The Quest for Genes Causing Complex Traits in Ocular Medicine

Successes, Interpretations, and Challenges

Sudha K. Iyengar, PhD

Gene mapping and positional cloning have gained acceptance as state-of-the-art methods to identify molecules that cause common complex diseases. However, the use of specialized technology, varying study designs, and misconceptions about the role of novel findings in genetics has led to confusion among basic scientists and health care professionals alike regarding the importance of these findings in molecular diagnostics and individualized medicine. To alleviate this confusion, the successes achieved in the past few years in mapping of genes for complex traits such as age-related macular degeneration and glaucoma are interpreted in the context of the appropriate population biology framework. The current article veers away from propagating the overly simplistic belief of a linear relationship between a specific gene and age-related macular degeneration, particularly one that equates possession of a specific risk allele as the only precursor to end-stage disease. Ascribing predictive properties to a single gene without consideration of its network partners, timing of action, or environmental correlates argues for a static view of gene action. Modern viewpoints of the mechanisms of action of a gene are contextual and encompass more cohesive frameworks, ranging from the developmental timing of action, to the genomic and environmental milieu. In this regard, gene mapping studies that have been so immensely successful in the gene detection phase of a study provide biased perspectives on the importance of these genes and the corresponding risk alleles in the general population because of their limited sample size and constrained design. To move the field of gene-based diagnosis forward, it will be necessary to conduct additional cohort and longitudinal studies using the original gene finding studies as a knowledge base to develop predictive models. In summary, while we have achieved great successes in finding genes for complex traits, the application of these findings to clinical medicine is not straightforward. The key question of who will develop disease in the future remains.

Arch Ophthalmol. 2007;125:11-18

The merger of 2 fields, traditional epidemiology, which focuses on the distribution and determinants of disease in human populations, and genetics, the study of inherited mechanisms of disease, led to the development of a broader allied field that draws on the strengths of both these disciplines. The goal of investigations in genetic epidemiology is to identify genes and to study their mechanisms of action in populations; other closely connected disciplines include molecular genetics and statistical genetics, neither of which has firm borders distinguishing it from genetic epidemiology. The theoretical basis for the field of genetic epidemiology was developed in the early 1900s by R. A. Fisher with the unification of theories from quantitative and qualitative modes of inheritance.1,2

CME course available at www.archophthalmol.com

Until recently, when the human genome was more fully characterized,1,14 the pace of the investigations and identification of disease genes remained slow. In the
past, the most prominent successes were limited to disorders with mendelian inheritance patterns and strong familial risks (also described as high recurrence risks), where a single gene carried the bulk of the disease burden (eg, paired box 6 and aniridia type 2). In contrast, common complex disorders or multifactorial disorders are characterized by multiple genes and environmental factors contributing to their etiology. Two international, large-scale endeavors, the Human Genome Project (HGP)\(^3,4\) and the International HapMap Project,\(^5,6\) have accelerated the speed at which disease genes are being discovered for both rare and common disorders. The HGP was a large-scale enterprise to sequence the nuclear genome, find, and annotate all the possible genes in the genome. The project was completed in 2001 and a draft assembly of the human genome is available on a Web server (http://genome.ucsc.edu and www.ensembl.org).\(^3,4\) The HapMap Project took on the task of further characterizing the genome where the HGP left off. The goal of the HapMap Project is to determine all common genetic variation (both in genes and outside genes) in several different populations worldwide.\(^5,6\) These projects have been supported by technological and methodological advances that often require specialty knowledge to interpret results.

The goal of this review is to provide an explanation of the different approaches used in statistical genetics and genetic epidemiology for health care professionals and researchers studying eye disorders to assist in the interpretation of the rapidly changing and complex literature.

**BINARY TRAITS**

Very often, diseases are assumed to be the outcome of a singular event or a convergence of multiple events into a singular outcome, and individuals are classified as “with disease” or “without disease” using specific nosology. Thus, the presence or absence of the disease may aggregate in families as a binary trait. To quantify the extent of the familial aggregation, the recurrence risk ratio may be used as a measure.\(^7\) The recurrence risk ratio (\(\lambda_r\)) for a specific relative type is the ratio of the prevalence of the disease in the relatives of the index case to the prevalence of the disease in the general population; relationships must be specified because distant relatives share less of the genome than first-degree relatives. In the situation of many complex traits where extended relatives are hard to obtain, the sibling relative risk ratio is often used to determine if sufficient power is available to map disease genes.\(^7,8\) As an example, the recurrence risk ratios to siblings of cases with age-related macular degeneration (AMD) are projected to be 3- to 6-fold higher compared with the general population.\(^9\) Obtaining sibling recurrence risk ratios higher than 2 argues that a pattern of familial clustering over and above the risk in the general population is present. While genetic predisposition is certainly one explanation for familial aggregation, it is certainly not the only explanation for obtaining relative risk ratios higher than 2. Shared environment may also contribute to higher risks of disease. As an example, smoking increases the risk of not just the individual who smokes but may affect other members of the household, especially if they are also genetically susceptible. Furthermore, prevalence is difficult to estimate for many complex traits because obtaining an unbiased sample of the general population is quite often a difficult endeavor. As described later, studies for gene mapping are biased toward collection of enriched families for disease or toward identification of cases and controls who may not ideally represent the frequencies in the general population. While these maneuvers are advantageous in finding disease genes, overlooking the ascertainment bias will skew the results when determining the population attributable risk (PAR) due to a gene.\(^10\) The PAR quantifies the proportion of the disease burden that can be eradicated or ameliorated if a risk factor is removed from the population. Thus, ascribing a PAR of 50% will suggest that 50% of the disease burden can be eliminated by taking away the risk factor. Ideally, to determine the PAR accurately, one would need a sufficiently large sample that is representative of the general population and is collected without regard to disease status.\(^11\) This brings the genetics question back to a population level and to investigations in epidemiology.

**QUANTITATIVE TRAITS**

Assessment of a continuous measure, such as intraocular pressure or cholesterol levels, can be used to appraise the familial aggregation of a trait that may relate to the disease process. These measures, which correlate with the disease process but do not fully represent all facets, are sometimes called endophenotypes or intermediate traits.\(^12\) These traits are assumed to be more sensitive measures of salient aspects of the disease process and hence might be easier to map than the disease itself. Correlation coefficients can measure similarity in values of a continuous measure between 2 relatives. Similar to the concept of sibling relative risk, the variance in the phenotype for quantitative traits can be parsed into genetic and environmental components. These calculations of the heritability (the additive genetic component) would enable one to embark on molecular mapping studies.

Heritability is a population-specific concept and is often misinterpreted and misrepresented. Hypothetically, if every individual present in a population possessed 2 copies of a specific disease allele (homozygous for disease), the heritability for that gene in that population would be zero because there is no variation at that locus and it is not feasible to contrast individuals with variability in their genetic content at that locus. A measurable heritability (\(>15\%\) conventionally) suggests that a gene for that disease is segregating in the population and that a sample from the population can be subject to mapping experiments. Interpretation of heritability estimates assumes an expert knowledge of the methods used in the calculations and the disease itself. There are a number of meth-
ods to calculate heritability, but each method uses slightly different pieces of information, specifically the information on additive and dominance portions of variance. Additive variance can be explained at the level of the allele. The DNA sequence is made up of 4 bases: A, G, T, and C. The order in which the bases occur in the natural sequence can be determined through a technique called sequencing, as was done in the HGP. The sequence is not identical in all individuals but is subject to variation. The variant form of any part of the sequence is generically called an allele. Alleles can be simple changes in a single base (eg, A vs G) or can be fairly complex involving large segments of the DNA. At a particular locus with 3 alleles, A1, A2, and A3, each allele has a specific value it would impart to the phenotype. As an example, if one considers systolic blood pressure, the A1 allele may correspond to a value of 118 mm Hg, whereas the A2 allele may correspond with a value of 140 mm Hg and the A3 allele may correspond with a value of 110 mm Hg. Thus, an individual with a genotype of A1A1 would have an average reading of 118 mm Hg, while an individual with the A1A2 genotype will have a higher reading of 129 mm Hg. This hypothetical example demonstrates that each allele acts on a linear scale and additively. Dominance variance characterizes the interaction of alleles when the joint action deviates from the simple linear relationship. For complex traits, both additive and dominance effects at a locus play a role in disease etiology.

Both binary and quantitative traits have been used in ocular genetics, but the earlier focus has been predominantly on binary traits. This trend is now shifting to examine quantitative traits. While not often described extensively, ordinal traits that measure disease progression or severity via juxtaposition of multiple traits in a specific series can also be used as a proxy for a direct quantitative measure. Here, caution needs to be heeded when performing disease mapping experiments because the assumption is that all steps in the model are equal, unless the steps are weighted to reflect the underlying pathophysiology. So for a 5-step scale, the change from 0 to 1 is the same as the change from 1 to 2 and so on. As long as each increase reflects minute changes in the natural history of the disease, the method of using the contrived scale is valid. In fact, using a binary indicator in a quantitative trait linkage analysis (1 = presence of disease, 0 = absence of disease) is a specific case of this method.

Calculations of relative recurrence risks, heritability, or segregation analysis are traditional methods used to establish the feasibility of a genetic mechanism causing the disease and are the prima facie evidence that precedes gene mapping studies. Segregation analysis, a method for formal model fitting using phenotypic data in pedigrees, has been used sparingly since the advent of advanced molecular methods because it is time-consuming and prone to many uncertainties. The majority of these studies make no allusions to any specific genes, nor do they require molecular genetic data.

**LINKAGE STUDIES**

Linkage analysis has been extensively used in the mapping of genes for ocular disorders and has led to the identification of genes for both rare and complex ocular traits. Linkage has customarily been used for the mapping of rare disorders through the collection of larger families, with many individuals being “affected.” The minimum requirement is a single pair of affected siblings but can include more extended families, with second- and third-degree relatives contributing to the overall evidence for or against a particular gene or chromosomal region. The clinical data on affection status are contrasted between individuals who are affected and unaffected. In the same statistical test, the molecular contrasts are provided by markers (sequence variants) that can trace inheritance patterns in chromosomes between members of the same family. These markers can interrogate a single gene, an entire chromosome or a specific chromosomal location, or the entire nuclear genome. The mitochondrial genome has a maternal mode of inheritance and its own set of markers. The expectation is that if a gene for the disorder exists, then individuals in a family who were affected would preferentially inherit the affected portion of a chromosome from their parent(s). Chromosomes or pieces of chromosomes that were inherited equally by affected and unaffected members of a family (ie, segregating randomly according to the laws of Mendel) are not associated with disease.

The summary statistic describing evidence for or against linkage between a disease locus and a marker was developed by Morton and is called the LOD (logarithm of the backward odds) score statistic and can be described as the ratio of the likelihood for the data at a particular recombination fraction over the likelihood of the data under no linkage.

\[
\text{LOD score (Z)} = \log_{10} \frac{\text{Likelihood of observed data if loci are linked at value of } \theta}{\text{Likelihood of observed data if loci are not linked at value of } \theta = 0.5}
\]

where \( \theta \) is the recombination fraction.

The original method and many of its adaptations relied on assumption of specific modes of inheritance (eg, autosomal dominant, autosomal recessive) to test the hypotheses of linkage, the so-called model-based methods. Subsequently, model-independent methods were developed that used allele sharing between members of a family without making assumptions regarding the mode of inheritance. A LOD score threshold of 3 using the traditional model and 3.6 when performing a genomewide scan using model-independent methods is considered sufficient evidence to conclude that linkage exists between a chromosomal segment and disease. For multifactorial diseases, very often the burden of a LOD score of 3.6 is not met after performing a genome scan, resulting in reservations that genes for complex diseases can be mapped. This is exemplified by examining all the linkage scans for AMD. Each individually did not reach statistical significance (ie, a LOD score of 3.6 on...
1q31), but a meta-analysis showed that the LOD score on 1q31 was the second strongest signal across studies; the strongest was the locus on 10q. When the linkage scans were originally published, the use of a variety of linkage methods and the corresponding statistics (eg, LOD score, maximum LOD score statistic, non-parametric LOD score) that were used to achieve the same goal were often confusing to readers unfamiliar with the methods because the statistics reported did not appear to be directly comparable. However, most can be converted to a familiar P value through use of specific referent distributions.

Prior to the popularization of single nucleotide polymorphisms (SNPs) as the markers that are common, cosmopolitan, and easy to genotype, microsatellites spaced at even intervals on each of 22 autosomes and on the X chromosome were used for genomewide scans. Such scans have been performed for both monogenic ocular disorders and complex ocular traits and ranged in cost, density of coverage, and markers genotyped. The current SNP scans consist of approximately 6000 or more markers, with an average intermarker distance of about 0.64 megabase (Mb) (1 Mb = 1 million bases). Microsatellite scans were less dense (about 400 megawide) with an average intermarker distance of 10 centimorgans or roughly 10 Mb. The replacement of microsatellites with SNPs enabled greater automation and reduced genotyping error, but microsatellites, the majority of which were commonly used for genetic studies, had more than 2 alleles and carried more information than most SNPs. The increase in the density of SNP to microsatellite coverage is to partially compensate for the loss of information because the majority of SNPs are biallelic. The current maps are increasing in density as more SNPs are publicly made available through government-funded and commercial ventures.

A requirement for linkage analysis is a family unit with at least 2 or more affected members. Study designs have varied from collection of large families with many affected individuals (families with the greatest genetic load), to modestly sized families, to a single affected sib pair, to distant relatives in isolated populations. In some cases, homozigosity mapping for recessive diseases using pooling techniques has also been useful, but this method is unlikely to gain popularity for multifactorial traits because the assumptions being made when pooling samples are quite vital to the success of the experiment. Affected parent-offspring relationships were included in the traditional model-based methods for linkage but do not provide linkage information via model-free linkage methods. These relatives have other uses, such as relationship testing and fine mapping, and should be collected whenever feasible. Strategies for enrichment of disease-bearing individuals vary with the complexity of the disorder and the availability of family members for participation in the study. For example, if a disease has a late onset, then finding surviving relatives who reside in the same geographic region may prove to be a difficult task, especially in larger metropolitan cities in the United States. For a late-onset disease, obtaining truly unaffected individuals may also be difficult, especially for diseases where the possession of the disease allele is not commensurate with developing the disorder (variable penetrance).

One method to alleviate concerns of misclassification in affection status is to use an ordinal or quantitative measure that captures additional information beyond a simple description of affected and unaffected in families. Further, this clinical (phenotypic) information should be captured uniformly within and across the families, using epidemiologic principles of objective measurement and standardized techniques that have been validated, whenever feasible. For example, using intraocular pressure as an intermediate trait for glaucoma may provide additional information during the gene mapping process, while recognizing that normotensive glaucoma may also segregate in the family. Therefore, by mapping genes for only this feature, the full extent of the genetic variation that causes glaucoma will not be characterized. Difficulties in collection of enriched families owing to variable penetrance, complex modes of inheritance, and changes in the disease status with age are often cited as reasons why alternate methods to map disease genes (eg, association studies described later) are selected in lieu of linkage mapping. In these cases, collection of endophenotypes as surrogates for disease may provide some necessary information.

In summary, while quite successful for traits showing mendelian inheritance patterns, linkage analysis had not proven very tractable for most multifactorial diseases, and many ocular disorders and traits lag in gene identification. The difficulties in mapping these diseases have been owing to genetic heterogeneity (different loci segregating in different samples), modest sample sizes that have not been sufficiently large to extract the linkage signal (ie, small effect size of the disease gene), phenotypic heterogeneity, and epistasis (interaction between genes). Thus, gene discovery for most ocular traits is still in the early phase, the exception being AMD, which has had successful breakthroughs in disease gene mapping despite its complexity. Two loci, complement factor H and LOC387715, were identified as major candidate genes underneath linkage peaks on chromosomes 1q and 10q, following linkage experiments. More convincingly, the majority of the studies published have shown some support for both loci. Further, the linkage scans for AMD have been surprisingly concordant in identification of multiple regions likely to harbor a candidate gene beyond the loci on chromosomes 1 and 10, which showed the best evidence for linkage in a meta-analysis for AMD. Similarly, there seems to be a convergence of evidence on chromosomes 5q and 14q for glaucoma susceptibility, although a disease locus has not yet been identified. Therefore, the prospects of using family-based methods to map disease genes for ocular diseases are quite favorable.

With the advent of high-density genome scans for association, many skeptics of linkage scans have advocated abandoning linkage stud-
susceptibility. In contrast to linkage, association scans have been broadened to genomewide association testing. The family-based association analysis was developed by Spielman and colleagues to guard against population stratification. The latter method cleverly uses information from linkage and is the ideal middle ground between strictly disequilibrium test. It also retains information from linkage and is the ideal middle ground between strictly family-based linkage and case-control methods.

Association testing can either determine if a gene is involved in disease through interrogation of specific variants, such as if a sequence variant (eg, the Y402H polymorphism at complement factor H for AMD or specific variants in optineurin or myocilin) is responsible for the disease burden in a particular population or sample. Recently, this framework of scanning candidate genes has been broadened to genomewide association tests, whereby the entire genome is inspected at predefined intervals to determine if association with a genomic fragment can be established. The method to capture the association ranges from the standard fare in epidemiology, such as χ² tests and logistic regression modeling, to more sophisticated tests. As an example, a general test that takes advantage of some of the useful properties of family-based association designs has been proposed; it uses moving windows to examine consecutive SNPs genomewide. The family-based association testing methods have other advantages as well. A wider range of hypotheses can be tested, including parent-of-origin effects, skewing of sex ratios, and other complex modes of inheritance.

Many companies are offering genotyping products that interrogate 100,000 to 500,000 SNPs simultaneously, and denser scans are in production stages. This phenomenon, coupled with the cost reduction for genotyping, has led to a paradigm shift from linkage studies to association studies. Association studies are beguilingly portrayed as the panacea to gene mapping woes, but similar to linkage studies, association studies also have their own weaknesses. First, isolation of a genetic variant(s) that shows stark frequency contrasts between cases and controls does not provide sufficient grounds to assume a causative role for the variant. The variant may not be causal but simply in linkage disequilibrium (assort in the population on the same chromosomal segment) with the disease variant. Second, a more grievous error can occur when cases and controls are unintentionally drawn from genetically different populations or even subpopulations with distinct properties, such that a false-positive signal is generated because of confounding when conducting association tests. As described earlier, the transmission disequilibrium test was proposed as a safeguard against such false positives, but with a large number of statistical tests being conducted using dense genome scans, a proportion of such false positives is expected. Methods to protect case-control studies from confounding related to population stratification have also been proposed.

However, the best defense against false positives is replication of the experiment in a second independent sample. For most diseases, collection of a second sample is generally prohibitively expensive because, typically, collection of clinical and demographic data is more expensive. In an effort to alleviate some of these problems, efforts to support community collaborations are ongoing, with public availability of data sets gaining popularity, despite ethical dilemmas posed by the exposure of individual genetic data.

Convenience sampling of cases and controls from clinical practices can also have an effect on the ability to identify disease genes through association testing because of biased representation of individuals from the population. When a variant associated with disease is identified, this result may not be por-
table to other populations or even to other samples derived from the same population. Finally, the foundation for genomewide association studies is built on the assumptions proposed in a population genetic theory of dispersal of human populations more than 100,000 years ago. This theory is called the common disease common variant hypothesis, and its assumptions, advantages, and shortcomings have been reviewed previously. Its greatest weakness is that it may not have sufficient power to identify rare causative variants for disease. In this instance, family-based methods will prove more effective if the right types of families are identified.

As previously described, the environment also plays a significant role in complex diseases. Embracing this comprehensive view, the National Institute of Environmental Health Sciences has launched a project similar in scope to the HapMap project, the Genes and Environment Initiative. The goal of the initiative is to understand how the genome and environment parameters interact to cause disease. In the past, gene and environmental investigations were limited to a few genes or exposures. The scope of this project broadens the paradigm to genomewide investigations. In this scenario, epidemiologic studies (eg, cohort studies) have a very important role if biological samples are available for assessment of DNA variation. These can form the knowledge base of prospective cohort studies.

INTERPRETATION OF RESULTS OF LINKAGE AND ASSOCIATION STUDIES

An important issue germane to the discussion of genomewide linkage and association tests is that of multiple testing. The genome has so much variability at the level of the population that 100,000 to 1 million or more markers may be required to cover it adequately. Each of these markers is then assessed for its correlation with disease, creating a problem with the large number of statistical tests being conducted. Although, more methods with more liberal thresholds have been proposed to deal with the multiple testing problem, the larger problem of discriminating signal from noise still remains, especially in smaller data sets.

Interpretation of the results of linkage and association scans is not an easy endeavor particularly when investigators arrive at discordant results. Guidelines suggested for contextual interpretation of these studies (eg, the Lander and Kruglyak guidelines for interpreting linkage studies) have been misinterpreted as the gold standard and good studies abandoned when the experiment merely suggests that additional data were needed to meet the burden of proof. In the case where the results from several experiments (ie, multiple genome scans or association studies of a candidate gene) are not always concordant, the actual variants or markers, the samples, the population from which the sample derives, and the statistical and molecular methods used need to be carefully compared for similarity and dissimilarity. As in the case of calpain 10 and diabetes mellitus, an apparent controversy in the results across studies can be reconciled through joint analysis and meta-analysis. Calpain 10 is a gene associated with type 2 diabetes and was discovered through a linkage scan for type 2 diabetes in the Mexican American population. This gene was shown to be of biological importance, but the genetic data were not supported by replication studies. However, aggregation of data across multiple samples showed that there was soft evidence in multiple samples that did not meet statistical significance individually but cumulatively met the threshold for significance. Replication is a very important attribute of epidemiologic studies, and meta-analyses may be viewed as the mechanism to combine data across heterogenous samples when collection of a new cohort of sufficiently large size is not feasible.

CONCLUSIONS AND FUTURE DIRECTIONS

The current state of the art is multistaged designs that encompass data from linkage, association, and molecular experiments when feasible. Neither linkage nor association mapping require a priori knowledge of the function of the gene prior to embarking on gene mapping investigations, but some general knowledge about the potential function of an unknown gene is certainly an advantage. In the future, ocular genetic studies may rely on multicenter and community collaborations for mapping experiments.

Despite the optimistic forecast for gene mapping, many unresolved issues remain. Some of these issues are formulated into simple questions that deserve further scientific thought. Are most ocular diseases caused by a few common variants that can be used to predict disease status? Should variants that show association signals in non–gene bearing chromosomal regions be ignored? Will interactions between genes and between genes and environment mediate the bulk of the modifiable risk for ocular disease? The field of comparative genomics for complex ocular disorders is in its infancy. Do we anticipate that the allelic spectrum of mutations will encompass copy number variants? And most importantly, can the discoveries made at the bench be translated into therapies that cure inherited ocular disorders?

Submitted for Publication: September 6, 2006; final revision received September 28, 2006; accepted September 28, 2006.

Correspondence: Sudha K. Iyengar, PhD, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Bldg 1315, 10900 Euclid Ave, Cleveland, OH 44106-7281 (ski@case.edu).

Financial Disclosure: None reported.

Funding/Support: Dr Iyengar is supported by grants EY015814 (Fine Mapping of Genes for Age-Related Maculopathy) and EY016482 (A Multicenter Study to Map Genes for Fuchs Dystrophy) from the National Eye Institute.

REFERENCES


2. Bailer AJ, Piegorsch WW. Estimating integrals using quadrature methods with an application...