The Polymerase Chain Reaction and Sequence-tagged Sites

Norman A. Doggett

Polymerase Chain Reaction

The polymerase chain reaction (PCR) is an in vitro method for selectively amplifying, or synthesizing millions of copies of, a short region of a DNA molecule. The reaction is carried out enzymatically in a test tube and has been successfully applied to regions as small as 100 base pairs and as large as 6000 base pairs. In contrast, DNA cloning is a nonselective in vivo method for replicating DNA fragments within bacterial or yeast cells. Cloned fragments range in length from several hundred to a million base pairs. (See "DNA Libraries" for further discussion of DNA cloning.)

PCR is particularly important to the Human Genome Project as a tool for identifying unique landmarks on the physical maps of chromosomes. The PCR can be used to detect the presence of a particular DNA segment in a much larger DNA sample and to synthesize many copies of that segment for further use as a probe or as the starting material for DNA sequencing.

Figure 1 illustrates the polymerase chain reaction. The reaction mixture contains:

- A DNA sample containing the target sequence.
- Two single-stranded DNA primers (short sequences about 20 nucleotides long) that anneal, or bind by complementary base pairing, to opposite strands of DNA at sites at either end of the target sequence. Such short DNA sequences are called oligonucleotides and can be synthesized in a commercially available instrument.
- A heat-stable DNA polymerase, an enzyme that catalyzes the synthesis of a DNA strand complementary to the target sequence and can withstand high temperatures.
- Free deoxyribonucleoside triphosphates (dATP, dGTP, dCTP, and dTTP), precursors of the four different nucleotides that will extend the primer strands.
- A reaction buffer to facilitate primer annealing and optimize enzymatic function.

The polymerase chain reaction proceeds by repeated cycling of three temperatures:

- Phase 1: Heating to 95°C to denature the double-stranded DNA, that is, to break the hydrogen bonds holding the two complementary strands together. The resulting single strands serve as templates for DNA synthesis.
- Phase 2: Cooling to a temperature between 55°C and 65°C to allow each of the primers to anneal (or hybridize) to its complementary sequence at the 3′ end of one of the template strands.
- Phase 3: Heating to 72°C to facilitate optimal synthesis, or extension of the primer strand by the action of the DNA polymerase. The polymerase attaches at the 3′ end of the primer and follows along in the 3′-to-5′ direction of the template strand catalyzing the addition of nucleotides to the primer strand until it either falls off or reaches the end of the template strand (see "DNA Replication" in "Understanding Inheritance").

The figure shows the materials in the reaction mixture and the first three cycles of the reaction. The DNA synthesized in each cycle serves as a template in the next. Note that an exact duplicate of each strand of the target sequence is first created during cycle 2. Each subsequent cycle doubles the number of those strands so that after n cycles the reaction will contain approximately 2^n copies of each strand of the

Figure 1. The Polymerase Chain Reaction

Reaction mixture includes DNA sample: two single-stranded primers, each with a 20-base sequence identical to the 5' end of one strand of the target sequence; heat stable *Taq* polymerase; and deoxyribonucleocide triphosphates (dNTPs).

Phase 1
Denature unamplified DNA at 95°C to form single-stranded templates.

Phase 2
Anneal primers to template at about 60°C.

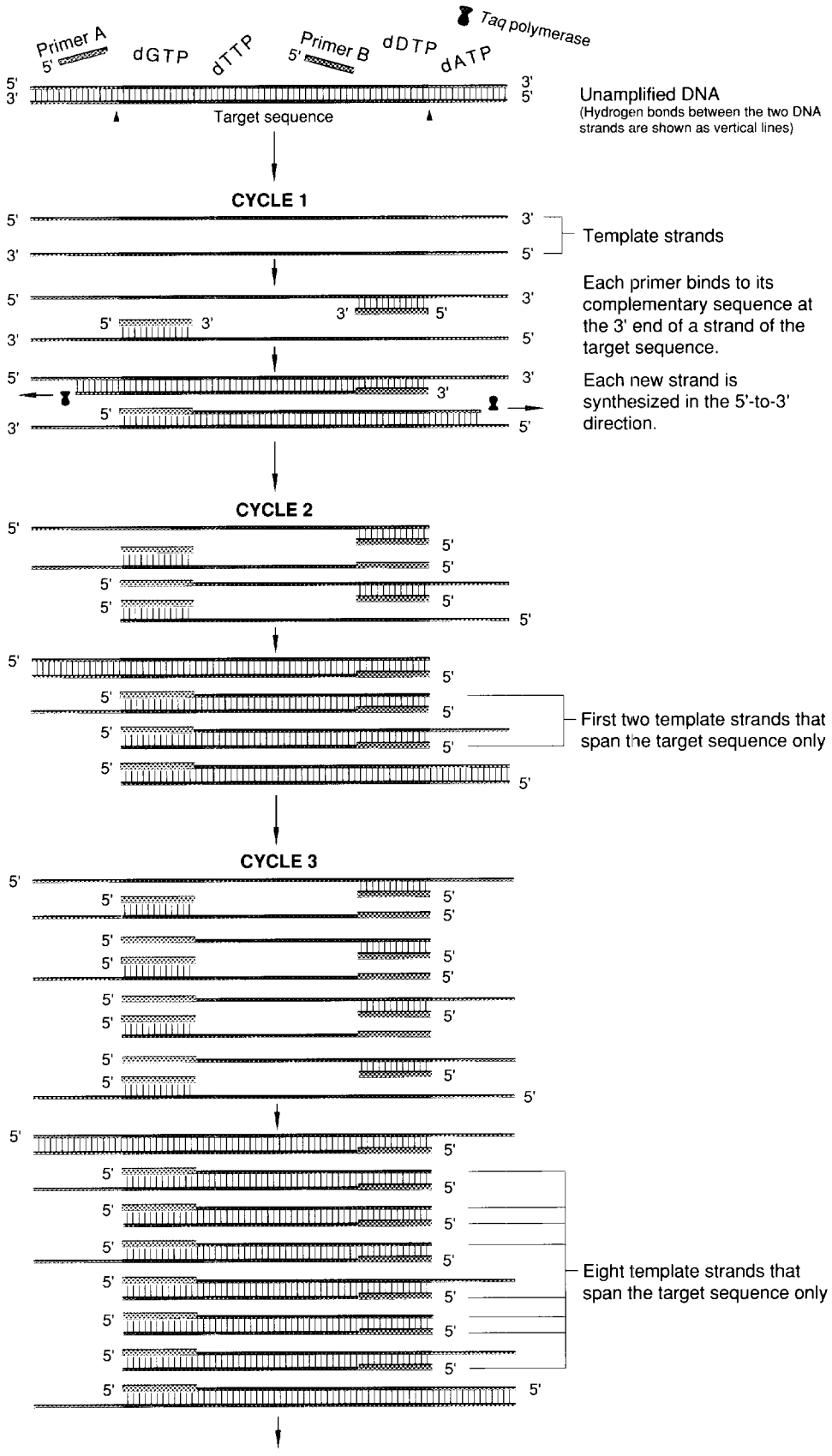
Phase 3
Synthesize new strands at 72°C.

Phases 1 and 2
Denature products of Cycle 1 and anneal primers to template strands.

Phase 3
Synthesize new strands.

Phases 1 and 2
Denature products of Cycle 2 and anneal primers to template strands.

Phase 3
Synthesize new strands.



Continue for 20 to 30 cycles to produce over 10^6 copies of target sequence.

target sequence. Typically the chain reaction is continued for 20 to 30 cycles in microprocessor-controlled temperature-cycling devices to create between roughly 1 million and 1 billion copies of the target sequence.

Taq polymerase, a heat-stable polymerase isolated from the bacterium *Thermus aquaticus* found in hot springs, is used in the reaction. The annealing temperature for the second phase of each cycle is chosen to be approximately 5°C below the temperature at which the primers no longer anneal to the target sequence. That so-called melting temperature varies depending upon the primer sequence. In particular because G-C base pairs (which have three hydrogen bonds) remain stable at higher temperatures than A-T base pairs (which have only two hydrogen bonds), primers containing mostly Gs and Cs have a higher melting temperature than those containing mostly As and Ts. The annealing temperature must be chosen carefully because if the temperature is too low, the primers will bind to sites whose sequence is not exactly complementary to the primer sequence resulting in the amplification of sequences other than and in addition to the target sequence. If the temperature is too high, the primers will not bind to the template strands and the reaction will fail.

Typically the initial DNA sample contains from 3,300 to 333,000 copies of the human genome (or 10 nanograms to 1 microgram of total genomic DNA). However, when working properly, the PCR will selectively amplify a unique target sequence contained in a single copy of the genome (6 picograms of DNA) isolated from a single cell. To evaluate the specificity of the reaction, that is, whether or not the reaction amplified a single target region, the reaction products are separated on a gel using electrophoresis. If a single region has been amplified, the gel will contain a single intense band containing the synthesized copies of the target sequence. The location of the band on the gel indicates the length of the amplified region. If more than one intense band appears on the gel, then more than one region of the genome was amplified by the reaction and the sequence of the primers appear more than once in the genome.

Sequence-tagged Sites

A sequence-tagged site (STS) is a short region along the genome (200 to 300 bases long) whose exact sequence is found nowhere else in the genome. The uniqueness of the sequence is established by demonstrating that it can be uniquely amplified by the PCR. The DNA sequence of an STS may contain repetitive elements, sequences that appear elsewhere in the genome, but as long as the sequences at both ends of the site are unique, we can synthesize unique DNA primers complementary to those ends, amplify the region using the PCR, and demonstrate the specificity of the reaction by gel electrophoresis of the amplified product (Figure 2).

Operationally, a sequence-tagged site is defined by the PCR used to perform the selective amplification of that site. The PCR is specified by the pair of DNA primers that bind to the ends of the site and the reaction conditions under which the PCR will amplify that particular site and no other in the genome.

STSs are useful because they define unique, detectable landmarks on the physical map of the human genome. One of the goals of the Human Genome Project is to find STS markers spaced roughly every 100,000 bases apart along the contig map

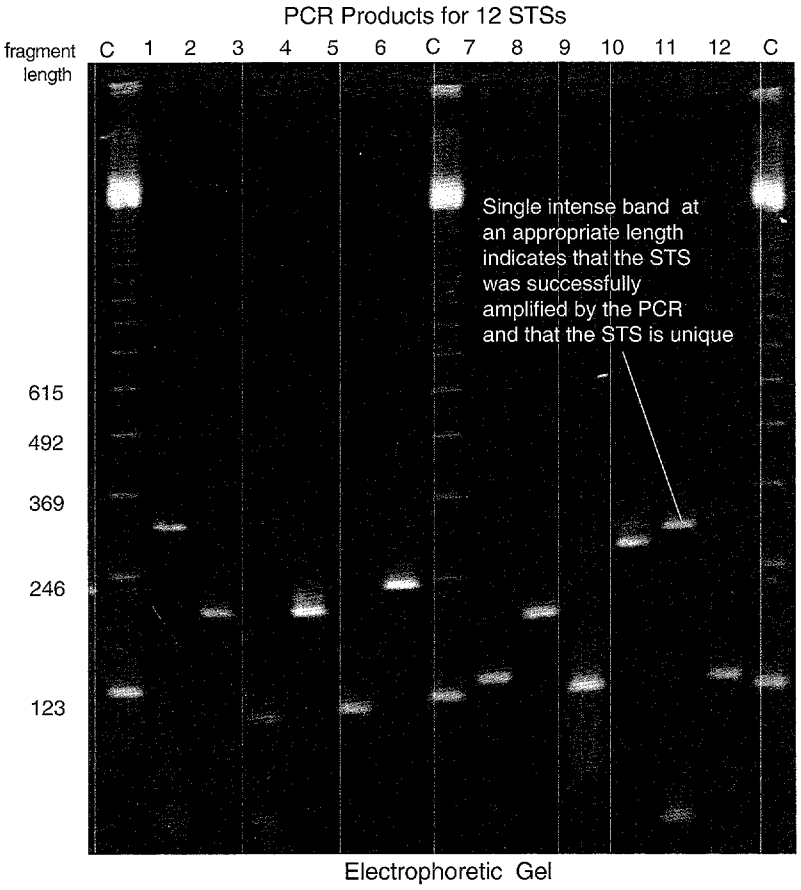
of each human chromosome (see "Physical Mapping—A One-dimensional Jigsaw Puzzle" for a description of contig maps). The information defining each site will be stored in a computer database such as GenBank. That stored information will include the PCR primers, reaction conditions, and product sizes as well as the DNA sequence of the site. Anyone who wishes to make copies of the marker would simply look up the STS in the database, synthesize the specified primers, and run the PCR under the specified conditions to amplify the STS from genomic DNA. As described below, copies of the STS can be used to screen a library of uncharacterized clones and identify a clone containing the marker. Therefore, a database of such landmarks will eliminate the need to store and distribute a permanent set of DNA clones or probes for the physical maps.

Figure 3 outlines the procedure for finding an STS marker. One begins by sequencing a 200- to 400-base region of a cloned DNA fragment. The rough sequence can be obtained from a single run of a DNA sequencing gel (see "DNA Sequencing"). The sequence is then examined to find two twenty-base regions separated by 100 to 300 base pairs that might serve as unique primers for a PCR (see Figure 4). The primers are synthesized and then the PCR reaction is run on genomic DNA to see whether the reaction results in the selective amplification of the targeted region. If it does, then the amplified region becomes an STS. In our work at Los Alamos, we found that about half of the sequences we obtained from randomly selected clones yielded an STS.

STS Markers for Physical Mapping

STSs are being used to find pairs of overlapping clones for the construction of contig maps of human chromosomes. Since each STS is a unique site on the genome, two clones containing the same STS must overlap and the overlapping region must include the STS.

Figure 2. STSs from Chromosome 16



To check that a sequence-tagged site (STS) is a unique sequence on the genome, the polymerase chain reaction (PCR) defining that site is carried out on total genomic DNA and the products of the reaction are separated on a gel by electrophoresis. If the reaction amplifies that site and no other, all reaction products will have the same length (known from the sequence of that site) and will appear together as a single intense band on the gel.

At Los Alamos, twelve different STS markers are amplified in parallel by the PCR, and the products are separated on twelve separate lanes of a gel. The presence of only one intense band in each numbered lane of the gel shown above indicates that the STS is indeed a unique site. Fainter bands near the bottom of a lane are residual primers remaining after the PCR. The sizes of the amplified products are measured relative to a ladder of standard fragments (with known lengths that are multiples of 123 base pairs) that have been separated by length in the gel lanes marked C.

Before overlap can be detected, clones containing the same STS must be identified from among a collection of clones in a DNA library. If the individual cloned fragments have been permanently arrayed on nitrocellulose or nylon membranes,

Figure 3. Steps in Developing an STS Marker

Either create a chromosome-specific library of M13 clones, or pick a clone from the end of a cosmid contig, digest the cosmid clone with a restriction enzyme, and clone the restriction fragments in M13 cloning vectors.

↓

Sequence 200 to 400 base pairs of DNA from an M13 clone. The rough sequence determined from a single run on a DNA sequencing machine is sufficient for identifying an STS. (By "rough" we mean an average error rate of 1 in 100 bases.)

↓

Compare the sequence to all known repeated sequences using computer algorithms to help identify regions likely to be unique.

↓

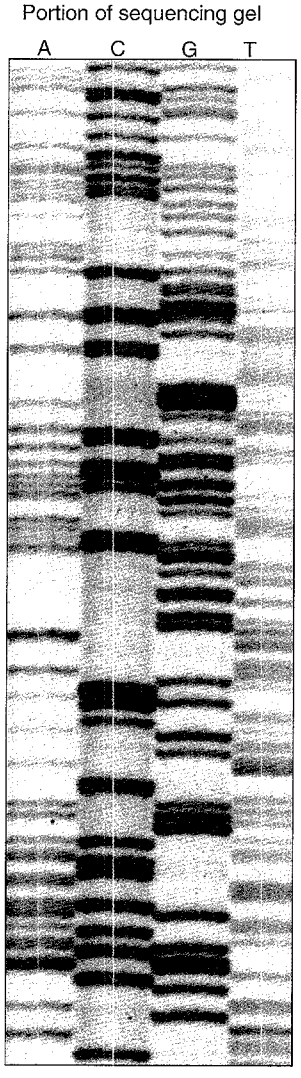
Select two primer sequences from the unique regions that are separated by 100 to 300 base pairs. Gs and Cs should comprise 45 to 55 percent of the bases in each primer sequence, and the melting temperatures of the two primers should differ by less than 5° C (see example in Figure 4).

↓

Synthesize the primers and use them to run the PCR on genomic DNA isolated from human cells. Analyze the amplification products by agarose gel electrophoresis to evaluate the specificity of the reaction.

↓

A functional STS marker will amplify a single target region of the genome and produce a single band on an electrophoretic gel at a position corresponding to the size of the target region.



then clones containing a particular STS may be identified by hybridization to copies of an STS marker. First, copies of the STS are generated from genomic DNA by the PCR. The amplified copies are labeled with radioactive ^{32}P, denatured, and then applied to the membranes containing the arrayed collection of cloned fragments. The labeled markers will hybridize only to those clones containing DNA sequences complementary to those of the markers. Clones that are positive for the STS are imaged as dark spots on x-ray films that have been exposed to the membranes containing those clones.

A more rapid screening method involves dividing a library of clones into pools and using PCR to interrogate each pool for the presence of the STS. In the PCR-based screening method, primers are synthesized for each STS, and many pools are screened in parallel. If a particular pool of cloned fragments supports PCR amplification of the STS target sequence, then at least one particular clone in the pool must contain the target sequence. Using a clever pooling scheme described below, the identification of which pools support amplification will result in the identification of the particular clone or clones containing the STS.

STS Markers for the Chromosome-16 Physical Map

In line with the five-year goals of the Human Genome Project, the Los Alamos effort to construct a physical map of chromosome 16 includes developing STS markers spaced, on average, at 100,000-base-pair intervals along the chromosome. At present about 60 percent of chromosome 16 is covered by contigs made up of cosmid clones. On average each cosmid contig spans a distance of 100,000 base pairs. We are developing STSs by sequencing regions from the clones that lie at either end of each contig. Thus far a total of 325 sequences have been obtained from such clones and about 100 of these have been developed into STSs. The STS markers will be stored in GenBank so that anyone who wants to regenerate the markers and use them to identify clones containing those markers may do so.

The STS markers from the end clones of our cosmid contigs are serving several purposes. First, they are being used to screen a library of YAC clones for clones that may overlap two different cosmid contigs and therefore close the gap between them.

Our library of 550 YACs is specific for chromosome 16. That is, the YACs contain DNA inserts from human chromosome 16 only. Since those inserts have an average size of 215 kb, the total YAC library represents a one-time coverage of the DNA in chromosome 16. The construction of such chromosome-specific YAC libraries is an important breakthrough for physical mapping and is described in "Libraries from Flow-sorted Chromosomes."

We have partitioned the YACs into pools and are using a PCR-based screening strategy to identify YACs containing each STS. Our pooling scheme, devised by David Torney in the theoretical biology group at Los Alamos, has the advantage of detecting false positive and false negative results from the PCR (see "YAC Library Pooling Scheme"). Once a YAC clone containing an STS is identified, a PCR technique (known as inter-ALU PCR) is used to generate a set of probes from that YAC. The probes are hybridized to our arrayed library of cosmid clones. If clones from two different contigs yield positive hybridization signals, then the YAC must bridge the gap between the two contigs. So far we have identified 30 YACs containing the STSs from end clones of cosmids. These YACs and seventy-five others have been hybridized to the cosmid clones resulting in the closure of sixty-five gaps in the contig map of chromosome 16..

The same STSs are being used to localize each of our cosmid contigs to an interval on chromosome 16, defined by a series of mouse/human somatic-cell hybrids containing various portions of chromosome 16. Collaborators David Callen and Grant Sutherland of Adelaide Children's Hospital in Southern Australia have collected a panel of 50 hybrid cells that divide chromosome 16 into 50 intervals with an average size of 1.7 million bases. Using a hybridization-based method and, more recently, our STSs and a PCR-based strategy, they have screened the DNA in each hybrid cell and thereby localized each of 70 contigs to one of the 50 intervals defined by the hybrid-cell panel. Those 70 contigs represent about 10 percent of chromosome 16.

Figure 4. Example of an STS

Rough Sequence—347 Bases (lower case letters indicate uncertainty in the base call)

```
                                                    ┌ Primer A
5   5'-GAATTCCTGA  CCTCAGGTGA  TCTGCCCGCC  TCGGCCTCCC  AAAGTGCTGG
51     GATTTACAGG  CATGAGGCAC  CACACCTGGC  CAGTTGCTTA  GCTCTCTAAG
101    TCTTATTTGC  TTTACTTACA  AAATGGAGAT  ACAACCTTAT  AGAACATTCG
151    ACATATACTA  GGTTTCCATG  AACAGCAGCC  AGATCTCAAC  TATATAGGGA
201    CCAGTGAGAA  ACCAATGTCA  GGTAGCTGAT  GATGGGCAAa  GGgATGGGgA
251    CTGATATGCC  cNNNNNGACG  ATTCGAGTGA  CAAGCTACTA  TGTACCTCAG
301    CTTTtCATCT  tGATCTTCAC  CACCCATGGg  TAGGTGTCAC  TGAAaTT-3'
                3'-CTAGAAGTG  GTGGGTACCC  AT-5' ──── Primer B
```

		Melting Temperature
Primer A	5'-GTT TCC ATG AAC AGC AGI CAG-3'	69.4°C
Primer B	5'-TAC CCA TGG GTG GTG AAG ATC-3'	68.7°C

The STS developed from the rough sequence shown above is 171 bases long. It starts at base 162 and runs through base 332. Primer A is 21 bases long and lies on the sequenced strand. Primer B is also 21 bases long and is complementary to the shaded sequence toward the 3' end of the sequenced strand. Note that the melting temperatures of the two primers are almost equal. A computer algorithm was used to pick out the two primer sequences and to calculate their melting temperatures.

STS Markers for Genetic-linkage Mapping

So far we have suggested that an STS yields the same product size from any human DNA sample. However, STSs can also be developed for unique regions along the genome that vary in length from one individual to another. The PCR that amplifies

Figure 5. Polymorphic STSs—Highly Informative Markers for Linkage Analyis

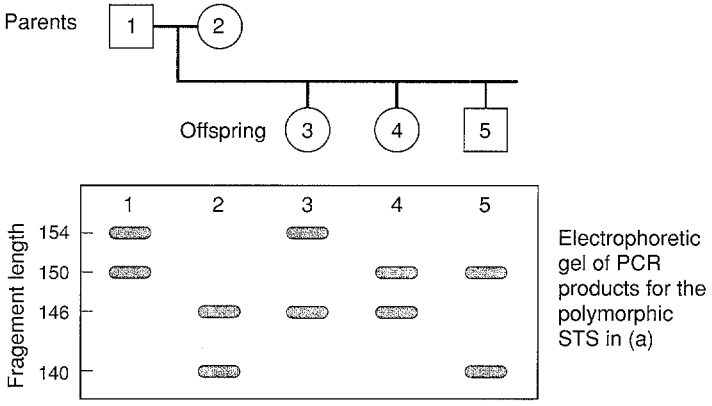
(a) A Polymorphic STS

5'- unique sequence GTGTGT . . . . . GT unique sequence -3'

$(GT)_n$

The number *n* of GT repeats varies among the population.

(b) Inheritance of the Polymorphic STS shown in (a)



A variable locus containing a short repeated sequence, such as the dinucleotide repeat $(GT)_n$, flanked by two unique sequences can be developed into an STS. An example is shown in (a). The size of the amplified product for that STS will vary depending on the value of *n* at that locus, and therefore the STS is polymorphic. Each individual carries two copies of the STS marker, one on each chromosome of a homologous pair, and each copy may have a different value of *n* and thus be a different allele of the polymorphic STS.

The inheritance of the polymorphic STS in a five-member family is illustrated schematically in (b). The electrophoretic gel shows the PCR products for the STS from each family member. The two alleles carried by the father are different from the two alleles carried by the mother. The children inherit one allele of the STS from each parent.

Because markers developed around such repeat sequences have many alleles, the likelihood that a given individual is heterozygous for such a marker is high. As explained in "Classical Linkage Analysis," at least one parent must be heterozygous for two different markers (or genes) in order to establish linkage between the two. Thus markers that have many alleles are likely to be *highly informative* for linkage analysis. (See "Informativeness and Polymorphic DNA Markers.") Polymorphic STSs will help to attain the five-year goal to construct a genetic-linkage map of highly informative DNA markers spaced at genetic distances of 2 to 5 centimorgans along each chromosome of the human genome. Moreover, these STSs are easily located on the physical map and thus provide a convenient means for aligning the linkage map with the physical map of a chromosome.

the variable region will yield different product sizes depending on which variations of the region are present in the genome of a given individual. An STS from a variable region is, by definition, a polymorphic DNA marker, which can be traced through families along with other DNA markers and located on genetic-linkage maps (see "Modern Linkage Mapping").

Figure 5(a) shows an example of a unique region that has variable lengths and can be developed into a polymorphic STS. At either end of the region is a unique sequence about 20 nucleotides long that can serve as a primer sequence for the PCR. Between those two sequences is a simple tandem repeat, $(GT)_n$ (or GT repeated in tandem *n* times). Such dinucleotide repeats are scattered throughout the human genome as are tri-, tetra-, and pentanucleotide repeats. Moreover, the number *n* of tandem repeats at a given locus along a chromosome is an inherited trait that tends to vary widely among the population. Thus each such variable locus has many different alleles (or forms), each one defined by the number *n* of tandem repeats between the unique sequences.

STSs are being developed for this abundant class of variable regions. Since the varying sizes of the PCR products from a polymorphic STS correspond to the alleles of that marker, PCR followed by gel electrophoresis of the amplified products is the method of detecting which alleles of the marker are carried by an individual [see Figure 5(b)].

Polymorphic STSs are particularly useful because they can serve as landmarks on both the physical map and the genetic-linkage map for each chromosome, and thus they provide points of alignment between the different distance scales on these two types of maps.

At Los Alamos we have identified the location of $(GT)_n$ repeats as part of our fingerprinting and mapping strategy (see "The Mapping of Chromosome 16"). We are now developing these regions into STSs for use in linkage mapping. ■