Imagine ancient explorers finding a giant island to explore. They initially can only touch the periphery as they sail around and try to imagine what lies within. With crude instruments, they make estimates of the size of the island and map more precisely what little they can reach. As more adventurers follow, more of the island is uncovered and fairly accurate maps of villages and roads are developed, but the precise map of the whole island is beyond their reach. When the population of settlers is large enough, the roads and cities cover the island, but the large expanses between are barely explored. It is not until decades later that new technologies are developed which allow rapid imaging and computer analysis of the terrain of the entire island and the complete map is available.

This is the same story as the Human Genome Project. When scientists first began to sequence, they worked on single genes which encompassed only a few thousand base pairs. While they could map the individual genes on the chromosomes and generate a series of markers on the chromosomes, the majority of the sequence was undetermined. As more scientists sequenced more genes, the map began to fill in, but there were large regions with no identified genes and therefore no sequence data.

Within the last decade, technology has been developed which allows the rapid sequencing of large regions of chromosomes. Scientists do not need to have marker genes to guide them, computers will analyze thousands of different sequence pieces and assemble them together into the final complete map. How this process is done is the subject of this chapter.

While the name "Genome Project" gives the image of large pieces of DNA, perhaps whole chromosomes being manipulated and sequenced in whole, the reality is that even the best sequencing machines can only read several hundred base pairs at a time with a high level of accuracy. So the key to mapping the entire genome is to break it down into little pieces, analyze each little piece and then reassemble the small parts of sequence information back into the large contiguous sequence which is called a 'contig'. And although the emphasis on the Genome Project is sequence analysis, it is necessary to integrate locating transcribed regions and using them as markers to the whole project. Only with these signposts can the accuracy of the sequence be assured.

When the US Geological Survey set out to map the topography of the entire United States, they first set and marked a series of reference points. Brass cylinders were set at each point with the longitude, latitude and altitude of that position engraved on top. Once the markers were set, surveyors could use these fix points to determine the location of the surrounding terrain. In the genome, the positions that stand out are the genes or the actively transcribed regions. Therefore, the first step in mapping the chromosome is to position as many transcripts as possible. This is done by Expression Sequence Tagging. To do this, random cDNAs are generated and a few hundred base pairs of sequence is determined. The cDNA is then mapped to its chromosomal location by in situ hybridization. This now fixes a few hundred base pairs of information, enough that it is very unlikely to be duplicated at another location by random events, to a specific chromosomal location. The advantage of using cDNA is that transcribed genes are usually in the class of unique DNA and not likely to be repetitive, so the few hundred base pairs of sequence is likely to be found only at the one chromosomal location. If a contig of random clones is later found to contain these few hundred base pairs, the computer can match it and instantly determine its chromosomal location.

EST's are also useful since they comprise the majority of the transcripts from a given organism. They therefore give a useful estimate of the number of active genes in the organism. The EST chromosome map will indicate if there are active or inactive regions of a chromosome, which will allow the sequencing project to be centered on the regions rich in transcripts and expressed genes.

The next step is to break up the DNA into pieces which can easily be fed into automated sequencing machines and yet maintain enough information that each piece can be differentiated from the thousands or millions of other pieces that will be generated by fragmenting an entire genome. It might seem that the most efficient method would be to cut a chromosome into precise pieces and start at one end and work straight through. However, it is actually important to fragment several copies of the genome into pieces with different breakpoints to generate segments with overlapping ends. If the DNA is fragmented using a system which cuts at the same places on every chromosome, such as complete digestion with a restriction endonuclease, there will be nothing at the end to indicate which is the next piece on the chromosome. The only marker available is sequence similarity which can only be used if there are random ends to give regions of overlap. The end result is that any region on the chromosome may be sequenced as part of several overlapping pieces. This may seem wasteful, but is useful since it increases the accuracy of the sequence determination. An automated sequencing machine can determine up to 500 bp of sequence in a single reaction, but the accuracy is lowest at the very beginning and last 20% of the read. Overlapping fragments increase the likelihood that most regions will be determined at least in part from the most accurate regions of sequencing runs. Secondary structure induced by hybridization of two parts of a single-stranded DNA molecule hybridizing to each other can alter the mobility of the products of a sequencing reaction. This can result in a 'compression' where products from the reactions for different nucleotides run at nearly the same mobility and make it difficult to determine their exact order. The point at which the secondary structure can form and the bases affected will differ depending on which strand of a piece of DNA is being used as the template. Therefore, compressions can often be resolved by aligning reads made in opposite direction. The random overlapping fragments are inserted in random direction relative to the sequencing primer binding site in the plasmid, which results in most parts of a chromosome being read off both strands which helps to eliminate artefactual base pair reads resulting from compressions.

It might seem that the size of the genomic fragments used to seqeunce would not matter, as long as they were large enough to contain significant data. However, full genomic sequencing is often done with two different sized collections. One group is centered around 10,000 bp in length while the other is 1-2,000 bp in size. The smaller clones generate more widespread sequence information. The larger clones confirm the EST scaffold by generating two sequences (each end) which are a large, but set distance apart. If there are potential alignment problems with the smaller fragments, for instance in regions with duplications, the larger fragment containing the region can be sequenced to align all the parts.

With powerful software to sort through sequence data, it has not always been necessary to set up a scaffold of EST sequence data before attempting to assemble the complete sequence. Using a method called 'shotgun sequencing', random short pieces of the genome are assembled and sequenced. Enough sequences are read to theoretically

read the entire genome multiple, usually 6-10 times, over. This ensures a high probability that every sequence will be present at least once to prevent any gaps. Computers then begin the process of finding the regions of overlap between the short pieces and assembling them into larger and larger units until the entire genome is covered. The multiple redundancy allows a high level of certainty in the final compilation. This method is much faster since it eliminates the multiple steps of EST cloning, sequencing and mapping and requires only small sized genomic fragments for sequencing. This method works very efficiently for bacteria and other simpler, smaller genomes. The difficulty lies in the analysis of complex genomes which often contain repeated region. The human genome contains tens of thousands of copies of

One popular misconception is that when a genome project is declared "complete" that every single base pair has been determined. Just as there are sites on Earth that cannot be approached for mapping without extraordinary effort, there are regions of the genomes which defy standard sequencing methods. These regions include highly