Urol Res (1998) 26:361–369

INVITED EDITORIAL

Jeffrey R. Gingrich · Mitchell S. Steiner

A gene hunting we will go!

Received: 16 April 1998 / Accepted: 26 June 1998

Abstract Two general approaches toward the identification of new genes which may contribute to specific developmental, physiologic, or pathologic processes can be pursued: (1) a strategy of positional cloning or reverse genetics and linkage analysis in which affected individuals or families are studied to identify defective genes, and (2) an immediately focused approach which uses available partial information about a gene to screen cDNA libraries toward the isolation and characterization of the complete gene. The general principles that form the basis of how these two strategies aid the gene hunter in the pursuit of new genes is discussed. The modern day urologist should become familiar with these techniques and the promise they hold for the future diagnosis and treatment of urologic diseases.

Key words Gene identification · Positional cloning · Expressed sequence tags · cDNA library screening

Introduction

The pendulum has swung the other way! The old adage that man is simply the product of his environment is not entirely correct. As our understanding of genetics continues to evolve, it has become quite clear that an individual's genes also play an important role. Genes not only dictate how we look physically, but they also influence diverse processes including personality, aging, and disease that were previously attributed to the environment. In fact, there is constant interaction between genes and the environment. The environment may mediate some effects through modification of gene expression (epigenetic phenomena). Not only can genetic mutations cause disease, but in addition they can even predispose or make us more susceptible to other diseases including cancer. As urologists, it has become imperative that we develop a better working knowledge of modern genetics. Ultimately, abnormal gene function resulting in urologic disease, including cancer, may be corrected by gene therapy, thereby restoring the normal cell phenotype.

The size of the genetic blueprint for the human body is immense. There are roughly 64 000 to 80 000 genes scattered amidst 3 billion nucleotide bases [3, 12, 26]. Only 2% of the total bases make up protein-coding portions of genes; the remaining 98% of total bases has no known function and is sometimes referred to as junk DNA [26]. To date, over 16 000 genes have been discovered with 70–80 new genes being discovered each month. As DNA sequencing technology continues to improve, the rate of gene hunting will continue to increase exponentially. The goal of the Human Genome Project initiated in 1990 is to sequence the entire human genome by the year 2005 at a cost of 211 million dollars a year [21]. The impact of The Human Genome Project will revolutionize medicine by laying the foundation for new diagnostic and therapeutic technologies including the ability to screen a person for thousands of genetic diseases by applying one drop of blood onto a special chip that can then be scanned by a computer. Using currently available technology, sequencing of approximately 80 million bases will be completed this year. Therefore, through new technology the annual production rate is expected to and must reach 400–500 million bases in just a few years if the human genome project expects to realize its goal [22]. In addition to obtaining the nucleotide sequence, these sequences must also be analyzed to decipher the genetic code.

Because direct DNA sequencing is expensive, slow and labor intensive, other gene hunting strategies have emerged and are being conducted in parallel. For the sake of simplicity, the genome can be considered analogous to a book where the chromosomes are chapters and the genes are single sentences made up by nucleotides containing the nucleotide bases adenine (A), guanine (G), cytosine (C), and thymine (T) instead of the 26

J. R. Gingrich · M. S. Steiner (✉)
Department of Urology,
Division of Urologic Oncology, University of Tennessee,
956 Court Ave., H220, Memphis TN 38163, USA

letter alphabet. Just like a sentence begins with a capital letter and ends with a period, the sequence encoding a gene begins with an ATG and ends with a UAA, UAG, or UGA stop and a poly-A tail. Thus, even if the paragraph contains additional random letters, or junk DNA, the sentences, or genes, can be found among the letters because of their expected structure. Consequently, two general approaches, which are not mutually exclusive, are currently used to hunt for genes. One strategy, which is in a sense somewhat of a "shotgun approach", is called positional cloning, or reverse genetics. This approach studies families who are affected by defective genes and then by linkage analysis finds the chapter and even the page that the putative diseased gene is located. Once the location of the gene has been mapped to that page of a chapter (subregion of a chromosome), then candidate genes that are known to reside in this location (sentences located on that page or even in a paragraph) are studied to determine whether they may be the responsible gene (sentence) in question [8, 26]. A second, more immediately targeted approach uses available parts of or information about a gene to screen a cDNA library [DNA copies of all the RNA messages from the genes (sentences) in the genome (book)] to find a match, to isolate or "clone" the gene, and then to obtain the DNA sequence of the entire cDNA. As you will see, both of these techniques merge together once the gene is identified as genetic databases and maps will be queried to determine whether the isolated (cloned) or mapped candidate gene is known or is novel and is then characterized. This current review seeks to introduce the reader to the general principles that are the basis of how these two strategies aid the gene hunter in the pursuit of new genes related to specific developmental, physiologic, or pathologic processes.

Positional cloning

Positional cloning is a powerful technique that assumes that there is no functional information about the gene in question and must locate the responsible gene on the basis of map position [8]. Positional cloning begins by linkage analysis of multiple affected families to identify the subregion of a chromosome where the gene may located. Then, genetic maps of known genes that reside in that subregion are analyzed to see if one of these candidate genes may be the actual gene. The success and efficiency of positional cloning depend on the quality of the DNA markers and the number of genes already mapped to that particular subregion of the chromosome [8].

Gene linkage analysis

Clues to where a particular gene is located may be derived from comparing the inheritance pattern of the mutated gene with that of DNA markers which have known chromosomal locations. The coinheritance, or genetic linkage, of a given disease gene and DNA marker suggests that they are physically close together on the chromosome, hence "linked." The genetic basis to this principle is that during meiosis the DNA replicates first within a cell committed to meiosis (DNA copies go from 2N to 4N) resulting in sister chromosomes (chromatids) that are attached at the centromere. In meiosis I, each set of homologous chromosome pairs (one set from each parent) independently assorts in a random fashion with respect to parental origin between the two daughter cells. The DNA does not duplicate again in these daughter cells (2N) in meiosis II, and like mitosis, the centromere of each chromosome divides and the sister chromatids segregate to form four germ cells (1N) [15, 27]. With independent assortment of chromosomes, 2^{23} different genetic combinations of germ cells are possible. Further genetic diversity occurs through another process called recombination. During meiosis I, the homologous nonsister chromatids undergo DNA crossover, that is, the breakage, swapping, and reunion of pieces of DNA between the homologous chromosomes inherited from the mother and those inherited from the father. The net result is two new hybrid chromosomes are formed. The closer the physical proximity of two gene loci, the less likely there will be a recombination event and the higher the probability that these genes will be coinherited. Hence, the degree of linkage is proportional to the distance between the gene (allele, locus) and the DNA markers on a given chromosome. These recombination events are measured in terms of centimorgans (cM) one of which is equal to one recombination event per 100 meioses. A rule of thumb is that 1 cM represents approximately 10^6 base pairs of DNA (1 Mb) [13].

By comparing the frequency of recombination between genetic DNA markers and the responsible gene, the degree of separateness can be deduced and statistically analyzed to determine gene linkage. If the marker and gene independently assort, then the gene and marker are not linked. If they cosegregate, then they may be linked. In linkage analysis, the coinheritance of the marker and gene are followed within a nuclear family. The probability that the observed inheritance pattern could occur by chance alone, i.e., unlinked is calculated. The calculations assume a particular degree of closeness of the gene and marker, and the ratio of 2 probabilities that no linkage versus a specified degree of linkage is determined. This ratio is expressed as the odds for that degree of linkage. As this ratio is logarithmic, it is denoted as Logarithm of the Odds, or LOD score. It is generally accepted that if this ratio is greater than 3, then the genes are linked with the odds being a 1000:1 chance that the responsible gene and marker are linked [9].

Genetic markers are important for positional cloning

Once a particular region of the chromosome has been identified by linkage analysis between known DNA markers and the responsible gene, then the known candidate genes that reside in that region are implicated in the disease. Unfortunately, of the 80 000 possible genes only about 16 000 have been discovered and even fewer mapped. Hence, for positional cloning to be successful it is imperative that new genes are identified and mapped. Although the Human Genome Project has taken on the ambitious task of sequencing the entire human genome, this is still 7 years away assuming they remain on schedule. In the meantime, there have been some major developments in approaches to gene identification, cataloging, and mapping that have proven to be invaluable to gene hunters using positional cloning.

Markers are sequences of DNA that have been mapped to chromosomal locations that encompass the entire genome with sufficient diversity between individuals (polymorphic) to allow the founder's alleles to be distinguished from other family members. Interestingly, the human genome differs slightly from individual to individual, and these variations may be exploited to be used as genetic markers. DNA sequences may be different in noncoding and coding regions alike. If the base change occurs within the coding region resulting in an abnormal protein, then this change represents a mutation. In contrast, if this base alteration occurs in the coding region with no functional consequence or if these are base differences in the noncoding region, then this is regarded as a polymorphism. The estimated frequency of nucleotide variation outside the coding region may approach 1:100 base pairs. These base changes may alter restriction enzyme sites (specific sequences of four to eight nucleotides that a bacterial enzyme recognizes and cleaves or "cuts" the DNA) so that the enzyme can no longer cut at that specific site, leaving a larger residual fragment of DNA following enzymatic digestion. Another form of DNA polymorphism is based on variations in the number of tandem repeated DNA sequences lying between two restriction enzyme sites. The number of tandem repeated units at these variable-number tandem repeat loci (VNTR) varies between 11 and 60 bp and may be different for each homologous chromosome. Accordingly, a restriction enzyme cut may result in different sized fragments depending on the number of VNTR between restriction enzyme cuts for each loci. Collectively, these polymorphisms that result in different sized chromosomal fragments following restriction enzyme digestion of the DNA have been called restriction fragment length polymorphism (RFLP). If this change is in one chromosome and not the other, then the two chromosomes can be distinguished from one another in such an individual. If heterologous, a short and long piece of DNA is obtained during the digestion and the patient is designated "informative" at this locus since this approach can be used for linkage analysis. On the other hand, if homologous (same sized fragment for both chromosomes) then the individual is not informative because the two homologous chromosomes cannot be distinguished from one another. The RFLP can be inherited like a gene being passed from generation to generation and can be used as a genetic marker. RFLP maps are available for identifying candidate genes.

Currently, RFLP linkage studies have essentially been replaced by polymerase chain reaction (PCR)-based microsatellite DNA markers, the genetic DNA marker of choice. Microsatellites are short tandem repeats of di-, tri-, or tetra-nucleotide repeats [4]. Comprehensive genetic maps have been developed containing greater than 5000 microsatellite markers spanning across the entire genome at a resolution of almost 1 cM [10]. These maps are available from the Centre d'Etude du Polymorphisme Humain and the Human Genome Database. In essence, PCR primers that flank these microsatellites (tandem repeats) may be used to amplify the region of a chromosome or the entire DNA genome. The variable sized fragments are then separated by gel electrophoresis and detected by radiolabeled probes [29]. If linkage is established, then the disease gene can be assigned to a rough chromosomal location spanning 10–20 Mb of DNA. Then, additional microsatellite markers that map in close proximity of the gene can further delimit the location of the gene to a 1-Mb region. Microsatellite markers are more abundant, easier to analyze, and more informative than RFLP [9].

Help from expressed sequence tags

Once a cDNA is isolated or cloned, it has to be sequenced. This is a laborious process that requires sequencing of tandem overlapping regions of DNA as current sequencing technology can only provide quality nucleotide sequence for 300–400 bases at a time. DNA sequencing must be performed forward ($5' \rightarrow 3'$) and reverse ($3' \rightarrow 5'$) directions three separate times to be considered accurate. Next, all of these short sequences of 300–400 bp DNA, also known as contigs, have to be reconstructed. The cost for sequencing a gene 2.5 kb in size is approximately US$4000. As you can see, gene hunters were truly in a rut as the cost was high and pace of sequencing slow.

In 1991 an ingenious cost-effective approach to gene hunting was reported by Adams et al. [2]. They selected all the random cDNA clones and performed a single automated sequencing read from one or both ends of the cDNA insert. This new class of sequence of a small fragment of the entire cDNA was termed expressed sequence tag, or EST. ESTs tend to be short, consisting of 400 bp and are relatively inaccurate with a 2% nucleotide error rate. No attempt is made to characterize or identify the entire sequence of the clone. It is also fully recognized that there may be many redundant and overlapping ESTs and sequencing artifacts including vector (the bacterial host DNA) sequence or other contaminants, but the value of putting some sort of address label on as many unknown cDNAs (genes) as possible was an important feat. In fact, these address labels have been collated since 1992 into a database called dbEST [7]. Merck & Co. and the Washington University

Genome Sequencing Center had contributed 508 945 human EST sequences by 1997 to the dbEST database [14]. Other groups have done the same so that currently there are over a million ESTs in the dbEST database. The number of ESTs placed in dbEST is expected to double every 18 months [6, 25]. GenBank developed a database search tool called BLAST which allows the scientist to query sequence similarity either deduced from a protein sequence or from another species against dbEST [6]. Another group called the I.M.A.G.E. consortium has been instrumental in collecting EST cDNA libraries, organizing these clones, and, for a small fee, making these clones available to researchers for further sequencing and characterization [18].

As there are only 80 000 genes, then there must be a large number of redundant and overlapping genes represented in the dbEST database. In an attempt to reduce redundancy, the Institute for Genome Research undertook a large scale DNA fragment assembly project so that similar ESTs, that is, gene address labels, were merged to make up 62 808 partial DNA sequences called tentative human consensus sequences (THC). However, a further 175 563 ESTs remained unmatched [1]. UniGene took another approach. They arranged only 3′ end sequenced ESTs and known full-length mRNAs from characterized genes into clusters that very likely represent distinct genes. Approximately 62 421 sets have been formed, which gives a rough estimate of the number of human genes found so far.

Although these approaches have been helpful in providing organization to the exponentially growing number of ESTs being entered into dbEST, it still does not help the gene hunter using positional cloning techniques. Rather, what is really needed is a mapped location within the genome for each of these ESTs. To overcome this problem, nonredundant sets of 3′ end sequences ESTs from UniGene were distributed to a Radiation Hybrid (RH) Consortium made up of the Whitehead Institute for Genome Research, the Sanger Center, the Stanford Human Genome Center, and the Wellcome Trust Center for Human Genetics who together sought to map these ESTs. The RH Consortium mapped the ESTs using radiation hybrid techniques and designated the mapped ESTs as sequence tagged sites, or STSs [24]. Over 16 000 STSs, roughly 16 000 individual genes, or one-fifth of the total number of genes were determined by this approach [26]. Given the density of this new genetic map, there is a 1 in 5 chance that the diseased gene will correspond to an EST that has already been localized – an STS. Thus, modern gene hunting strategies have centered around these STS markers and their maps.

Summary of how gene hunting through positional cloning is really done today

Positional cloning assumes that no functional information about the gene exists; and therefore, the responsible gene must be located on the basis of map position. The first step in positional cloning is linkage analysis of many affected families with multiple affected members. This process begins with clear phenotype determination, pedigree ascertainment, and obtaining DNA from these family members for study. DNA is usually obtained from blood samples by centrifuging the blood and isolating the buffy coat containing the lymphocytes. The DNA is then extracted from the lymphocytes. Next, highly polymorphic microsatellite markers that span 10–20 Mb are used to type the entire genome, or if known, a specific chromosome. Computer-based statistical analysis of the data is performed to determine whether there is significant cosegregation between the microsatellite markers and the diseased gene. If linkage is established (LOD > 3), then the position of the diseased gene has been localized to a subregion of a chromosome spanning 10–20 Mb. Additional polymorphic microsatellite markers are then used for that subregion of the chromosome to increase the resolution to a 1-Mb region. Now the putative gene locus must be identified, so genetic maps are then consulted to see which genes reside in this location. By definition, since function is not known, all of these genes in this region are considered candidate genes. Next, the gene hunter goes back to the affected families for the following: (1) to determine the presence of a mutation or change of gene expression for any of these candidate genes; (2) to confirm that the candidate gene has the same inheritance pattern as the disease; and (3) to show that the mutation is not a polymorphism found within the general population [8]. If these criteria are satisfied, then the gene and its location have been found. Today, gene hunters have become quite efficient and quick in finding candidate genes for diseases. For example, Wooster et al. [31] were able to map and clone the entire BRCA2 (breast cancer gene 2) in a span of 2 weeks! Important genes that have been found or pursued using positional cloning include the von Hippel-Lindau disease tumor suppressor gene (Fig. 1) and recently the hereditary prostate cancer 1 gene (HPC1) locus was mapped to chromosome 1q by using segregation analysis of high-risk prostate cancer families [17, 28].

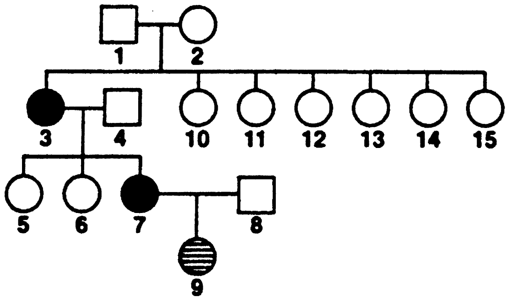

Fig. 1 Pedigree of a family affected with von Hippel-Lindau disease. *Filled circles* represent affected individuals. *Hatched circle* represents an individual predicted to be affected (From [17].)

Specific target gene cloning

In contrast to positional cloning strategies for gene hunting when no information is known about a gene, some clues to the structure and function of a gene may already be available. Genes which contribute to a developmental, physiologic, or pathologic process may be implicated at the level of expression, that is at the level of RNA and protein expression. For example, at the RNA level, gene expression may be relatively increased or decreased in the normal compared with the malignant state due to gene amplification, deletion, or mutation. Changes in gene expression may also be due to increased transcription of the messenger RNA from the DNA or due to changes in RNA stability which leads to increased or decreased breakdown of the RNA. Likewise, protein expression may increase or decrease due to changing levels of messenger RNA expression, changes in protein stability due to mutations, or changes in the expression of enzymes which degrade the protein. Isolating a gene directly from the genomic DNA can be relatively difficult since an allele may be found once within the whole genome. In contrast, RNAs are often expressed at high levels in specific tissues and relative to other genes within the tissue. An example of this is testosterone production in the testis. The identification of a tissue in which a gene is expressed at high levels dramatically increases the likelihood of isolating the RNA for that gene. The RNA can then be used to isolate the gene DNA from the rest of the genome.

cDNA library screening

As described below, many techniques to identify important new genes such as ddPCR (see below) and degenerate PCR provide only a partial cDNA sequence of a gene which must then be used to isolate the full length cDNA from a cDNA library. In other cases, no nucleic acid sequence may be known but the protein may be available. A full or partial peptide sequence can be obtained from the protein and based on that sequence a corresponding nucleic acid can be synthesized. Because several amino acids may be coded for by more than one codon the nucleic acid sequence cannot be exactly predicted. However, a potential nucleic acid sequence can be predicted, a corresponding nucleic acid synthesized and then used to probe a cDNA library.

In mammalian cells, a DNA template is required for *DNA polymerase* to duplicate DNA prior to cell division. Of the estimated 80 000 genes in the human genome only approximately 15% are expressed at any one time within an individual cell. The ability to follow changes in gene expression at the RNA level is based on the discovery that RNA viruses are capable of replication using an enzyme called *reverse transcriptase*. Reverse transcriptase is able to transcribe cDNA using an RNA template [5, 30]. Scientists use reverse transcriptase to transcribe collectively cDNAs from RNA which has been isolated from an organ, a tumor, or tissue culture cell line. Recently, techniques using laser-assisted microdissection have been developed allowing the isolation of RNA from single cells as they progress from a normal to malignant phenotype [11]. Each cDNA is then inserted into the DNA of a bacteriophage virus such as lambda (λ) forming a clone. The collection of all the cDNA clones is called a library. When the λ library is cultured with bacteria on an agar plate, a single phage infects a single bacterium, replicates, lyses it, and then infects and lyses more bacteria around it forming a bare region or plaque. The cDNAs, still within the λ DNA, can then be transferred by capillary action to a nylon filter. Many cDNA libraries are now constructed directly in the bacterial DNA without requiring the use of the λ phage. Utilizing these methods, cDNA libraries which are enriched for genes preferentially expressed within specific tissues are developed. Partial nucleic acid sequences obtained through any method can be labeled with a radioactive isotope and then hybridized to filters from a cDNA library. The plaques which hybridize strongly to the probe are then isolated from the agar, expanded, and sequenced to confirm that they contain the probe sequence and to determine whether they represent the full length cDNA sequence.

In situations where no nucleic acid or peptide sequence is known, an antibody to the protein product of the gene of interest may be available. When a cDNA library is constructed, the cDNAs can be inserted into the bacterial DNA in the correct orientation and in the correct reading frame such that the cDNA is expressed in the bacterium under the control of a gene normally expressed in the bacterium. This results in the production of a hybrid or fusion protein which is a combination of the bacterial and the inserted cDNA gene sequence. This type of library is called an "expression library." The library filters can then be screened with the available antibody to detect which phage or plasmid is producing the protein.

Yet another method to screen a library may be based on the functional interaction of the fusion protein if interaction of the protein of interest with other cellular or extracellular proteins has been previously characterized. The cDNA expression library filters may be screened with a radioactively labeled protein probe. For this type of screening, the protein must not only be expressed by cDNA clones within the *E. coli* expression library, but the normal interaction between the proteins must be maintained under the experimental conditions. Therefore, this method of library screening is somewhat more complex. Some proteins may not be functionally expressed in bacteria such as *E. coli*. They may require expression in mammalian cells such as COS or other specific cell types due to unknown factors. cDNA plasmids may also be successfully transfected into and expressed in mammalian cells. The protein expression by individual cells containing the appropriate plasmid may be selected for in culture utilizing specific antibody cell sorting through a technique known as panning. The cells expressing the protein can then be expanded, the plasmid isolated from the cells, and transfected into bacteria for expansion and sequencing.

Identification of preferentially expressed genes

The identification of preferentially expressed genes may be accomplished utilizing several different methods. One of the first methods developed is called differential hybridization. For example, to identify genes increasing in expression in benign prostatic hyperplasia (BPH) compared with normal prostate, a cDNA library from BPH and a library from normal prostate could be constructed. The BPH phage library would be plated and transferred to duplicate sets of filters. One set of filters would then be probed with radioactively labeled cDNA from normal prostate and one set with radioactively labeled cDNA from BPH. Plaques which only hybridize to the BPH library represent genes that are expressed in BPH and not in the normal prostate. The individual plaques can then be retrieved from the agar, cultured in bacteria to expand the amount of cDNA available, and then sequenced to determine known or previously undescribed genes that are preferentially expressed in BPH compared with normal prostate cells.

A newer technique to identify preferentially expressed genes called differential display polymerase chain reaction (ddPCR) has been described by Liang and colleagues [19, 20]. In this technique, reverse transcriptase and "oligo-dT" which is a PCR primer with 12–18 thymidines in sequence which anneals or attaches to the long adenine tail of mRNA are used to transcribe cDNAs within two different tissues or cell lines. For this purpose, the oligo-dT primer ends in either a one- or two-based anchoring adenine, guanine, or cytosine so that a cDNA is created from only a subset of the genes expressed in the tissues. The second, or upstream primer is an eight-nucleotide random primer. Mathematically, a combination of these three primers and 80 of the 8 nucleotide random primers would amplify every possible cDNA in the human genome. Each pair of primers will amplify approximately 50–100 mRNAs. A PCR reaction including [$\propto$-^{35}S]dATP is performed using a single random primer and the same oligo-dT primer to amplify the cDNAs. They are then electrophoresed through a polyacrylamide gel to separate each cDNA, the gel dried, and exposed to radiographic film. cDNAs which are displayed at increasing or decreasing intensity between samples represent genes which are changing in expression (Fig. 2). These bands can be isolated from the gel, reamplified, cloned into a plasmid vector, sequenced, and searched against databases to see if they are known or unknown genes. Confirmation of their differential expression by northern analysis is usually performed. ddPCR frequently generates only a partial cDNA sequence. Therefore, if the gene is truly differentially expressed and not previously characterized, the partial cDNA may then be radioactively labeled and used to screen a cDNA library to obtain a full length cDNA for further characterization.

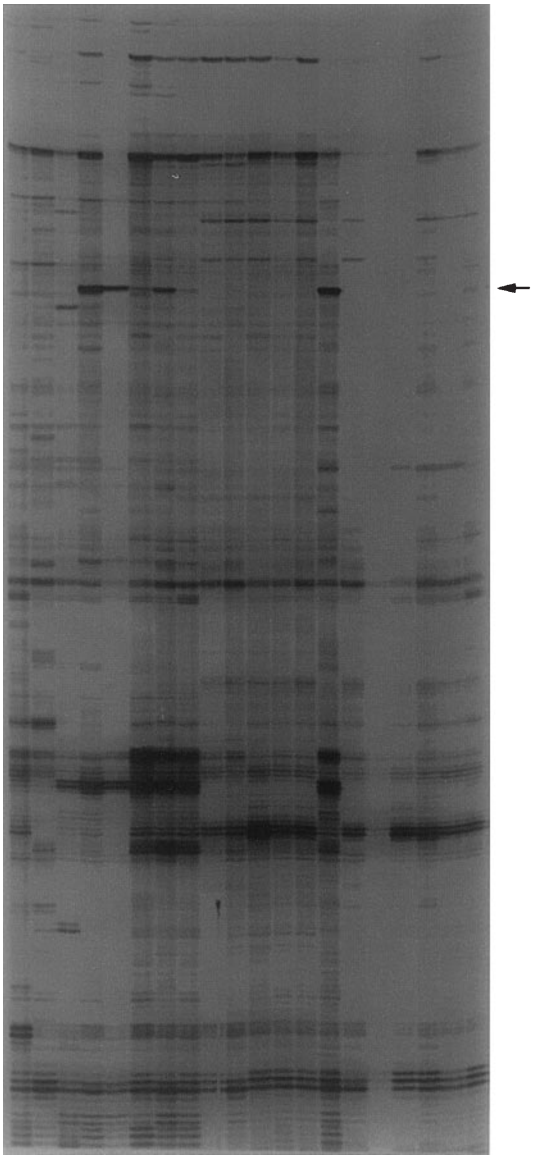

Fig. 2 Example of differential display PCR gel. The *arrow* indicates a gene with markedly different expression between tissues

More recently, the Cancer Genome Anatomy Project (CGAP) was initiated with the goal of achieving the comprehensive molecular characterization of normal, precancerous, and malignant cells. In the initial phase of the CGAP, efforts have been focused on lung, colon, ovarian, breast, and prostate cancer. cDNA expression libraries from microdissected tissues have been constructed and the sequences entered into public databases including dbEST. Several prostate libraries including normal prostate, BPH, prostatic intraepithelial neoplasia, and prostate cancer have been developed [16]. "Digital differential display" of gene expression may now be performed electronically through the CGAP

World Wide Web site to analyze changing expression patterns of specific genes between normal, malignant, and cancerous tissues. The techniques of ddPCR and differential EST analysis have recently been combined to identify genes involved in prostate cancer metastasis [23]. Utilizing this approach to compare several prostate cancer cells lines and normal prostate epithelium, transcription of the extracellular matrix protein hevin was downregulated in transformed prostate epithelial cells and metastatic prostate adenocarcinoma. This finding was confirmed by northern analysis and *in situ* hybridization studies. This type of technology represents a powerful new methodology for rapid identification of changes in gene expression which may prove important in our understanding of myriad urologic diseases.

Another strategy to identify new genes which are related to a known gene but unique to a specific tissue that may be a new member of a large family of genes also involves PCR methods. This strategy involves designing PCR primers based on a portion of sequence such as a phosphorylation or enzymatic site or region of a receptor that is consistently observed in the previously described family members. Alternatively, it may be a portion of sequence that is highly conserved between species. These primers are then used to amplify RNA or DNA from a new tissue or species. The PCR nucleic acid products obtained can then be cloned into a plasmid vector and sequenced to determine if they are unique. In many cases, an interesting gene may have been identified through studies in a different or less complex organism than humans. For example, a particular phenotype or characteristic may have been observed during the breeding of *Drosophila* (fruit flies). A mutation in a particular gene may then be identified that accounts for the observation. At that point, whether the gene is conserved and its function in other species such as humans becomes of major interest. In this way, a cDNA sequence identified from one species may be used as a probe to screen a cDNA library of another species.

Summary of gene hunting through specific target cloning

When clues to a candidate gene structure or function are available, they can be used to obtain more information from a cDNA library. All cDNA library screening methods require some unique information about the gene of interest to allow its isolation from the thousands of other genes expressed in the cell. This information may be as little as a small portion of the nucleic acid sequence obtained from ddPCR or based on a peptide sequence or as much as a complete cDNA sequence obtained from a different species which can be utilized as a probe for the cDNA library. Alternatively, at the protein level, a unique gene protein product epitope recognized by a specific antibody or a protein–protein interaction with another radioactively labeled known protein is required to successfully screen a cDNA library. After obtaining the cDNA the real characterization of the gene begins (Fig. 3).

Structural and functional characterization of a gene

Once a cDNA is isolated from a cDNA library, it must be sequenced in its entirety and the sequence compared against available nucleic acid sequences to determine whether the sequence has been previously reported or remains novel. The approximate expected full length cDNA size may be predicted from the cDNA length in another species and through northern analysis of 10–20μg of total cellular RNA which is electrophoresed and probed with the available partial cDNA. In addition, RNA is obtained from several other species and probed with the cDNA probe to determine whether the protein may be conserved between species. The cDNA sequence must be compared with a similarly isolated genomic DNA insert to identify intron/exon junctions. In addition, computer analysis to determine possible peptide sequences dependent on the translation reading frame may be performed to elucidate any protein motifs in common with other known proteins. Since protein translation is

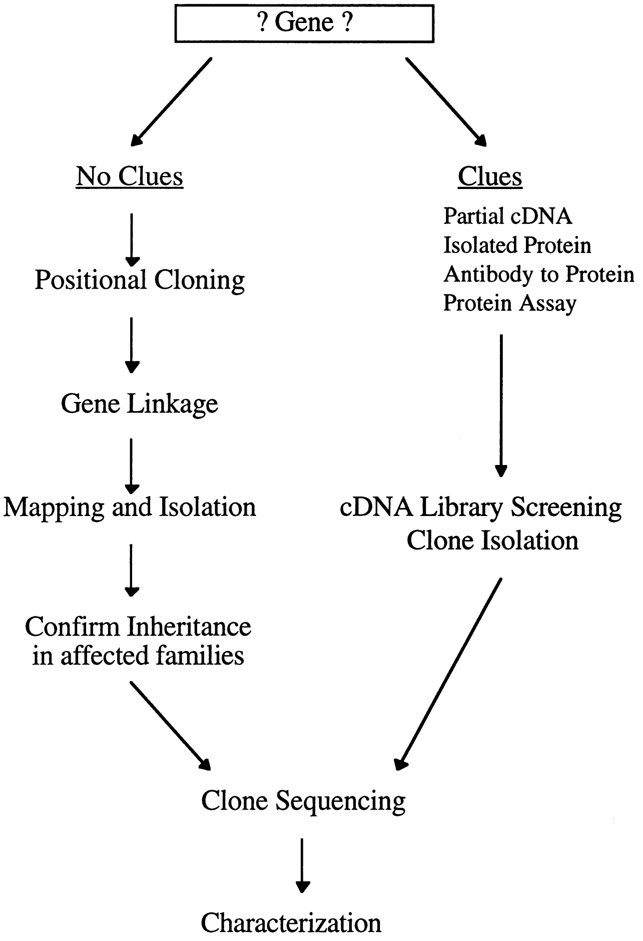

Fig. 3 Flow diagram summary of the primary steps in positional cloning and more immediately targeted approaches to gene hunting

Table 1 Selected Internet molecular biology sites

Cancer Genome Anatomy Project	http://www.ncbi.nlm.nih.gov/ncicgap/
BLAST Sequence Search	http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-last?Jform = 1
Foundation Jean Dausset CEPH	http://www.cephb.fr/bio/ceph-genethon-map.html
Human Genome Project	http://www.ornl.gov/TechResources/Human_Genome/home.html
I.M.A.G.E.	http://www-bio.llnl.gov/bbrp/image/image.html
Merck Gene Index	http://www.merck.com/!!uWJmV1oqauWJmq0oUl/mrl/merck_gene_index.2.html
Molecular Biology Computation Resource	http://mbcr.bcm.tmc.edu/databases.html
Multiple Sequence Alignments	http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-lign.html
National Center for Biotechnology Information	http://www2.ncbi.nlm.nih.gov/genbank/query_form.html
Sequence Utilities	http://dot.imgen.bcm.tmc.edu:9331/seq-util/seq-util.html
Stanford Human Genome Center	http://www-shgc.stanford.edu/
UK MRC HGMP-Resource Center	http://www.hgmp.mrc.ac.uk/homepage.html
Unigene	http://www.ncbi.nlm.nih.gov/UniGene/Hs.Home.html

initiated at a methionine (AUG) codon and because it is possible for a gene to have one or more translation start sites, a complete cDNA will necessarily contain one or more ATGs at its translation start site. Protein translations based on the ATG sites are performed to determine the longest possible reading frame before reaching a stop codon. It is quite apparent that ready access to a computer terminal with Internet capabilities has now become part of essential standard laboratory equipment. A list of several commonly used Internet sites for the modern laboratory is provided in Table 1.

Defining the structural characteristics of a gene is interesting and important, but the gene function must be delineated to ultimately determine its specific role in the developmental or pathologic process. Although a gene may appear to be primarily expressed within a specific tissue or process, upon further investigation it is usually expressed over a range of levels in several tissues or cell types. The timing (temporal) and location (spatial) expression of a gene can provide some initial clues to its potential function. Transfection studies of the gene into tissue culture cell lines can be utilized to investigate the effects of overexpression on numerous cellular processes including growth rate, differentiation, and tumorigenicity. Alternatively, antisense experiments which abrogate the function of the gene within cells may also suggest a gene function. Although many clues to gene function may be obtained from in vitro studies, in vivo experiments utilizing transgenic mice in which a gene is either overexpressed or "knocked out" in a living organism may be required to determine the ultimate functional role of a gene.

Conclusions

Through this review, the reader has been introduced to two general principles that are the basis of how researchers pursue new genes which may contribute to urologic disease: (1) a "shotgun" type of strategy called positional cloning or reverse genetics and linkage analysis which studies affected individuals or families to identify defective genes, and (2) an initially focused approach which uses available partial information about a gene to screen cDNA libraries toward the isolation and characterization of a gene. Although once entirely futuristic, glimpses of identifying patients with a genetic predisposition to urologic disease, predicting the probability and rate of further disease progression after a diagnosis has been determined based on molecular markers, and of treating urologic diseases through restoring normal gene function and/or regulation are now on the horizon. Although the majority of urologic surgeons will not be "gene hunters" who participate in the identification and characterization of these new molecular markers, they will play an important role in confirming the clinical utility of these new tests, explaining the implications to colleagues and patients, and in implementing new treatment techniques and strategies based on this information. Therefore, familiarity with these techniques and the promise they hold for the diagnosis and treatment of urologic disease is imperative for the modern day urologist.

References

1. Adams MD (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377:3
2. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651
3. Antiquera F, Bird A (1993) Number of CpG islands and genes in the human and mouse genomes. Proc Natl Acad Sci USA 90:11995
4. Ashley MV, Dow BD (1994) The use of microsatellite analysis in population biology: background, methods, and potential applications. In: Schierwater B, Streit B, Wagner GP, DeSalle R (eds) Molecular ecology and evolution: approaches and applications. Birkhauser, Basel, p 187
5. Baltimore D (1970) Viral RNA-dependent DNA polymerase. Nature 226:1209
6. Benson DA, Boguski MS, Lipman DJ, Ostell J (1997) GenBank. Nucleic Acids Res 25:1
7. Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST-database for "expressed sequence tags". Nature Genet 4:332
8. Collins FS (1995) Positional cloning moves from perditional to traditional. Nature Genet 9:347
9. Consevage M, Cyran S (1997) Basic elements of gene mapping and identification. Curr Opin Cardiol 12:288
10. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseas P, Marc S, Hazan J, Seboun E (1996) A compre-

hensive genetic map of the human genome based on 5264 microsatellites. Nature 380:152

11. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA (1996) Laser capture microdissection. Science 274:998
12. Fields C, Adams MD, White O, Venter JC (1994) How many genes in the human genome? Nature Genet 7:345
13. Green ED, Cox DR, Myers RM (1995) The human genome project and its impact on the study of human disease, 7th ed. McGraw-Hill, New York
14. Hillier L (1996) Generation and analysis of 280,000 expressed sequence tags. Genome Res 6:807
15. Kleckner N (1993) Meiosis: how could it work? Proc Natl Acad Sci USA 93:8167
16. Krizman DB, Chuaqui RF, Meltzer PS, Trent JM, Duray PH, Linehan WM, Liotta LA, Emmert-Buck MR (1996) Construction of a representative cDNA library from prostatic intraepithelial neoplasia. Cancer Res 56:5380
17. Latif F, Tory K, Gnarra J, Yao M, Duh F, Modi W, Geil L, Schmidt L, Zhou F, Li H, Wei MH, Chen F, Glenn G, Choyke P, Walther MM, Weng Y, Duan DR, Dean M, Glavac D, Richards FM, Crossey PA, Ferguson-Smith MA, Le Paslier D, Chumakov I, Cohen D, Chinault AC, Maher ER, Linehan WM, Zbar B, Lerman MI (1993) Identification of the von Hippel-Lindau disease tumor suppressor gene. Science 260:1317
18. Lennon G, Auffray C, Polymeropoulos M, Soares MB (1996) The I.M.A.G.E. consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151
19. Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967
20. Liang P, Zhu W, Zhang X, Guo Z, O'Connell RP, Averboukh L, Wang F, Pardee AB (1994) Differential display using one-base anchored oligo-dT primers. Nucleic Acids Res 22:5763
21. Macilwain C (1997) NIH and genome project set for more funds. Nature 388:316
22. Marshall E (1998) Physicists urge technology push to reach 2005 target. Science 279:23
23. Nelson PS, Plymate SR, Wang K, True LD, Ware JL, Gan L, Liu AY, Hood L (1998) Hevin, an antiadhesive extracellular matrix protein, is down-regulated in metastatic prostate adenocarcinoma. Cancer Res 58:232
24. Olson M, Hood L, Cantor C, Botstein D (1989) A common language for physical mapping of the human genome. Science 245:1434
25. Pruitt KD (1997) Webwise: navigating the human genome project. Genome Res 7:1038
26. Schuler GD (1996) A gene map of the human genome. Science 274:540
27. Simchen G, Hugeral Y (1993) What determines whether chromosomes segregate reductionally or equationally in meiosis. Bioassays 15:1
28. Smith JR, Freije D, Carpten JD, Gronberg H, Xu J, Isaacs SD, Brownstein MJ, Bova GS, Guo H, Bujnovszky P, Nusskern DR, Damber JE, Bergh A, Emanuelsson M, Kallioniemi OP, Walker-Daniels J, Bailey-Wilson JE, Beaty TH, Meyers DA, Walsh PC, Collins FS, Trent JM, Isaacs WB (1996) Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. Science 274:1371
29. Strausberg RL, Dahl CA, Klausner RD (1997) New opportunities for uncovering the molecular basis of cancer. Nature Genet 4:415
30. Temin HM, Mizutani S (1970) Viral RNA-dependent DNA polymerase. Nature 226:1209
31. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. Science 265:2088