Genetics of Childhood Disorders: VII. Cloning Genes of Interest

RACHEL SPARKS, B.A., PAUL J. LOMBROSO, M.D., AND JEFFREY R. GRUEN, M.D.

An estimated 100,000 genes lie on our 46 chromosomes. One third of these are thought to be specifically expressed in our brains at different periods over the course of our lives. Over the past few years, a small but growing number of genes have been cloned. Mutations in a number of these genes lead to a disruption of the normal CNS development. The ability to isolate genes is probably one of the most important recent achievements in neuroscience. This is a necessary first step before investigators are able to understand the normal role of these genes and how their protein products function in our brains. In the next several columns, we will discuss exactly how genes are cloned and how researchers learn more about the structure and function of the proteins they encode.

Last month, we reviewed how cytogenetic techniques have advanced to the point at which we can actually visualize chromosomal abnormalities such as a deletion. These physical disruptions point to the site on the chromosome that presumably has a mutation that contributes to the observed phenotype in the affected individual. These types of cytogenetic analyses reduce the search for a mutated gene from the several billion nucleotides on our total complement of DNA down to the millions of nucleotides contained within the deleted site. Considerable work remains, however, before the gene of interest is actually isolated, and various methods have been developed over the past decade to do exactly that.

Most of the techniques to isolate genes make use of "libraries." Two types of libraries will be described. The first is called a complementary DNA (cDNA) library. This type of library is produced from—and therefore contains—all the messages expressed in a given tissue. To construct such a library, one first isolates the pool of RNA messages from a tissue, copies these messages into the more stable cDNA molecule, and places the cDNA molecules into an appropriate vector. Several types of vectors exist, but in general they serve similar functions: to carry a single cDNA molecule and to facilitate their own replication, thereby allowing the investigator to make many copies of the library. A single message, or "insert," is carried in each vector. Each vector and its insert is termed a clone, and it is the total collection of cDNA clones that is called the library.

cDNA libraries consist of all the expressed sequences from a tissue of interest. The library will contain many copies of the most common housekeeping genes as these genes are abundantly expressed and many copies of their RNA messages were originally present when the RNA was isolated to make the library. The library will contain fewer copies of less commonly expressed genes. However, if the screening methods are powerful, one should be able to find even very rarely expressed messages. The ultimate goal in screening a cDNA library is to isolate a clone of interest. One then determines the nucleotide sequence for that clone, and the nucleotide sequence defines the precise amino acid sequence for the protein the gene encodes.

It should be noted that one can construct a cDNA from any tissue in the body. cDNAs can also be made from the same tissue at different time periods. For example, to identify genes that are transiently expressed during the development of the cerebral cortex, one would obtain cerebral cortices early in development, isolate the pool of RNA messages present in that tissue, convert these messages into cDNA molecules, and place the cDNA molecules into an appropriate vector. The end result would be a cerebral cortex cDNA library from early in development. Many of the clones contained in this library would also be found in an adult cerebral cortex cDNA library, as many genes are expressed at both times. However, the fetal cDNA library would also contain unique messages that are likely to be involved in the early development of the cerebral cortex.

A second type of library is called a genomic library. In this case, the goal of the investigator is to make a library from the total amount of chromosomal DNA. The DNA molecules are digested into smaller, more manageable pieces, and these smaller DNA sequences are placed into a vector. Special vectors have been constructed for this purpose, as the amount of DNA that must be packaged is considerably larger than for cDNA libraries. It is possible to replace large fragments of yeast chromosomes with segments of DNA from human chromosomes and have these segments replicated by the yeast. The library that is constructed using this type of vector is called a yeast artificial chromosome (YAC) library (Fig. 1). It allows researchers to package up to 1 million nucleotides into a vector. This larger stretch of genomic DNA may contain several genes, each with their respective exons, introns, and regulatory regions.

It has been appreciated for some time that many housekeeping genes have particular repetitive sequences near their regulatory regions. These sequences are enriched for a partic-

ular dinucleotide repeat (CpG) and are distinguishable from other nucleotide sequences by the extent of methylation that occurs. This chemical modification to the DNA molecule makes the region sensitive to digestion by rare cutting restriction endonucleases, enzymes that cleave DNA at specific nucleotide sequences. One of the enzymes that cuts here as a consequence of the methylation is called HpaII, and as a result these regions have been termed HTF (HpaII tiny fragments) islands.

It is possible to isolate some genes by taking advantage of these HTF islands. As was mentioned, the majority of HTF islands are associated with the regulatory regions of many genes. Isolating the downstream nucleotide sequence would

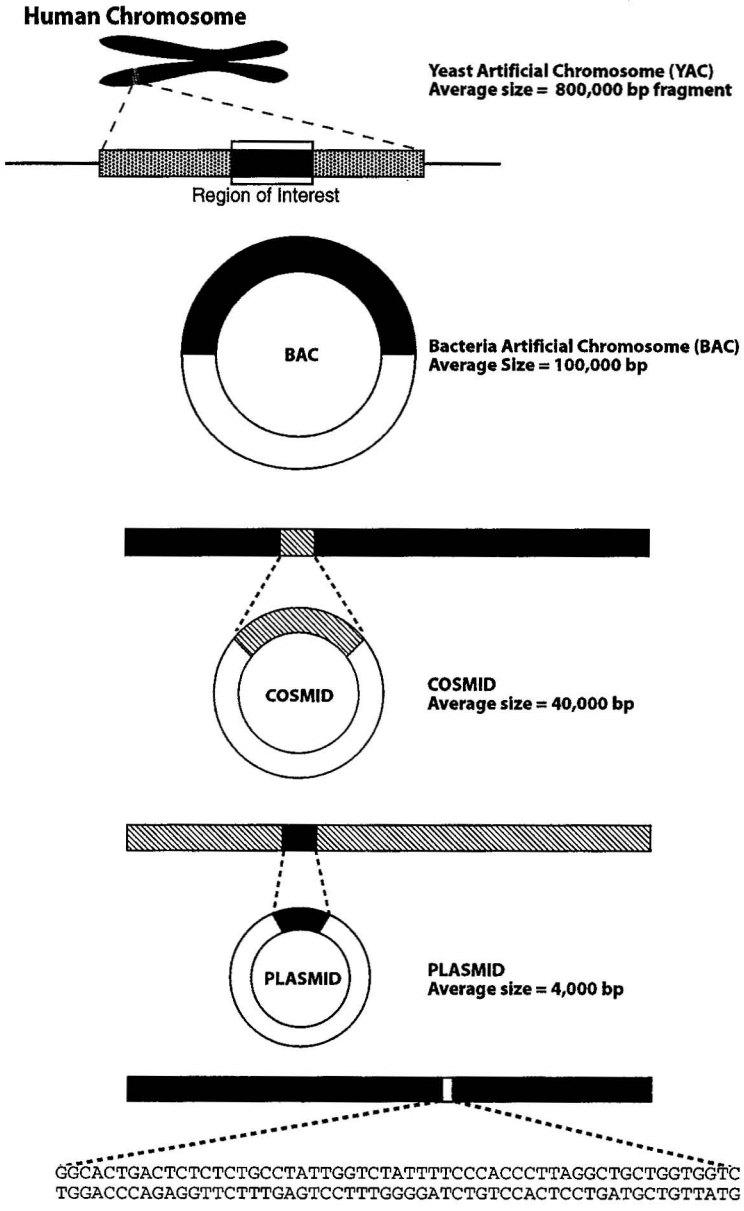

Fig. 1 Isolating a gene of interest requires dividing a particular chromosomal region into progressively smaller pieces until one can finally send the DNA off for sequencing. Individual human chromosomes are much too large to be cloned in their entirety and must be broken into smaller fragments. Genomic sequences of up to 1 million bases can be subcloned into YACs. To identify a specific gene, one first pulls out one or several overlapping YAC clones that span the region of interest. One then screens a BAC library using expressed sequence tags that have been found in the YAC clone. BACs may be further subcloned into cosmids or processed directly into plasmids. Either the cosmid or the plasmid is typically sent for sequencing. Overlapping clones are often required to obtain the entire nucleotide sequence that encodes a protein of interest.

disclose the likely coding sequence for a protein. After subcloning genomic fragments that lie near HTF islands, the isolated genomic clones are used as probes to screen cDNA libraries. The technique takes advantage of the fact that complementary nucleic sequences bind very tightly to each other. Screening a library in this fashion uses the genomic clone to bind to the cDNA that was actually transcribed from that region. Unfortunately, the regulatory regions of most tissue-specific genes do not have HTF islands. Hence this technique has limited potential to identify the majority of genes expressed in the CNS.

Additional methods have therefore been developed. One of these techniques takes advantage of chromosomal abnormalities that are present in a number of disorders. It is possible to visualize these abnormalities either by routine karyotyping methods or the more recently developed chromosomal painting technique (see last month's column). One can also hone in on candidate chromosomal regions by linkage analysis or association studies.

The region immediately surrounding such a suspected chromosomal region becomes the area that is likely to contain a gene of interest. But how does one actually go about isolating the gene? One method is to identify several YAC clones that span the chromosomal region of interest. The YAC DNA is then digested into fragments, and these fragments are immobilized on a nylon filter. The filter is probed with members of a cDNA library. The underlying principle is that clones that were originally transcribed from the chromosomal region of interest will bind to the fragments of genomic DNA bound to the filter. Once again, this is because complementary nucleic acid strands have an enormous affinity for each other. cDNA clones that represent messages expressed from other chromosomal regions will not bind to the YAC fragments immobilized on the filter, and these clones will be washed away. The bound clones are carefully removed and amplified. These clones then represent potential genes that were transcribed from the chromosomal region of interest.

If one were looking for genes that cause a particular disorder, then one would sequence that gene from several individuals with the disorder and compare the nucleotide sequence with the sequence found in unaffected individuals. Convincing evidence for that gene being the gene responsible for causing an illness comes when mutations are found among other affected individuals and not found in unaffected individuals.

This method is termed cDNA hybridization selection. Although it has been largely successful, there are several drawbacks. For one, some of the recovered sequences are false-positives and derive from transcribed sequences in other regions of the genome. In addition, ribosomal and repeat sequences, which normally represent up to 10% of the human genome, must be blocked so as not to be selected in the experiment. Finally, the selected cDNAs are rarely full-length and tend to be relatively short. To achieve a more complete representation of all transcribed sequences in a given genomic region, several cDNA libraries will need to be used, and several screenings performed. cDNA hybridization selection performed with several cDNA sources can typically identify 70% to 80% of the transcribed regions in a given genomic clone.

The method that was just describes relies on the expression of genes in order to isolate a gene of interest. If, for example, the cDNA library used to screen the genomic fragments was made from adult brain RNA, then genes expressed during early development will not be found. If that gene is the gene responsible for a disorder, it will not be isolated by screening the genomic filters with an adult brain cDNA library.

More recently, techniques have been developed that do not depend on the expression of genes for those genes to be isolated. One of these methods is termed exon trapping. The underlying principle relies on the fact that the vast majority of genes are processed after they are transcribed from DNA into RNA. One of the important steps in the processing of RNA messages is to bring together the exons that contain the protein coding sequence and to remove the large stretches of intervening, noncoding sequences. These intervening sequences are termed introns, and the procedure that removes the introns and joins together the exons is called splicing. A number of small nuclear proteins are responsible for bring about the splicing of genes into mature RNA messages. These proteins recognize conserved nucleotide sequences at either end of an intron that needs to be removed. When this signal is found, the small nuclear proteins bind tightly, and the exons of a gene are spliced together while the introns are removed.

In the expression-independent method for isolating genes, the nucleotide signals that indicate a splicing site are used to identify the exons on either side. Genomic DNA is digested into small fragments, and the fragments are cloned into a specialized vector that contains the first half of the splicing signal necessary for intron removal. If the genomic segment that is cloned into the vector contains the remaining half of the necessary splicing signal, a splicing event will occur. These splicing events can be detected through a simple assay, and it is a relatively easy matter to then isolate the clones identified. This technique is most useful when the chromosomal region that is suspected to carry a mutated gene has been identified. It is considerably easier to trap exons expressed in the much smaller area surrounding a translocation, for example, then to have to use the entire genomic sequence.

The largest drawback to exon trapping is the presence of cryptic splice sites in the genome. These sites will lead to the isolation of false-positive clones, or clones that do not represent exons. This necessitates the rescreening of candidate regions in order to guarantee that isolated clones are true-positives. In addition, this technique cannot be used to locate genes that lack introns. Despite these problems, exon trapping is a quick and

efficient way to locate coding regions within specific regions of genomic DNA.

With the expansion of computer databases and the sequencing of longer stretches of DNA, computers are playing an increasingly important role in identifying expressed sequences. Currently, only 3% of the genome has been sequenced, and the longest continuous stretch available in public databases is less than 1.5 million base pairs. However, over the next few years, genome centers anticipate sequencing millions of additional nucleotides of genomic DNA, and the completion of the sequencing of the entire human genome is scheduled for the beginning of the next millennium.

Personal computer programs exist that can quickly search genomic DNA sequences for the presence of HTF islands, regulatory sequences, promoters, splice sites, and open reading frames. Large sequence databases, such as GenBank at the National Center for Biotechnology Information, are available on the Internet. These databases are the repositories for DNA and RNA sequences from laboratories around the world. There are several useful tools available to search these databases.

As the human genome becomes well covered with genomic clones, it will be increasingly important to identify the transcribed sequences within those clones. The methods discussed here such as cDNA hybridization selection, exon trapping, and computer-based analysis of deposited sequence data will identify many human genes. We can then ask what their normal function is and how they cause disease when mutated.

WEB SITES OF INTEREST

http://www.ornl.gov/hgmis/publicat/primer/intro.html
http://www.ncgr.org/ncgr
http://ihs2.unn.ac.uk:8080/bbgenome.htm
http://www.kumc.edu/gec/hgpwww.html

ADDITIONAL READINGS

Bird A (1997), Does DNA methylation control transposition of selfish elements in the germline? *Trends Genet* 13:469–472

Ebeling M, Suhai S (1997), Molecular databases on the Internet. *J Mol Med* 75:620–623

Hochgeschwender U, Gardiner K (1994), *Identification of Transcribed Sequences.* New York: Plenum

Lovett M (1994), Fishing for complements: finding genes by direct selection. *Trends Genet* 10:352–357

Rowen L, Mahairas G, Hood L (1997), Sequencing the human genome. *Science* 278:605–607

Snyder E, Stormo (1993), Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res* 21:607–613

Accepted April 15, 1999.

Ms. Sparks was an undergraduate at Yale College, Dr. Lombroso is Associate Professor, Child Study Center, and Dr. Gruen is Research Scientist, Department of Pediatrics, Yale University School of Medicine, New Haven, CT.

Correspondence to Dr. Lombroso, Child Study Center, Yale University School of Medicine, 230 South Frontage Road, New Haven, CT 06520; e-mail: paul.lombroso@yale.edu.

To read all the columns in this series, visit the Web site at http://info.med.yale.edu/chldstdy/plomdevelop/

0890-8567/99/3810–1316©1999 by the American Academy of Child and Adolescent Psychiatry.

The Impact of Welfare Reform on Parents' Ability to Care for Their Children's Health. S. Jody Heymann, MD, PhD, Alison Earle, PhD

Objectives: Most of the national policy debate regarding welfare assumed that if middle-income mothers could balance work while caring for their children's health and development, mothers leaving welfare for work should be able to do so as well. Yet, previous research has not examined the conditions faced by mothers leaving welfare for work. *Methods:* Using data from the National Longitudinal Survey of Youth, this study examined the availability of benefits that working parents commonly use to meet the health and developmental needs of their children: paid sick leave, vacation leave, and flexible hours. *Results:* In comparison with mothers who had never received welfare, mothers who had been on Aid to Families with Dependent Children were more likely to be caring for at least 1 child with a chronic condition (37% vs 21%, respectively). Yet, they were more likely to lack sick leave for the entire time they worked (36% vs 20%) and less likely to receive other paid leave or flexibility. *Conclusions:* If current welfare recipients face similar conditions when they return to work, many will face working conditions that make it difficult or impossible to succeed in the labor force at the same time as meeting their children's health and developmental needs. **Am J Public Health** 1999;89:502–505. Copyright 1999 by the American Public Health Association.