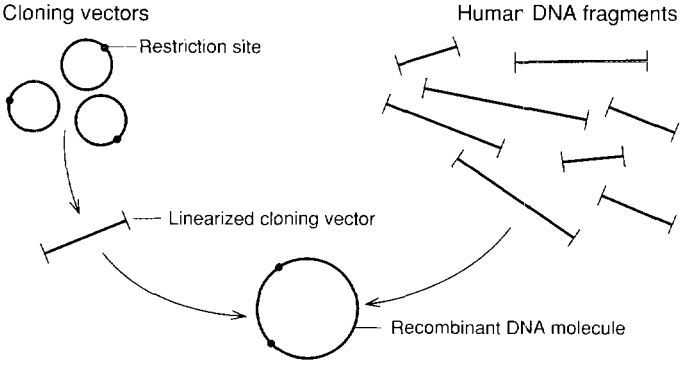
Construction of a Library of Cloned DNA Fragments

Step 1: (a) Isolate many copies of the human DNA molecule to be mapped.

(b) Partially digest the molecules with a restriction enzyme to create overlapping fragments.

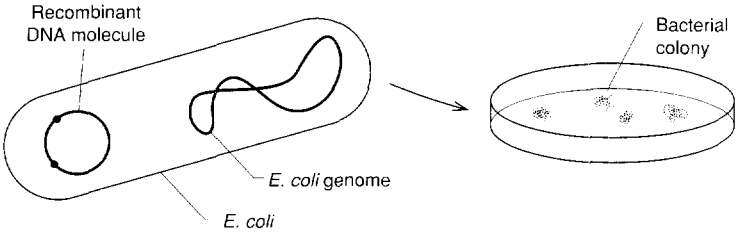


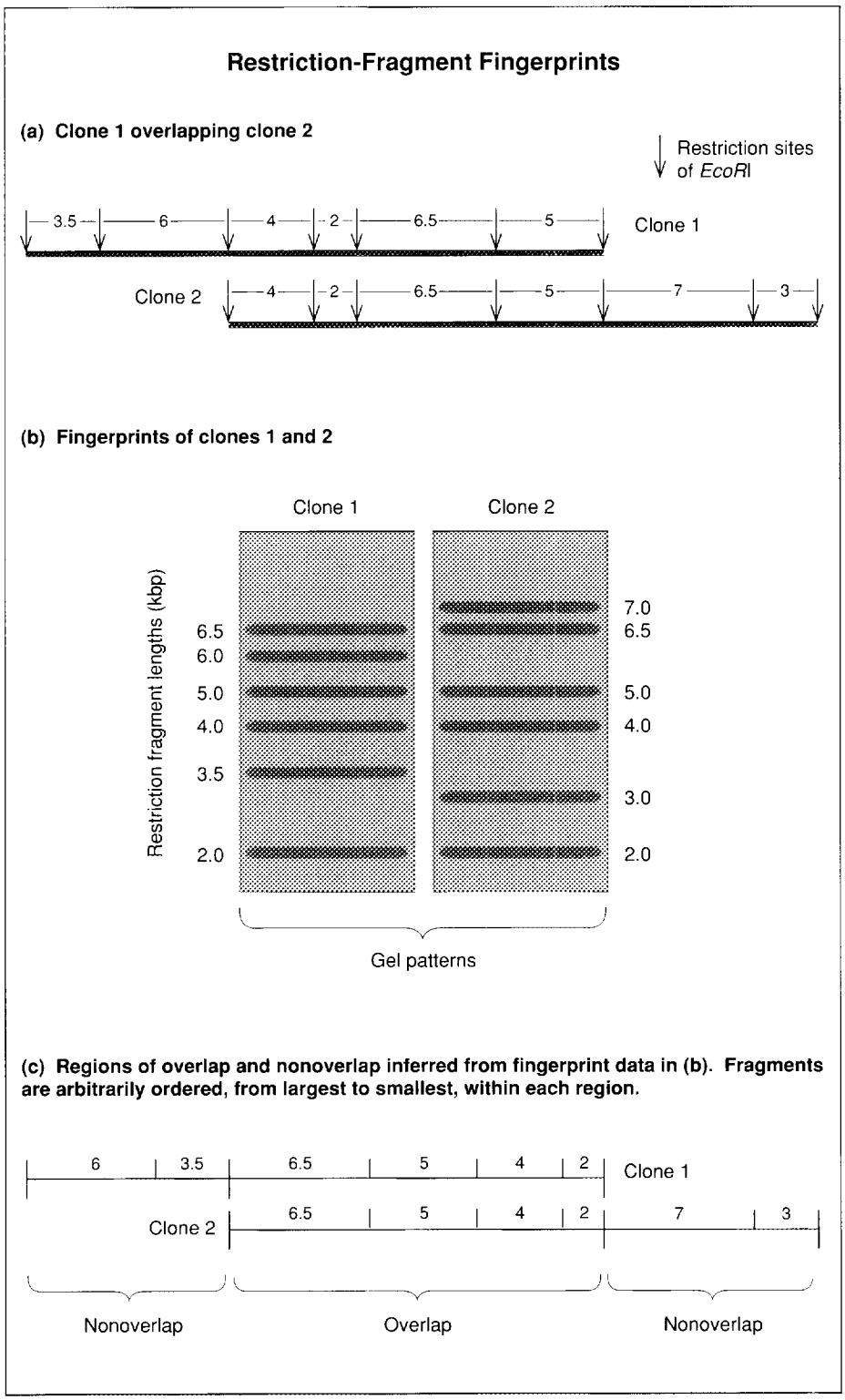
Step 2: (a) Linearize the circular cloning vectors with the restriction enzyme used in step 1b.



(b) Ligate cloning vectors and human DNA fragments to create recombinant DNA molecules.

Step 3: Facilitate the entry of recombinant DNA molecules into host cells, here the bacterium *E. coli*, and grow each host cell into an isolated colony, thereby producing many identical copies of that recombinant DNA molecule.





fingerprints. The simplest fingerprint of a cloned fragment is the one obtained by completely digesting about 10^{10} copies of the clone with a restriction enzyme and then determining the lengths of the resulting restriction fragments by gel electrophoresis. The restriction-fragment lengths determined from the gel constitute the restriction-fragment fingerprint of the clone.

Suppose we obtain restriction-fragment fingerprints of our clones by using the restriction enzyme *Eco*RI, which can cut DNA at every occurrence of the six-base-pair sequence GAATTC. Within a random sequence of the four DNA bases, any six-base-pair sequence occurs, on average, every 4^6, or about 4000, base pairs. Therefore the average length of the restriction fragments produced by *Eco*RI from a random sequence of the DNA bases is about 4000 base pairs. Now the sequence of bases in the human genome is not random, but nonetheless, the average length of the restriction fragments in the *Eco*RI fingerprints of a set of clones is about 4000 base pairs. Thus we expect that the human DNA inserts in two cosmid clones, each of which are, say, about 30,000 base pairs long, will have at least one restriction fragment in common if they overlap by more than about 15 percent.

Example: To illustrate the information content of fingerprints made by using the restriction enzyme *Eco*RI, consider two clones that are known to overlap as shown in part (a) of the figure. The cleavage sites for *Eco*RI are marked by arrows, and the distances between restriction sites are given in thousands of base pairs (kbp). Part (b) shows the restriction-fragment fingerprints obtained by completely digesting many copies of each clone with *Eco*RI. After several hours of electrophoresis, the restriction fragments of

each clone have separated into distinct bands, each band consisting of all the restriction fragments with a particular length. (The bands are made visible by staining, and each gel is calibrated with fragments of known lengths.)

The region of overlap between the two clones shown in the figure yields four restriction fragments with lengths of 4, 2, 6.5, and 5 kbp. Thus the fingerprints of the two clones have four bands in common at the gel positions corresponding to those lengths. Suppose these two fingerprints were the only information we had about the two clones shown in the figure. We might suspect that the clones overlap one another and that the overlap region included four restriction fragments with lengths of 2, 4, 5, and 6.5 kbp. We might then partition the restriction fragments into a region of overlap and two regions of nonoverlap as shown in part (c) of the figure. Note that we would have no way to impose any further ordering on the restriction fragments present in the fingerprint. Shown in (d) is a photograph of actual fingerprint data.

Question: Can we infer that two clones overlap solely on the basis of their restriction-fragment fingerprints?

Answer: Since a restriction-fragment fingerprint is, in essence, just a list of restriction-fragment lengths, it gives us no information about the order of the fragments within each clone. Also, we can't tell whether the restriction fragments of the same length in two different fingerprints are copies of the same fragment. So the fact that the fingerprints of two clones have one or more restriction-fragment lengths in common does not provide unambiguous evidence that the two clones overlap. On the other hand, by taking into account statistical properties of restriction-fragment lengths, we can estimate the likelihood of overlap given the data. David Torney of Los Alamos has developed a rigorous formulation of the likelihood calculation that takes into account the distribution of the distances between cleavage sites in the genome (the distribution of *Eco*RI cleavage sites appears to be a Poisson distribution with an average spacing of 4000 base pairs), the errors in the measurement of restriction-fragment lengths (about 1 percent), and all possible ways in which the two clones might overlap. Since the declaration of a false overlap would lead to the merging of pieces of the map that are not contiguous on the genome and since such mistakes are very time-consuming to correct, a conservative approach is to declare an overlap only if the likelihood of overlap is 90 percent or greater. Given the simple restriction-fragment fingerprints shown on the page opposite, two clones must overlap by about 50 percent to yield such high likelihoods of overlap. Thus small overlaps are typically not detected with this conservative approach. As described in "The Mapping of Chromosome 16," the Los Alamos mapping group has devised a fingerprint that includes information about the presence of repetitive DNA sequences on the restriction fragments in each fingerprint. That additional information facilitates the detection of much smaller overlaps and therefore requires the fingerprinting of fewer clones to complete the contig map.

Question: How are pairs of clones with a high likelihood of overlap assembled into contigs, sets of contiguous overlapping clones?

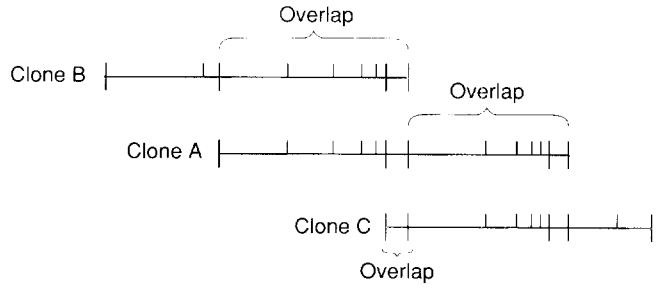
Answer: Given the uncertainties in fingerprint data, assembling pairs of overlapping clones into contigs from those data alone is a difficult computational problem. The

standard procedure is to find pairs of clones, link those pairs into groups, and then attempt to order all the restriction fragments within each group of clones in a self-consistent manner. The method is essentially an incremental approach. As each new clone is added to a contig, one tries to retain as much of the existing construction as possible even in the face of contradictory data.

A significant departure from the incremental procedure has recently been developed at Los Alamos. Map construction is treated as an optimization problem in which all available data are taken into account rather than only the data yielding high overlap probabilities. A description of this global approach to map construction is discussed in "Computation and the Human Genome Project." Here we illustrate the more standard procedure.

Example 1: Suppose that the fingerprints of clones A, B, and C reveal that clones A and B have five fragment lengths in common, A and C have six fragment lengths in common, and B and C have one fragment length in common. Furthermore, we have calculated from those data that the likelihood of A and B overlapping is 90 percent, of A and C overlapping is 95 percent, and of B and C overlapping is 10 percent. We would then assemble the three clones into a contig as shown in the figure, where some restriction fragments are placed in regions of overlap and the

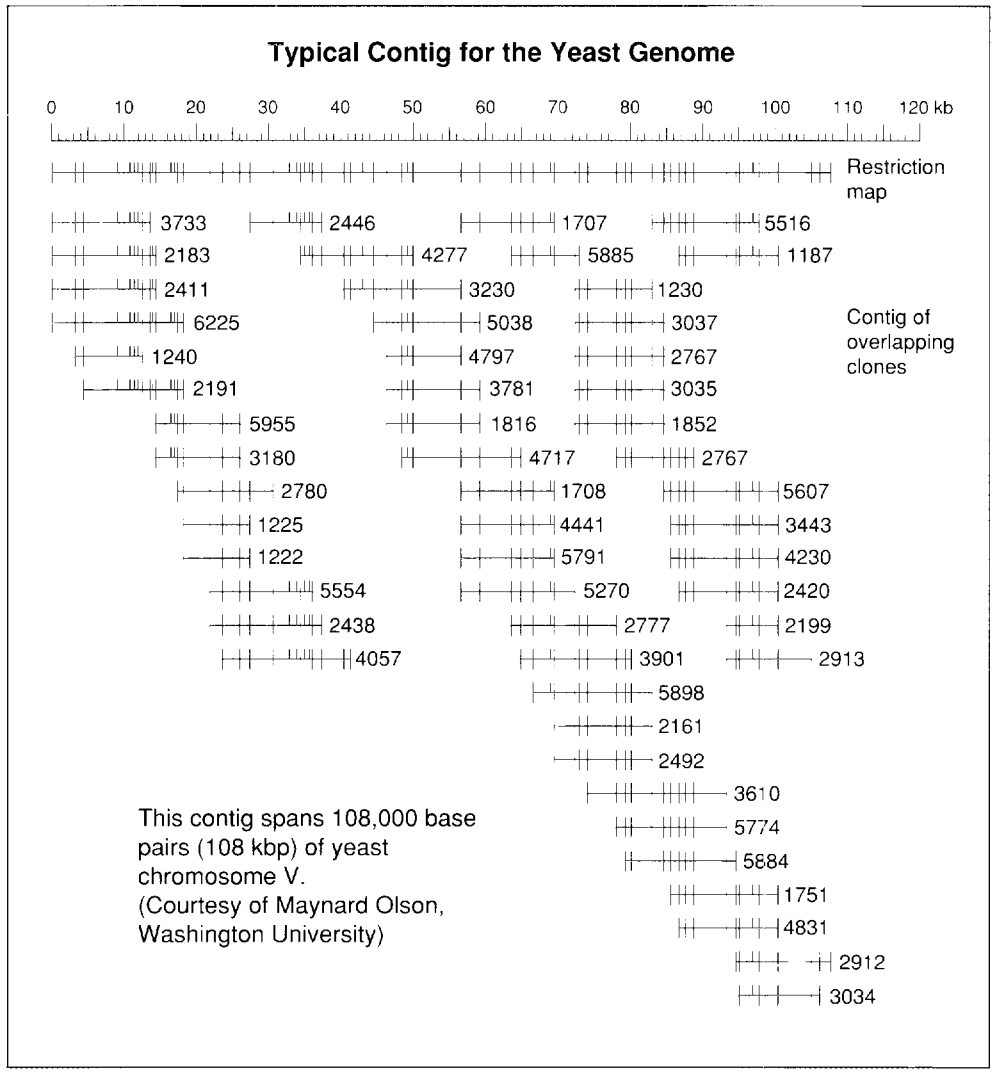
Assembly of a Contig



Likelihood analysis of fingerprint data suggests that clone A overlaps clone B and clone C and that clone B and clone C do not overlap. However, clone B and clone C do share one restriction fragment and that fragment can be placed in the overlap between clone B and C with no loss of consistency. Preliminary assignments of restriction fragments to overlap or non-overlap regions might be altered as more clones are added to the contig.

remaining ones are placed in the regions of nonoverlap. As we add other clones to the contig, we might have to revise the partitioning of the fragments into overlapping and nonoverlapping regions to construct a consistent ordering for the entire contig. Because of the uncertainties in fragment lengths and the possibility that fragments of equal length are not necessarily the *same* fragment, complicated computer algorithms are necessary to determine the most likely order of the clones in a contig. When the number of clones in a contig is much larger than the number required to span the region covered by the contig, we can order many of the restriction fragments that appear in each fingerprint and thereby help to avoid some false overlaps.

Example 2: Shown at right is a contig assembled on the basis of restriction-fragment fingerprints. The contig spans about 100,000 base pairs. Also shown is a restriction map deduced from the contig. The restriction map shows the order of and distances between restriction sites in thousands of base pairs or in kbp. The exact positions of some restriction sites (marked by the longer vertical lines that extend through the cloned fragments) have been determined by the fact that each lies at the end of one of the clones in the contig and therefore separates a region of overlap between two clones from a region of nonoverlap. Other restriction sites (marked by the shorter vertical lines) have been localized to a single overlap region but cannot be ordered further. Such sites have been arbitrarily located left to right on the contig in order of decreasing inter-site distance. This contig is representative of those used in constructing the recently completed physical map of the genome of baker's yeast (*Saccharomyces cerevisiae*). That map is, on average, eight clones deep. That is, any region is present in, on average, eight clones. Such great redundancy provided information about the order of a large fraction of the restriction sites and greatly reduced the chance of a false overlap.



Question: Do the disconnected contigs assembled by fingerprinting randomly selected clones steadily increase in length until they become connected?

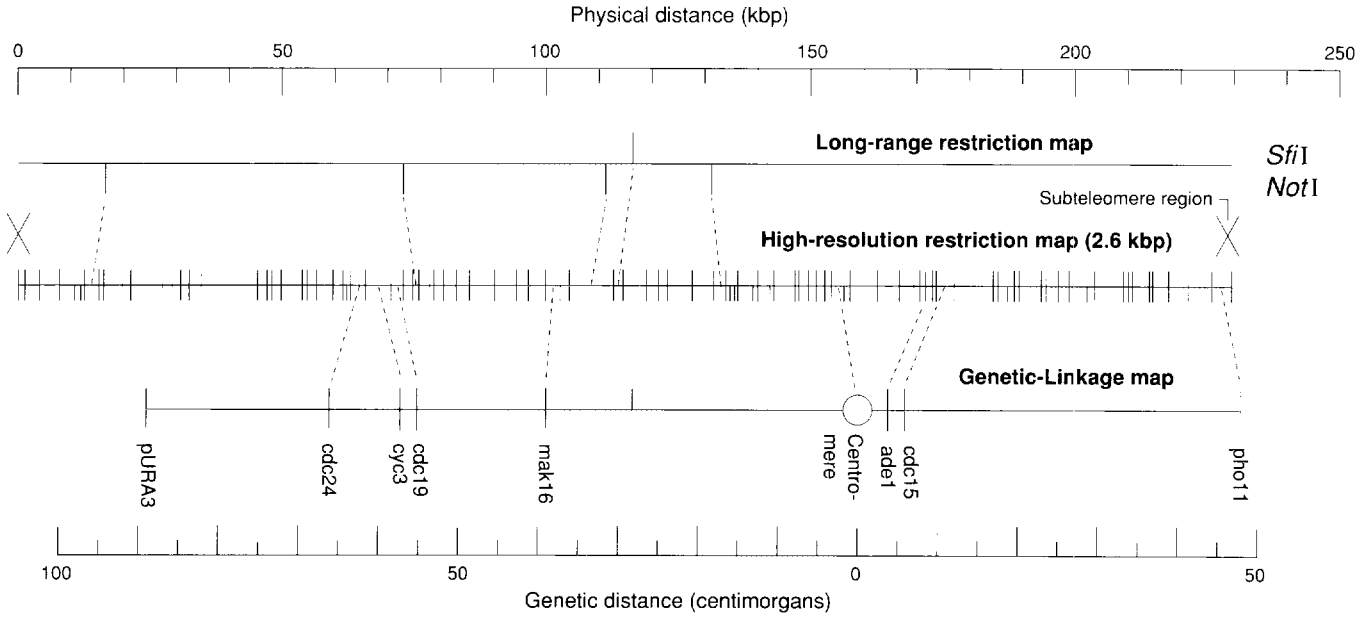
Answer: No. In a random fingerprinting strategy, both the numbers and sizes of the contigs grow fairly rapidly at first, but the rates of growth decrease after the existing contigs cover about two-thirds of the region to be mapped. The decrease in growth rate is due to the increasing probability that a randomly selected clone falls within a region for which a contig has already been assembled. Contig growth is also limited because small overlaps typically go undetected and some portions of the region being mapped may not have survived the cloning process. In fact, contigs assembled from cosmid clones typically stop growing after reaching lengths of 100 kbp.

Question: How do we order disconnected contigs along the chromosome and how do we check their accuracy?

Answer: Many types of lower-resolution maps can be used to position the contigs along a chromosome and to check that all the clones in a contig come from approximately the same region of the genome.

Example: The contigs constructed for yeast chromosomes, which had an average length of 100 kbp, were ordered relative to a high-density genetic-linkage map containing 400 markers spaced at an average physical distance of 30,000 base pairs. To check the integrity of each contig, the clones that form it were hybridized to very

Complete High-Resolution Restriction Map of Yeast Chromosome I



The high-resolution restriction map for yeast chromosome I was derived from a completed contig map of the chromosome. The Xs mark the beginning of the subtelomeric regions which are known to lie a few thousand base pairs away from the telomeres (ends) of the chromosome. Restriction sites for the thirteen-base cutter *Sfi*I and the eight-base cutter *Not*I and markers on the linkage map of chromosome I are localized to particular restriction fragments on the high-resolution restriction map. (Courtesy of Maynard Olson, Washington University)

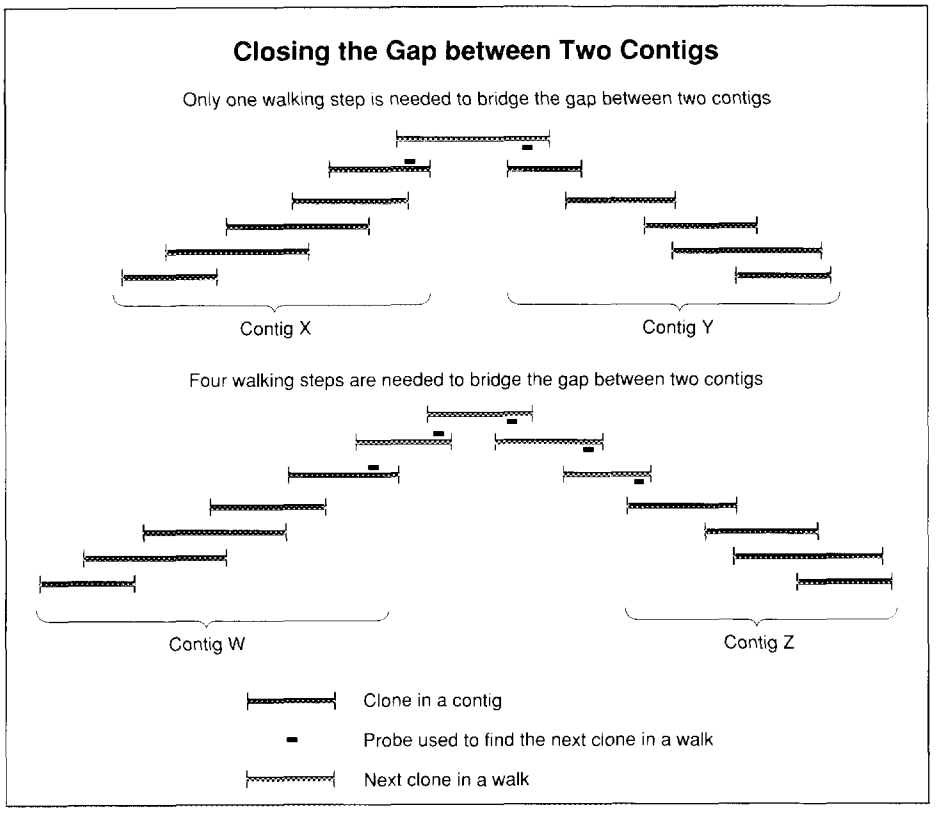
long (over 100,000 base pairs) restriction fragments of DNA that had been separated by pulsed-field gel electrophoresis. If the clones assigned to a contig do in fact come from a single region of the genome, it is likely that all of them will hybridize to a single large fragment on the gel.

The figure shows the high-resolution restriction map deduced from the completed contig map of yeast chromosome I. Also shown is the alignment of the restriction map with two other maps: (1) the genetic-linkage map and (2) a long-range restriction map showing the distances between the eight-base restriction sites of the enzyme *Not*I and the thirteen-base restriction site of *Sfi*I. (The latter map was constructed using pulsed-field gel electrophoresis.) Markers on the genetic-linkage map and restriction sites on the long-range restriction map have been localized to particular restriction fragments on the contig map. Those correspondences are indicated by dotted lines.

The contigs being assembled for human chromosomes are being checked by a variety of techniques including in-situ hybridization and hybridization to the DNA from hybrid cells containing increasingly longer portions of the chromosome being mapped (see "The Mapping of Chromosome 16").

Question: After the contigs are ordered and checked for accuracy, how do we fill in the gaps between the contigs?

Answer: As mentioned earlier, the fingerprinting of randomly selected clones is not an efficient way to fill in the gaps between contigs after the existing contigs cover a large fraction of the region being mapped. Instead it is time to employ a directed strategy. One directed strategy involves identifying unique regions within the clones at the ends of a contig and using those regions as probes to pick out other clones that will extend the contig. If the contigs cover a very large fraction (95 percent) of the region being mapped, a single probe from the end of a clone may identify a new clone that spans the distance between two existing contigs and thus merges them into one. If not, then one must continue stepwise by creating an end probe from each added clone and screening the library of clones to find the next clone that extends the contig a bit farther. This procedure is called walking, and it is extremely time-consuming. Nevertheless, it has been used successfully to complete physical maps of the *E. coli* and yeast genomes. Those genomes are relatively small (containing 5 million base pairs and 13 million base pairs, respectively), and the gaps between contigs were small before walking was attempted.



Example: The figure illustrates the merging of two contigs by either a single clone or several walking steps.

CAVEAT: A physical map is a very difficult puzzle to complete. As mentioned in the round table (see pages 108–109 in "Mapping the Genome"), the generic clone-to-fingerprint-to-contig cycle, which is amenable to automation and improved data-analysis algorithms, is only a small fraction of the work. The rest of the work required to close gaps between contigs and to track down inconsistencies such as the branching of one contig into two or more contigs involves many standard molecular-biology procedures, which, in the case of the human genome, must be carried out on an unprecedented scale. It is estimated that the completion of the yeast map took about 20 person-years of work, and the mapping of *each* human chromosome will take about 100 person-years. Further, mapping of human chromosomes presents some new challenges.

- An average human chromosome is ten times the size of the yeast genome, and the increased size calls for more efficient mapping strategies, such as working with larger clones.
- Unlike the genomes of yeast and *E. coli*, human DNA contains repetitive elements that require a new fingerprinting strategy to avoid inferring overlaps between clones containing long repetitive stretches of DNA near their ends.
- Experience has shown that regions containing repetitive sequences are often lost in the cloning process. Consequently, parts of the puzzle of each human chromosome may be missing, in which case completion of the map will require specialized techniques.

These challenges are being met in a variety of ways including the use of YAC cloning vectors, which accept DNA inserts eight to ten times larger than the inserts accepted by cosmids, and the use of STS markers, which, unlike restriction-fragment fingerprints, identify unique landmarks on the map and therefore eliminate the need for complicated probabilistic analyses to infer overlap between two clones. ■