Commentary

From mutation mapping to phenotype cloning

Jon J. Jonsson and Sherman M. Weissman

Department of Genetics and Boyer Center for Molecular Medicine, Yale University, New Haven, CT 06520-8005

Current methods of detecting known mutations in genes are efficient and have in many instances been incorporated into the armamentarium of diagnostic medicine. Many genetic diseases, however, including autosomal dominant diseases with high mutation rates and genetically lethal X chromosome-linked diseases, result from heterogeneous and often private mutations in the respective genes. Methods of identifying new mutations in genes are still technically demanding and lack sensitivity. This has limited their utility in phenotype–genotype correlations and in clinical medicine.

Short of sequencing, single-stranded conformational polymorphism and heteroduplex analysis are the most commonly used nucleic acid-based methods to map mutations in genes. They rely on detecting differences in electrophoretic mobility of gene segments carrying mutant versus normal sequences. Screening for mutations by selective cleavage of mismatched base pairs in heteroduplexes has also been attempted. Development of methods using chemical reactivity of mismatches with carbodiimide had limited success in part because the modification was not readily detected. Improved detection of mismatches using chemical cleavage was reported more recently (1) but sensitivity in detecting single base mismatches was still suboptimal (2). Early attempts to develop enzymatic methods for mutation detection using S1 nuclease to cleave single-stranded DNA in mismatched heteroduplexes (3) also lacked sensitivity in detecting single base-pair changes.

The above methods recognize mismatches by reacting preferentially with single-stranded DNA and may be confounded by transient separation of strands in double-stranded DNA (breathing), particularly in regions rich in A-T base pairs (4). The development of reagents with direct reactivity toward mismatches and with minimal reactivity toward transient single strands would have theoretical advantage. This has been attempted with DNA repair enzymes which recognize modified bases in mismatched DNA (5, 6). In this issue of the Proceedings Youil et al. (7) describe an advance in mapping of mutations, using T4 endonuclease VII for the enzymatic detection of mismatches in heteroduplexes comprising strands representing both normal and mutant alleles (7). The main function of endonuclease VII, produced by gene 49 of bacteriophage T4, appears to be resolution of branched DNA structures (8). Its other activities, including selective cleavage of double-stranded DNA containing mismatches, have also been carefully characterized (9). Youil et al. report that digestion with T4 endonuclease VII detected 17 out of 18 single base-pair mutations by cleavage of at least one strand in the heteroduplex; a set of A-A/T-T mismatched heteroduplexes was the only mutation not cleaved in either strand. Some nonspecific nuclease activity of endonuclease VII was noted, although it did not confound mapping of mutations. T4 endonuclease VII has great potential in mutation screening, but new and exciting genome applications of resolvases and other efficient mismatch detection enzymes loom on the horizon as well.

Recent advances in human genetics have been nothing short of spectacular in identifying genes responsible for monogenic traits. In particular, the isolation of genes of unknown function by deriving information about their position in the genome ("positional cloning") has been a powerful new approach to identifying genes associated with monogenic conditions (10). Many human traits, however, including most common diseases, have a complex etiology presumably reflecting interactions between one or several genes and the environment. Identifying genes associated with complex phenotypes by established positional cloning methods becomes progressively more difficult as the number of contributing genes increases and as the effect of each individual gene to the phenotype decreases. In addition, the task can be greatly complicated if the effect of each gene is modified by the genetic background in which it operates. Current methods in molecular genetics are still inefficient in identifying genes in these situations, especially when the pathophysiology is poorly understood and candidate genes cannot be readily identified. New approaches to directly identify genes without requiring knowledge of their function or position in the genome would be a crucial advancement toward understanding the genetics of complex traits. Several recent papers, including the one by Youil et al. (7), represent technological progress toward development of methods with just that capability.

The concept of identical-by-descent (i.b.d.) sequences was developed by early geneticists and its use in mapping genes—for instance, by sib-pair analysis—was recognized (11). The conceptual leap from relying on knowing either the gene's function and/or its genome position for its isolation to the direct experimental isolation or mapping of an unknown gene based on its phenotype alone was first made convincingly by Sanda and Ford (12). They presented three arguments in support of a novel approach of identifying genes directly by isolating i.b.d. sequences. (i) The genomic sequences from two unrelated individuals are different (13). (ii) Given low mutation rates, i.b.d. segments in genomes of two related individuals are identical in sequence. (iii) Through segregation and recombination of chromosomes, the i.b.d. segments become progressively fewer and shorter with increasing numbers of meiosis separating two relatives. The i.b.d. sequences should therefore contain genes contributing to a phenotype which is shared by distant relatives. i.b.d. sequences could in principle be isolated in the laboratory by mixing, denaturing, and annealing genomic DNA from affected and distant related individuals and enriching for heteroduplexes which contain no mismatches while discarding homoduplexes and heteroduplexes with internal mismatches (given that the genomic DNA is split in fragments long enough to contain sequence differences when inherited from different ancestors). Sanda and Ford attempted this approach to gene identification by modifying mismatched heteroduplexes with carbodiimide. Following cloning of heteroduplexes into plasmid, only unmodified plasmid—i.e., the ones containing i.b.d. sequences without mismatches—would replicate in bacteria defective in repair enzymes. The principle of the approach was sound but the technology at that time was not sufficiently advanced to allow efficient isolation of i.b.d. sequences.

The prospects for direct isolation of genes associated with a phenotype but of unknown function and genome position were revived by two recent papers describing major methodological advances in the field. Brown and associates (14) described the technique of genome mismatch scan-
oning (GMS), which embodies some of the principles outlined by Sanda and Ford. Brown and coworkers Dam methylated genomic DNA from one of two yeast strains and, after hybridizing the genomes of the two strains, enriched for hemi-methylated heteroduplexes by using their resistance toward methylation-sensitive restriction enzymes. Subsequently the duplexes were treated with MunHLS enzymes from Escherichia coli. These replication-associated repair enzymes introduce single-stranded nicks in the unmethylated strand (i.e., the nascent strand immediately after semiconservative replication) close to mismatched bases. Cleaved homo- and heteroduplexes were subsequently separated from intact heteroduplexes by their sensitivity to digestion with exonuclease III followed by selective depletion of DNA containing single strands. GMS resulted in a 20- to 100-fold enrichment of i.b.d. sequences after crossing of two different strains of the yeast Saccharomyces cerevisiae. The signal from radiolabeled products, when applied to an array of phage A clones from a library of yeast genomic fragments, was sufficiently strong to identify i.b.d. sequences in the majority of instances. One inherent strength of this approach is that the signal from many adjacent fragments is integrated to give a stronger reading.

Simultaneously with the development of GMS scanning, Wigler and associates (15) introduced representational difference analysis (RDA), a powerful technique employing subtractive hybridization. Prior to this study subtractive hybridization had been used successfully in human genetics to clone fragments contained in an intragenital deletion in the X chromosome in a boy with three genetic diseases: Duchenne muscular dystrophy, chronic granulomatous disease, and retinitis pigmentosa (16). Genomic DNA from this individual was used to deplete homologous sequences in genomic DNA from a human cell line with 49XXXXY karyotype and retain material greatly increased in fragments corresponding to the deleted region in the X chromosome. In RDA restriction enzyme digestion of genomic DNA is followed by addition of adaptors and polymerase chain reaction (PCR) amplification. Only short fragments will be amplified efficiently, so this step generates a simplified representation of the genome estimated to contain between 1.8% and 15% of the original genomic complexity, depending on the restriction enzyme used. After reannealing of two different genomes the generation of homoduplexes is dependent on relative concentrations of amplicons in both the target and competitive driver genomes. After hybridization the target homoduplexes are selectively PCR amplified. Repetition of this procedure results in efficient isolation of a restriction fragment represented in much greater abundance in tester amplicons than in driver amplicons. RDA was shown in the original article (15) to be able to selectively amplify sequences which were present at one copy per haploid genome in the tester genomic DNA but absent from the driver genome. RDA was also shown to isolate fragments generated by two restriction sites which are much farther apart in the driver genome than in the target genome, since PCR amplifies smaller fragments more efficiently then larger ones. This happens when a particular restriction site is missing from the driver genome due to a restriction fragment length polymorphism (RFLP), but genetic rearrangements including deletions, duplications, and translocations may also create this situation. RDA therefore has the potential of quickly identifying genes contributing to carcinogenesis by selectively amplifying genetic material corresponding to rearranged sequences in tumor cell DNA after its competition against the patient's constitutional genomic DNA.

RDA was subsequently shown to correctly map loci differing between congenic strains of mice as well as to map autosomal recessive (AR) phenotypes in a mouse strain by crossing that strain with another strain expressing the corresponding dominant (AD) phenotypes (17). Pools of F2 progeny expressing the AR phenotype contained AD genotypes throughout the genome except at the locus in question, where they contained only the AR genotype. Driver made from genomic DNA from F2 progeny expressing the AR phenotype therefore effectively competed against all target DNA from the AD parental strain except at the locus of interest, allowing RDA-mediated isolation of genetic material mapping to the locus of interest.

Recently, Rosenberg et al. (18) described another method, RFLP subtraction, designed to purify restriction fragments from a complex genome if they do not have a counterpart of the same size class in a competing genome. RFLP subtraction used gel purification rather than PCR for size fractionation of DNA fragments. The authors took a slightly different approach for selective amplification of fragments unique to the target DNA. Before PCR amplification they performed three repetitive cycles of annealing target with biotinylated driver and depleting biotinylated sequences with avidin binding. Once fragments unique to the target had been concentrated sufficiently, the method involved self-annaling of target sequences, depletion of unannealed sequences, and selective PCR amplification of target homoduplexes as in RDA. Impressively, after competing two different mouse strains the authors found that 21 of 22 clones of RFLP subtraction products represented unique RFLPs between the two strains and only one clone contained a repetitive sequence present in both strains. It is possible that the added steps in RFLP subtraction compared with RDA increased the power and specificity of the procedure.

One potential and very important application of emerging techniques in studying complex genomes is the direct isolation of disease genes based on linkage disequilibrium. The predominant reason for genetic markers being in linkage disequilibrium is thought to be identity-by-descent. In principle these genetic markers could be identified by isolating genome sequences, which are more often identical among patients with a specific phenotype than among random individuals. Two separate pools of genomic DNA, one from a group of patients and one from a group of unaffected individuals, would be created. Each pool would be denatured, reannealed, and subjected to mismatch cleavage followed by selective enrichment of intact duplexes. In both pools there would be variation in how polymorphic each genomic fragment is and in the recovery of intact duplexes. The products of each pool would therefore be likely to contain widely different amounts of the various genome segments. But there should be no difference in relative abundance of products of homologous fragments between the two pools except that the patient product should contain a much larger amount of fragments representing disease gene loci, since they were more likely to be spared from mismatch cleavage. The next step in the method would be subtractive hybridization using patient product as target and unaffected individual product as driver to get selective enrichment of fragments which are present in greater abundance in patient than in control individual products. The product of this procedure should be greatly enriched for DNA segments in linkage disequilibrium contributing to the phenotype. Careful consideration needs to be given to the quantitative differences in variation for this approach to work. The number of individuals included in each pool, length of the individual genome fragments, and the degree of disequilibrium expected are among the numerous quantitative aspects of the procedure that need to be addressed in the experimental design.

Although RDA has been proven to identify a specific locus in highly inbred mouse strains, the challenge is to develop methods that are applicable to highly outbred populations such as man and with sufficient resolution to uncover significant linkage disequilibrium in such populations. Limitations of current methods need to be addressed with the expectation that further steps in this approach to gene identification becomes experimentally feasible in human genetics, where the sequences in link-
age disequilibrium are expected to be much shorter than the i.b.d. sequences in experimental crosses of different mouse strains. Although GMS is effective in yeast it has not yet been shown to work on mammalian genomes, which are much more complex and contain repetitive sequences. In addition, the unique sequences in the human genome are less variable between individuals than yeast sequences. Recognizing mismatches in heteroduplexes by an enzyme such as T4 endonuclease VII, which can cleave both strands of mismatched DNA, offers an advantage over enzymes such as MutHLS, which only cut a single strand in a duplex, in that PCR amplification can be readily incorporated in the procedure to recover intact duplexes. Given recent progress with PCR such as stretch PCR (19) it is likely that PCR can be modified to generate a broader representation of a complex template mixture.

RDA and RFLP subtraction, by isolating small restriction fragments present in one genomic DNA but not another, test only a very small fraction of the genome in each experiment and therefore leave most differences between two genomes undetected. Making these techniques more sensitive to other sequence differences would be advantageous. Incorporating steps into methods based on subtractive hybridization to increase their power to amplify differences in abundance of sequences between two genomes as opposed to essentially detecting only sequences present in one genome but absent from another would be a necessary prerequisite for their use in identifying genes based on linkage disequilibrium. Improving techniques to analyze complex genomes to make them less sensitive toward repetitive sequences and other problem regions in the genome would also greatly increase their applicability.

We propose the term “phenotype cloning” to describe the isolation of genes by virtue of their effect alone and without requiring prior knowledge of their biochemical function or map position. Phenotype cloning, in contrast to positional cloning, has the potential to save a great deal of repetitive work involving analysis of multiple markers. The development of phenotype cloning could result in a dramatic reduction in the amount of effort involved in cloning disease genes and other genes, and this approach, by collectively testing the whole genome, could in theory have greater resolution and sensitivity than current approaches in positional cloning.

It is said of individuals either good or bad “You will know them by their fruits” (Matthew 7:16). What would be the impact on genetics if we could say the same about genes?