FINDING AND INTERPRETING GENETIC VARIATIONS THAT ARE IMPORTANT TO OPHTHALMOLOGISTS

BY *Edwin M. Stone MD PhD*

ABSTRACT

Purpose: To explore two approaches for making the human genome more accessible and useful to practicing ophthalmologists.

Methods: DNA samples were obtained from patients with inherited eye diseases, and these samples were screened for sequence variations in known disease genes with a combination of single-strand conformational polymorphism analysis and automated DNA sequencing. Data from this screening were then used to evaluate strategies for productively narrowing the sample space as well as for estimating the pathogenic potential of variations that were discovered in individual patients. For the latter purpose, a universal nomenclature for pathogenic potential was proposed based upon the segregation of disease alleles and the evolutionary conservation of specific residues as reflected by a substitution matrix known as blosum 62.

Results: Sequence variations were found to be unevenly distributed among disease-associated genes, such that screening strategies could be refined to discover more than 50% of clinically important sequence variations with only 10% of the effort. The use of the blosum 62 matrix was more statistically powerful than our previous method of estimating pathogenic probability.

Conclusions: The size of the human genome requires that clinical questions be very carefully focused if they are to be meaningfully answered in a reasonable amount of time and with a reasonable amount of resources. By examining the behavior of known disease genes, one can design strategies for significantly focusing the sample space and for more effectively interpreting the variations that are found.

Trans Am Ophthalmol Soc 2003;101:431-478

SUMMARY

The progressive recognition over the past 100 years of the "central dogma of biology" (that everything we see as the structure and function of an organism can be linked to the sequence of its genome) has had a profound impact on all branches of medicine, including ophthalmology. Today, the massive amounts of data streaming out of the human genome project promise to significantly improve our ability to diagnose, counsel, and ultimately treat our patients. However, sequence variations that cause eye disease are not uniformly distributed in the genome or in the population, and as a result, the yield of most experiments that are designed to find these variations is not linearly related to the time or resources that are consumed by the experiment. Moreover, many sequence variations in the human genome do not affect the structure or function of the individual that harbors them in any detectable way. Even variations that are observed in well-established disease genes do not always cause disease. A simple, broadly applicable method for sorting sequence variations according to the probability that they cause disease will be essential for us to be able to fully harness the power of the information carried within our genomes.

From the University of Iowa Carver College of Medicine, the Howard Hughes Medical Institute, Iowa City. This work was supported by the National Eye Institute, the Foundation Fighting Blindness, the Carver Endowment for Molecular Ophthalmology, Research to Prevent Blindness, the Grousbeck Family Foundation, the Knights Templar Eye Foundation, and the Howard Hughes Medical Institute.

HYPOTHESES

1. By examining only the portions of a sample space that are the most likely to harbor a desired finding, one can dramatically increase the speed and decrease the cost of identifying clinically relevant sequence variations.
2. Evolutionary information derived from a large number of proteins can be combined with information about the segregation of sequence variations in families to accurately estimate the probability that specific sequence changes observed in individual patients truly cause their disease.

INTRODUCTION

In the early days of the Molecular Ophthalmology Laboratory at the University of Iowa, I received a blood sample that had been drawn from a child suspected to have some type of inherited eye disease. Wrapped around the sample and held in place with a rubber band was a scrap of paper that said only: "Run the chromosomes." I think that the kindest translation of that terse message would be: "I believe that my patient's disease is caused at least in part by variations in her DNA, and I would like you to help me find these variations and use them to assist me in caring for my patient." It is instructive to consider the magnitude of this physician's request. The haploid human genome consists of over 3 billion individual units of information (nucleotides or base pairs), and for a simple mendelian disease, even a single variation at one of these 3 billion sites could be responsible for this child's disease. At each nucleotide position, there are four choices: A, T, C, and G. If every nucleotide in the haploid human genome were the size of a common penny, three billion of them placed side-by-side along the equator would circle the earth 1.5 times. Assuming that these pennies were all minted during a 4-year period, finding the single nucleotide change responsible for this patient's disease would be the same as circling the globe at the equator one and a half times looking for a variation in the year on one of the pennies.

To put this into an economic perspective, tabulating the sequence of "years" in this 3 billion "pennies" just once was the primary goal of the Human Genome Project, which took thousands of scientists, billions of dollars, and over 10 years to complete.[1,2] However, even with all the resources of the human genome project at one's command, it would still be impossible to answer the physician's question as I rephrased it. The reason for this is that there is an enormous amount of "normal" variation in the genomic sequence of human beings. Any two random individuals will differ from one another by at least one of every 1,000 nucleotides.[3] Thus, even if it were technically possible to compare the genomic sequence of this young patient to that of a "normal" individual, it would yield at least 3 million differences, any single one of which could potentially cause the patient's disease.

How then can clinicians possibly hope to probe the human genome in individual patients in a clinically meaningful way? A partial answer to this question can be found in a second example that was sent to me at about the same time as the first. In this case, I received a sample from a clinician on the island of Guam who told me that he had seen a 12-year-old boy who had had perfectly normal vision until about 6 months earlier, at which time the patient had suddenly lost the central vision in his left eye. Associated with this loss in vision was swelling of the optic nerve head on the affected side. The initial clinical diagnosis had been optic neuritis. However, approximately 3 months after the first episode, the patient suffered a similar episode in his other eye so that he then had 20/800 vision in both eyes. A blood sample was included in the same package as this clinical information, and I was able to use this sample to examine the status of three specific nucleotides in his mitochondrial DNA. Less than 24 hours later, I was able to confidently make a diagnosis of Leber hereditary optic neuropathy (a disease that occurs in less than 1 in 100,000 people per year) in a patient that I had never personally seen and that lived half the world away.

How is it that with similar samples (anticoagulated venous blood) and similar methods (polymerase-chain-reaction-based DNA sequence analysis), we were able to provide a very specific and clinically relevant answer for less than $200 in less than 24 hours in one case, while in another we could not have provided any clinically relevant answer even if we had billions of dollars and 10 years to devote to it? The simple answer is that there are orders of magnitude with more clinically relevant information in 85 words of clinical description than there are in 3 billion nucleotides of unfocused DNA sequence analysis. Does this mean that molecular biology has nothing to offer clinicians? Far from it. But, no matter how comprehensive the clinical information is from a given patient (or how recognizable the clinical pattern is), a number of questions will often remain about a patient with a rare eye disease when one is limited to only clinical information. How many clinicians have ever personally made the diagnosis of Leber hereditary optic neuropathy? Among those who haven't, how confident would they be in making that diagnosis in the clinic tomorrow? The bottom line of these examples is that, as in most other branches of medicine, it is the combination of clinical information and laboratory analysis that allows a physician to arrive at the correct answer in the shortest amount of time and with the least expenditure of resources. More specifically, carefully obtained clinical information can be used to focus a molecular question to the point that with very reasonable expenditures of time and other resources, one can provide a clinically meaningful answer that is often many-fold more specific than one could provide with clinical information alone. In the case of the "run the chromosomes" patient, the total absence of clinical information meant that the "sample space" in which the question was posed consisted of 3 billion nucleotides. In contrast, the excellent clinical description of the boy from Guam narrowed the sample space to only three nucleotides. This billion-fold reduction in the size of the problem was made possible by combining excellent clinical information with

a previously recognized "genotype/phenotype correlation."

Although using clinical findings to reduce the size of a sample space is something that clinicians do subconsciously every day, there are several features of genomic sample spaces that make them hard to generalize to one's experience with other types of clinical tests. The first is the sheer size of the potential sample space. There are approximately 2,500 different "blood tests" that are performed by the Pathology Department of the University of Iowa Hospitals and Clinics. Even if one ordered every one of these tests on a given patient, it would be equivalent to studying only three millionths of the patient's genome in an unfocused way. The second difficult feature of genomic data is its abstractness. Liver enzymes and serum electrolytes have a degree of concrete reality to a physician that a single nucleotide polymorphism 4,000 base pairs upstream from a given gene will never have—and yet any of these could be highly predictive of the presence of an important disease. Perhaps the greatest challenge with genomic sequence information, at least in the present decade, is the paucity of normative data and the sensitivity of the data we have to variations in ethnicity. As I will discuss more fully below, sequence variations that are very "abnormal" for one population can be quite "normal" for another.

These challenging features of the genomic sample spaces lead to a few practical rules for clinicians and basic scientists who want to venture there in search of answers about human disease.

1. One's question must be focused in order for it to be answerable with a realistic expenditure of resources in a reasonable length of time. The corollary to this rule is that the more a question can be focused, the faster and less expensively it can be answered.
2. One must always study a control group in exactly the same manner (same lab, same time, same methods) as a patient group; and one must take every reasonable precaution to limit (or at least control for) variations in ethnicity between these groups.
3. When trying to predict which sequence variations cause disease and which do not, one must decide upon the criteria that one is going to use for analyzing a given data set before examining the data set. As a corollary to this, one must not use a criterion that is based upon the data themselves unless one is certain that this criterion is statistically fair (that is, that it would not yield a result that is favorable to the hypothesis using random data). Ideally, one would use a standard set of criteria with inherent biological and statistical validity.

In this thesis, I will consider different ways that a human molecular genetic question can be productively focused, and I will present data to show the effects of this focus on the answers to specific questions. I will also propose a set of standard criteria that one could use to estimate the likelihood that a given sequence variation causes disease in a variety of experimental situations. However, before moving on to these topics, I will close this introduction by presenting an example of the ill effects of using the data themselves as part of the criteria for separating "real mutations" from "non-disease-causing polymorphisms."

Suppose that a group has just discovered a new gene and would like to determine whether it causes a specific rare autosomal dominant disease. They have available to them 200 patients that exhibit this phenotype and 200 control individuals selected from the same clinic population. The gene is very large and requires 50 different polymerase chain reactions (PCRs) to study a single individual. The investigators realize that many variations in the genome do not cause disease. They reason that disease-causing variations will be the ones that are present only in the patients (and are absent in the controls), while non-disease-causing polymorphisms will be likely to be distributed haphazardly between these two groups. Before conducting the experiment, they decide to define a disease-causing mutation as one that is present in patients but absent in the controls. A large experiment is then performed that analyzes all 50 parts of the gene in every individual (20,000 PCR reactions). Suppose the data from this experiment look like those of Table I. Each portion of the gene that is amplified by an individual PCR reaction is known as an amplimer. The amplimers that contain the "disease-causing variations" are shown in bold in the table; and, when these variations are summed, one finds that the patients exhibit 14 disease-causing variations while the control patients exhibit none (Total 1, Table I). Analysis of these data with Fisher's exact test reveals that the patients and controls have a significantly different number of disease-causing variations with a P value of $<.0001$.

What is wrong with this analysis? Although it is true that sequence variations that truly cause a rare disorder will be absent or seriously depleted in a control population of this size, one cannot use this as a criterion for selecting individual variants for contingency table analysis without biasing the outcome. The data shown in Table I were not in fact generated in a molecular genetic experiment of 400 individuals but were generated by a research assistant in my laboratory flipping a quarter twice for each cell in the table. The numbers reflect the number of "heads" that were obtained during these two coin flips. If one totals all of the results of this experiment without using the biased

TABLE I: COIN FLIP EXPERIMENT ANALYZED WITH FISHER'S EXACT TEST

AMPLIMER	PATIENTS	CONTROLS	AMPLIMER	PATIENTS	CONTROLS
1	1	2	26	1	1
2	0	1	27	2	1
3	1	1	28	0	2
4	2	1	29	2	2
5	**2**	**0**	30	2	2
6	1	1	31	2	1
7	1	1	32	1	1
8	0	1	**33**	**1**	**0**
9	2	1	34	1	1
10	1	1	35	0	2
11	1	1	36	1	2
12	0	0	**37**	**1**	**0**
13	1	1	**38**	**1**	**0**
14	1	1	39	0	1
15	0	2	**40**	**1**	**0**
16	1	2	41	1	1
17	**2**	**0**	42	1	2
18	**2**	**0**	43	0	0
19	1	1	44	2	1
20	0	1	**45**	**1**	**0**
21	0	1	46	1	2
22	2	2	**47**	**2**	**0**
23	1	1	**48**	**1**	**0**
24	0	2	49	0	2
25	1	1	50	1	2
			Total 1	**14**	**0**
			Total 2	**50**	**52**

"disease-causing variation" criterion, there were, as expected, 50 "heads" observed in the "Patients" column and 52 in the "Controls" column ($P = 1$).

There is a certain set of experimental conditions that increases the likelihood of making the type of error shown in this example. The first element of risk is the size of the gene, which increases the number of amplimers that could individually exhibit a skewed response by chance. The second characteristic is a sample size that is large enough to allow some variations to be observed but small enough that the variations that are observed are not seen in both patients and controls (that is, one of the populations will exhibit a small number of "positives," while the other population will exhibit zero). Finally, the gene must exhibit sufficient "allelic diversity" that many of the amplimers tested will have non-zero results. For a more complete discussion of this allelic diversity issue, readers are directed to an article by Webster and coworkers.[4]

Returning to the hypothetical example of the 50-exon gene, would it still be possible for several of the amplimers in the gene to harbor true disease-causing mutations while other amplimers harbored only non-disease-causing polymorphisms that would obscure a true association between the gene and the disease? If so, how could the investigators have detected a relationship between the gene and the disease without falling into this "circular argument"?

The answer is that the investigators should have chosen a criterion that was independent of the distribution of variations between patients and controls when deciding which variations should be considered disease-causing and which ones should be considered non-disease-causing polymorphisms. One such criterion that we have employed for many years requires that a sequence variation change an amino acid and further requires that this amino acid change would be expected to result in an alteration in the charge, size, or polarity of the translated protein.[5] If all of the bold rows in Table I had resulted from the application of this charge-size-polarity rule, then despite the fact that the overall number of "changes" in patients and controls were equal, the statistically significant result in "Total 1" would be a valid observation and would strongly support this gene's involvement in the disease in question.

It is notable (and confusing to many) that once a gene has been statistically significantly associated with a disease—using unbiased criteria that were selected before examination of the data (practical rule No. 3 above)—it *is* valid to use the distribution of an individual variation between patients and controls as one of the criteria for deciding whether that specific variation is likely to cause disease. In the latter situation, one is simply

correctly applying the general knowledge that a true disease-causing variation will not be equally distributed between patients and controls (and that a "highly penetrant" disease-causing variation will be present in controls at a much lower frequency than in affected patients). This is an "iterative" argument, not a circular one. This will be discussed more fully in the section entitled "Estimating Pathogenic Probability."

THE CONCEPT OF SAMPLE SPACE

In my laboratory, as in many others like it, the basic unit of work is the PCR amplification of a small bit of a single individual's DNA, followed by analysis of this amplified DNA for evidence of sequence variation. It is useful to think of this quantum of effort as a cube of unit volume: one amplimer of one gene screened in one patient requires one unit of effort. It is obvious to anyone that it requires the same amount of work to screen ten patients for variations in a single amplimer of a single gene as it does to screen one patient for variations in ten different amplimers of a single gene or one amplimer of one gene in ten patients. What is less obvious to most people when they are imagining the types of genomic experiments that they would like to do is the enormous effect on the volume of the sample space when one increases all three axes at once. That is, if it costs $5 to screen one patient for variations in one amplimer of one gene, how much does it cost to screen 400 patients for mutations in 20 amplimers in 200 genes? Eight million dollars!

Because of this geometric reality, most experiments have to be constrained to a very small number on at least one of these axes to make them achievable with a realistic amount of resources in a realistic amount of time. One way to do this is to consider one's primary goal in doing the experiment and selecting a point or two on one axis that is most closely connected to that goal, thereby converting an impossible three-dimensional problem into a tractable two-dimensional one. Another way to do this is to use existing knowledge and prior experience to rank the values of each axis to maximize the probability that their corresponding cells in the sample space will contain important findings. Returning to the $8 million experiment, if one is able to use prior experience to choose the 40 patients (from the 400), and the two amplimers (of the 20) and the 20 genes (of the 200) that are most likely to harbor a meaningful sequence variation, the cost of the experiment is reduced to only $8,000—a thousandfold reduction in cost. If on each of the axes, the desirable findings are clustered such that 50% of the findings are associated with only 10% of quantity represented by that axis (patients, genes, or amplimers), then one can find 12.5% of all there is to find in the total sample space by screening only the first thousandth of it! To put it another way, there is a 1,250-fold lower cost per mutation found in the "best thousandth" of the screening space than there is in the space as a whole.

Later, I will discuss the ways in which one can best choose how to constrain the screening space given the underlying purpose of the experiment, and I will also provide some experimental evidence that clinically relevant sequence variations are clustered in such a way that a 90% reduction of one of the axes of the sample space can be achieved with a reduction of only 50% in valuable findings. However, before discussing ways in which one might prioritize different goals, it will be useful to consider what some of those different goals might be.

WHY STUDY DISEASE GENES?

What things can one hope to accomplish by iteratively comparing the genomic sequence (genotype) to the clinical appearance (phenotype) of human beings?

1. One can identify genes that are important for the normal structure and function of the eye and whose mutation results in known eye diseases.
2. By thoughtfully interpreting these gene discovery data, one can identify some basic biological phenomena and/or pathophysiologic mechanisms that are currently unknown.
3. One can use the improved understanding of basic developmental and physiological processes to devise interventions for disease. Some of these interventions will be in the form of conventional treatments, such as small-molecule drugs or surgery, while others will be a new class of treatment in which artificial genes are actually expressed in a diseased tissue for the purpose of improving its function or protecting it from further injury. Once designed, these treatments can be tested in animal or in vitro models that are identified or created using the specific genetic information that is identified in the gene discovery step.
4. One can use the observed correlations between genotype and phenotype to help make diagnoses in individual patients with inherited eye disease. When these correlations are strong, we can predict specific outcomes with high accuracy many years before those outcomes occur. One can also use this type of information to give individual family members more accurate understanding of their risk for having an affected child. This information can in some cases be coupled with other interventions, such as preimplantation diagnosis[6] and in vitro fertilization, to allow couples with a very high risk of having a child affected with a very severe disease to have their own biological children with a risk of disease that is no higher than the general population.

5. One can sample large numbers of individuals with specific phenotypes to identify mechanistically homogeneous groups of human beings who can, in turn, be invited to participate in clinical trials of the treatments that are shown to be successful in the animal or in vitro models. These genotypically homogeneous populations can also be carefully studied from a clinical standpoint, and in some cases a specific natural history associated with the specific genotype can be elucidated.

With respect to this last point, I believe that genotypically homogeneous groups of patients will play an essential role in the development of novel therapies for the following reasons. First, it is well recognized that many important clinical entities, such as retinitis pigmentosa, age-related macular degeneration, and glaucoma, are genetically and pathophysiologically heterogeneous.[7-9] Thus, it seems unlikely that a single type of therapy would be effective against all the different forms of any of these diseases any more than a single type of antibiotic would be likely to be effective against all types of bacteria. Figure 1 extends this antibiotic analogy to illustrate the effect that mechanistic heterogeneity would have on the search for a new drug. In panels A and B, two antibiotics are tested sequentially against pure cultures of two different microorganisms. It is quite easy to see that each of these drugs is very effective against one of the microorganisms but not the other. If one mixes these organisms together before testing with the antibiotic discs (panel C), it is harder to see the therapeutic effect of either drug; and if this were one small part of a large screen for new drugs, one might decide that neither of these agents was sufficiently promising for further testing. If one mixes ten different organisms together (panel D), the efficacy of the two drugs that was so apparent in the first two panels is completely obscured. Having access to molecularly homogeneous patient populations is analogous to having pure cultures of bacteria to test antibiotics against. Of course, we hope that any new drug will be broadly efficacious, but it seems unwise (and unnecessary) to count on this broad efficacy when doing so might cause us to miss the value of an important class of new compounds.

The other very real value of genotypically homogeneous patient populations is the ability to control the timing of a therapeutic intervention as well as the timing of the assessment of its effect. Consider the hypothetical disease process shown in Figure 2. Knowledge of the natural history of this specific genetic subtype of disease allows us to predict the kinetics of visual loss in patients who are not yet severely affected. Suppose that an animal model of this disease suggests that treatment anytime during its course will arrest the disease process. Suppose that business considerations of the drug manufacturer require that efficacy be demonstrated in less than 3 years with the smallest number of patients possible. What patients should be included in the trial? If patients are included who are 3 or more years younger than the age of the first downward inflection of the vision curve (or older than the age of plateau of the vision curve), there will be no demonstrable effect of the drug (even if it is 100% effective) during the course of the 3-year study. In contrast, selecting a group of patients who are just beginning to lose vision (and who are known to have the same molecular form of the disease as the animal model that benefited from the treatment) will give the maximum possible difference between patients and controls and thus be able to detect the efficacy of the drug in the shortest time possible with the smallest possible patient group.

PRIORITIZING GOALS FOR INDIVIDUAL EXPERIMENTS

Many of the foregoing reasons for collecting and analyzing genotype and phenotype information overlap one another within families, laboratories, and clinics. This overlap tends to obscure a fairly important fact: the different groups that are interested in one or more of these applications of genomic information tend to prioritize these applications in dramatically different ways. To a significant degree, this overlap of goals is a good thing because there are many opportunities for synergy among hypotheses, investigators, families, and funding agencies. However, something that is less obvious but very real is that, given any amount of resources, these very desirable goals are in competition with one another. In order to make progress toward any one of them, one must, to some extent, limit progress toward another.

If one were in the position of deciding how to allocate resources (time, personnel, equipment, reagents) to achieve the various outcomes listed in the preceding section, how would that best be done? Would one try to divide the resources equally among the different goals? Would one try to favor the needs of the individual over the needs of society, or the reverse? Would one try to consider the way that these individual goals are interrelated so that one doesn't discover hundreds of genes without ever solving the problem of delivering gene-based diagnosis or therapy to patients on a national scale? I raise these philosophical questions to make the points that (1) there are clearly definable constituencies with clearly definable (although not always articulated) goals; and (2) because resources are limited, these goals are in competition with one another.

The importance of these points in the context of this paper is that, at the level of an individual experiment, one must clearly define a goal in order to maximize the likelihood of achieving the goal and of achieving it most

efficiently. In my experience, it is often preferable to perform two different experiments that are each aimed at achieving the goal of a specific group than it is to perform some type of hybrid experiment that is designed to try to achieve the goals of two competing entities at the same time.

To explore this issue of competing goals a bit more specifically, it may be useful to consider three different "interest groups," each with a specific size, composition, goal, and set of resources. The first is an individual patient with a rare inherited eye disease. The second is the group of all patients and families affected with this disease as represented by a private foundation dedicated to supporting research into this disorder. The third is the population of a large country or perhaps society as a whole as represented by a federal funding agency that is administered by the government. How might these constituencies prioritize the valuable possibilities of genomic experiments? What effect would this prioritization have on the design of specific experiments?

First, consider the parents of an infant with a recent clinical diagnosis of the autosomal recessive condition known as Leber congenital amaurosis (LCA). These parents have as much interest in the overall progress of medicine as anyone, but right now their interest is primarily in their child. They would like a molecular diagnosis right away to confirm the clinical impression of LCA, and they are asking their genetic counselor whether it will be possible at some point for them to have preimplantation diagnosis to avoid having a second affected child. Although they would never phrase it in this way, the genomic challenge from their point of view is: "How can you find the mutations that are responsible for my child's disease as rapidly and inexpensively as possible?"

At the present time, there are six genes that have been convincingly shown to cause the LCA phenotype in humans: guanylate cyclase, RPE65, RPGRIP, CRX, CRB1, and AIPL1.[10-16] A screen of the entire coding region of these six genes by either automated DNA sequencing or single-strand conformation polymorphism analysis (SSCP) requires the study of 108 different PCR products (amplimers). How should one proceed to get the answer that this family wants as rapidly and inexpensively as possible? Should the genes be screened one at a time, or all six at once? If screened sequentially, which one first? The largest? The smallest? In alphabetical order? Should the entire coding sequence of every gene be screened?

Figures 3 and 4 are interesting to consider in the context of these questions. Figure 3 shows all of the sequence variations with a reasonable likelihood of causing disease (ie, with an EPP of 2 or 3; see section "Estimating Pathogenic Probability," that follows) that we detected in the coding sequences of these six genes in a cohort of over 300 LCA probands (the mutations themselves are tabulated in Appendices E through J). In Figure 3, the genes are arranged in alphabetical order, and within each gene the amplimers are arranged as they are in the genomic sequence, with the 5′ end of the gene to the left. The value depicted on the y-axis is the total number of potentially disease-causing sequence variations found in each individual amplimer during the screen of the entire cohort. Figure 4 shows the same data displayed in a different way. In this case, the amplimers have been ranked from left to right according to the number of potentially disease-causing variants that they were found to contain. That is, the amplimer containing the highest number of total variations (13) is placed at the left, and the 32 amplimers that did not harbor a single variation are placed to the right.

Assuming that the new patient in our current example was drawn from a similar population as the cohort that we have already screened, each y-axis value (mutations per amplimer) of Figure 4 is proportional to the probability of finding one of the patient's two mutations in the amplimer with that y value. Careful inspection of this figure shows that the ten amplimers that are most likely to contain a mutation are nearly equally distributed among four different genes. Also, one can see that these 10 amplimers (which comprise only 8% of the total coding sequence of these six genes) contain nearly half of all the mutations that we have found to date. To put these graphs into the context of the preceding "sample space" discussion, the fact that a single patient is being tested converts the screening problem from three dimensions to two. Then, by sorting the amplimers according to the probability that they will contain this patient's mutation, it allows one the opportunity of discovering 50% of all there is to find with only 10% of the effort of a "complete screen."

For those who feel a compulsion to "be complete," it is important to recognize the unfortunate truth that screening the entire coding sequences of these six genes by SSCP does not in any way represent a truly complete mutation screen of all LCA genes. To begin with, these genes are collectively responsible for only about a third of all cases of LCA,[16] and even within these genes, there are undoubtedly mutations that exist in the promoter and other noncoding regions that would not be detected by a PCR-based assay of the coding sequence.

The red line in Figure 4 depicts another meaningful value, the cost of finding each mutation in each amplimer. This is obtained by dividing the cost of screening the entire cohort with each amplimer, by the number of mutations found in that amplimer. Looking at the data in this way, one can see that it costs very little to find mutations in the first 10 amplimers (at the left of the figure), while the cost per mutation is so high that we cannot even

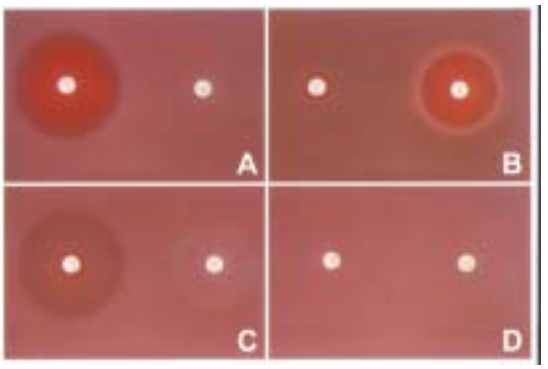

FIGURE 1
Antibiotic susceptibility analogy for the value of homogeneous populations for drug discovery. Panels A and B show the effects of two different antibiotics on the growth of two different pure bacterial cultures with different antibiotic susceptibilities. When these bacterial cultures are mixed before subjecting them to antibiotic susceptibility testing (panel C), the effects of the antibiotics are more difficult to discern. When ten different bacterial cultures are combined before testing (panel D), the efficacy of the antibiotics is undetectable even though 10% of the bacteria are highly sensitive to each of the two antibiotics (see text).

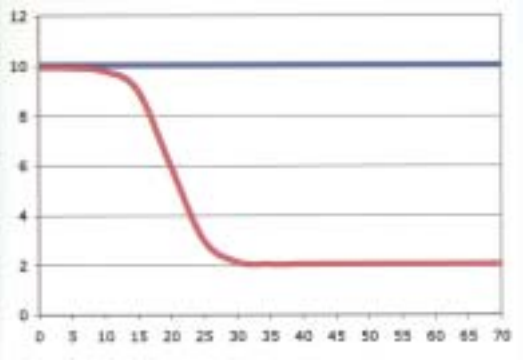

FIGURE 2
Importance of timing for assessing intervention. A measurement of visual function is given on the y-axis and decreases with age in this hypothetical disease (red curve). Intervention before vision loss can result in prevention of disease (blue line). Treatment at age 5 and assessment at age 8 or treatment at age 35 and assessment at age 38 would detect no benefit, whereas treatment at age 15 and assessment at age 18 would detect a large benefit (see text).

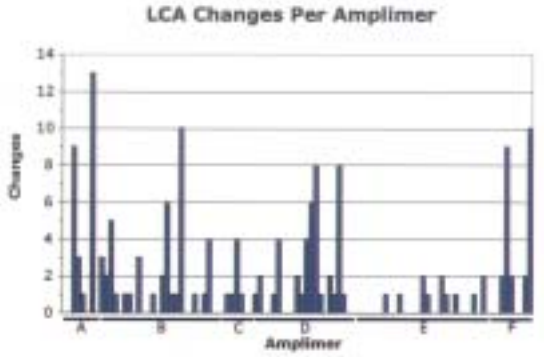

FIGURE 3
Results of screening a six-gene sample space in over 300 individuals. The coding regions of six genes involved in Leber congenital amaurosis are represented in genomic order along the x-axis (each gene depicted as a bar: A-AIPL1, B-CRB1, C-CRX, D-RetGC, E-RPGRIP1, F-RPE65). The number of sequence variations with a reasonable likelihood of causing disease (ie, with an EPP of 2 or 3; see text) found in each individual amplimer during the screen is given on the y-axis.

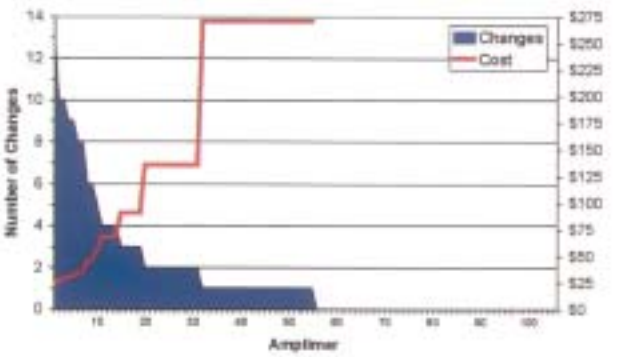

FIGURE 4
Focusing the sample space. The data from Figure 3 are now displayed with the amplimers ranked according to the number of variations detected within each amplimer during the initial screen. Displayed in this way, the graph illustrates that the likelihood of finding a mutation diminishes as more amplimers are screened. More than 50% of the variations are concentrated in less than 10% of the sample space. The red line shows the increasing cost of finding each new mutation as the screen proceeds into regions of the sample space with less and less likelihood of harboring a variation (see text).

estimate it for the amplimers to the right. Suppose the parents are willing to pay $500 for the 17.5% chance of finding a discoverable mutation in the first 10 amplimers (ie, 50% of the 35% maximum chance of finding an LCA mutation with our current level of knowledge). Would they be willing to pay $5,000 for the remaining 17.5% opportunity that is present in the next 50 amplimers? Would they be willing to pay another $5,000 for the less than 1% chance in the remaining 40 amplimers? Should every patient be screened on every amplimer even if two high-probability mutations (EPP of 2 or 3) have been found on separate alleles of the same gene early in the screening process? That is, just to make sure that nothing is missed, should every patient have a $10,000 test? After six more LCA genes of equal size have been discovered, should every patient have a $20,000 test just to make sure that nothing is missed? I think not. As I will explore more fully in the discussion section of this paper, I believe that to maximize the benefit derived from genetic testing, we should tailor the tests that we perform to the specific research or clinical situation. The main points to be made in this section are that (1) 100% diagnostic certainty is not attainable with this type of testing and (2) trying to attain 100% diagnostic certainty for large numbers of people will

be prohibitively expensive and time-consuming.

Before leaving this "single patient" example, it is interesting to consider whether all genes or diseases are equally susceptible to the type of focusing of the sample space that is depicted in Figures 3 and 4; and, if not, what the properties of the genes and diseases are that allow such focusing to be performed. In our experience, genes are not equally amenable to focusing of the sample space. Some of the reasons for this (size and allelic diversity) are intrinsic to the gene itself, but a significant additional factor is a clinical one. Specifically, the number of genes that need to be considered in any given clinical situation varies widely depending on the phenotype of the patients. If a clinician is sure that a patient is affected with LCA, the screening problem can be limited to the genes shown in Figures 3 and 4. If recessive retinitis pigmentosa is also being considered, a much larger (and intrinsically less focusable) screening space will be involved. For discussion and decision-making purposes, I find it useful to divide screening opportunities into nine different categories based on the results of the initial screening of each gene in a series of different phenotypes. The nine categories can be visualized as a "three by three" table (Table II) with three divisions on the "clinical axis" and three divisions corresponding to patterns of sequence variations on the "gene axis."

For this purpose, the three clinical categories are defined by the product of two values. The "specificity" of a given phenotype (which allows one to narrow the list of genes that need to be screened on clinical grounds) is multiplied by the fraction of that phenotype caused by detectable mutations in the gene (or group of genes) in question. These three categories are named by letters corresponding to the following products: (A) >75%, (B) 10% to 75%, and (C) <10%. The other characteristic of the specific genotype-phenotype relationship is the mutation distribution among the genes themselves. Again we can recognize three categories, this time named with numbers:

1. Five or fewer detectable amin-acid-changing sequence variations in the specified gene are responsible for >90% of patients with the specified phenotype that is attributable to that gene.
2. >5 but <50 sequence variations are collectively responsible for >90% of the disease attributable to that gene or genes—or any number of variations occur that are restricted to <30% of the coding sequence of the gene(s).
3. >50 different amino-acid-changing variations are responsible for <90% of the disease associated with these genes. These changes are distributed throughout the gene(s).

Table II provides examples of some specific genotype-phenotype relationships that we have observed and how they fall into these nine different categories. The A1 corner of the table contains gene-phenotype pairs that provided the most useful clinical information per unit of effort, while the C3 corner of the table contains gene-phenotype pairs that are the most expensive (in time and other resources) to screen. Table III provides even more detail about the mutation screening assays of 32 genes that we have screened in large numbers of patients and gives the screening category for each of these genes.

The second "interest group" that should be considered is a private foundation dedicated to finding a cure for LCA. What is their interest in genomic screening likely to be? Given the promising results of gene therapy in the Briard dog model of LCA reported by Jean Bennett, Al Maguire, Sam Jacobson, and their coworkers,[17] it would be very reasonable for a private research foundation to be interested in identifying human beings who might be willing to be subjects in a clinical trial of RPE65 gene therapy. With limited resources available for this genotyping, they might want to restrict their efforts to the nine RPE65 amplimers with the greatest likelihood of harboring a mutation. Moreover, since the initial clinical trials will probably involve individuals over the age of 18, they might want to further restrict their screening to patients in this age group. To put this in a "sample space" perspective, limiting the experiment to one gene reduces the problem from three dimensions to two. Further restriction of the experiment to only nine of the 14 amplimers of this gene and to only patients between the ages of 18 and 45 with the clinical diagnosis of LCA would allow this foundation to screen every such patient in the United States for about $60,000.

Finally, consider a federal funding agency supported by taxpayers' money and monitored by elected government officials. This organization is charged with the responsibility of advancing science and medicine as a

TABLE II: GENOTYPE/PHENOTYPE RELATIONSHIPS

	A	B	C
1	EFEMP1 Malattia Leventinese		EFEMP1 macular dystrophy
2	VMD2 Best disease	CHM Choroideremia	GLC1A primary open-angle glaucoma
3		ABCA4 AR Stargardt disease	ABCA4 cone-rod dystrophy

TABLE III: MUTATION SCREENING ASSAYS FOR 32 GENES ASSOCIATED WITH INHERITED EYE DISEASE*

GENE	EXONS			AMPLIMERS FOR EXONS SCREENED				DISORDER	CAT	TABLE
	C	NC	SCR	T	N	S	A			
ABCA4	51	0	51	51	0	17	34	AR Stargardt	B-3	Appendix L
								AR cone-dystrophy	C-3	-
ACHM2	7	0	7	15	2	9	4	Achromatopsia	B-2	-
ACHM3	18	3	15	16	0	2	14	Achromatopsia	B-2	-
AIPL1	6	0	6	9	0	2	7	Leber congenital amaurosis	B-3	Appendix E
BigH3	17	0	2	2	0	0	2	Corneal dystrophy	B-2	-
CHM	15	0	15	17	1	-	-	Choroideremia	B-2	-
CLN3	15	0	1	1	0	0	1	Batten disease	A-1	-
Col2A	54	0	52	52	-	-	-	Stickler syndrome	B-3	-
CRB1	12	0	12	27	0	2	23	Leber congenital amaurosis	B-3	Appendix F
CRX	3	0	3	7	0	0	7	Leber congenital amaurosis	B-3	Appendix G
								Cone-rod dystrophy	C-3	Appendix G
EFEMP1	12	2	1	1	0	0	1	Malattia Leventinese	A-1	-
ELOVL4	6	0	6	6	1	0	5	AD Stargardt	A-1	-
GLC1A	3	0	1	4	0	1	3	Juvenile open-angle glaucoma	A-1	Appendix K
								Glaucoma	C-2	Appendix K
GUCY2D	19	1	19	22	0	2	20	Leber congenital amaurosis	B-3	Appendix H
Mito 3460	1	0	1	1	0	0	1	Leber hereditary optic neuropathy	A-1	-
Mito 11778	1	0	1	1	0	0	1	Leber hereditary optic neuropathy	A-1	-
Mito 14484	1	0	1	1	0	0	1	Leber hereditary optic neuropathy	A-1	-
Myo7A	47	2	49	49	5	3	41	Usher syndrome I	B-3	-
NDP	2	1	2	4	1	0	3	Norrie disease	A-2	-
NR2E3	8	0	8	9	1	3	4	Enhanced S-cone syndrome	A-2	-
OPA1	28	1	28	28	1	1	26	Dominant optic atrophy	B-2	-
RDS	3	0	3	7	0	1	6	AD retinitis pigmentosa	B-2	Appendix C
								Pattern dystrophy	B-2	-
RHO	5	0	5	9	0	0	9	AD retinitis pigmentosa	B-2	Appendix B
ROM1	3	0	3	6	-	-	-	Pattern dystrophy	C-2	-
								Retinitis pigmentosa	C-2	-
RP1	3	1	1	1	0	0	1	AD retinitis pigmentosa	B-2	Appendix D
RPE65	14	0	14	14	0	0	14	Leber congenital amaurosis	B-3	Appendix I
RPGrip	23	1	24	32	1	10	21	Leber congenital amaurosis	B-3	Appendix J
RS1	6	0	6	6	0	0	6	X-linked retinoschisis	B-2	-
TIMP3	5	0	4	5	1	4	0	Sorsby dystrophy	B-2	-
USH2A	22	1	23	23	1	0	23	Usher syndrome II	B-3	-
VHL	3	0	3	5	1	2	2	Von Hippel Lindau	B-2	-
VMD2	10	1	10	12	0	4	8	Best disease	A-2	-

*Various features of genomic structure, distribution of known mutations, reliability of single-strand conformation polymorphism (SSCP), and clinical utility (see Table II) of 32 assays are summarized. Genomic structure and distribution of known mutations are summarized by: C=coding exons; NC=noncoding exons; SCR=exons we have chosen to screen. The structure and robustness of the SSCP assay is summarized by: T = total amplimers attempted (~200-300bp in size each); N = the subset of "T" that could never be amplified despite redesign of primers and multiple buffer and polymerase chain reaction (PCR) conditions; S= the subset of "T" that would "sometimes" yield useful data; A = the subset "T" that "always" amplified and yielded useful data. In the disease column, AD = autosomal dominant, AR=autosomal recessive. CAT= category of genotype/phenotype relationship as derived from Table II and described in the text.

whole (and at some level of doing the most good for the most people possible). Although this agency is undoubtedly interested in gene therapy of LCA (a disease that affects fewer than one in 10,000 people) to some degree, they are also very interested in age-related macular degeneration (AMD) (which affects 1 in 3 people over the age of 75)[8,18-21] and glaucoma (which affects 1 in 30 people over the age of 40)[22]. All other things being equal, this agency will probably be more interested in spending their resources for experiments designed to find new genes responsible for AMD and glaucoma than for experiments designed to find additional individuals with mutations in known LCA genes. For the purpose of this discussion, their goal might be summarized as, "We would like to gather statistically significant evidence for the involvement of a number of novel genes in AMD, glaucoma, and photoreceptor degeneration while consuming the smallest amount of resources possible for each individual discovery." Given the uneven distribution of mutations seen in the genes that cause LCA, how might one focus a

candidate gene screening experiment to find genes for AMD or glaucoma?

To develop a strategy for this ambitious goal, it seems reasonable to assume that a strategy that would work for AMD would also work for glaucoma or the photoreceptor degenerations group (PDG). Thus, we can devise one strategy and simply execute it three times to accomplish the overall goal. The next step would be to try to limit and/or rank each of the three axes of the sample space in an effort to make the three problems tractable.

Before considering which of the three axes is most or least amenable to constraint, one must realize that there is a major difference in this example when compared to the first two: the paucity of genotype/phenotype information at the beginning of the experiment. That is, in the first two examples, it was already known that a set of genes was associated with a given phenotype and the problem was how to find additional examples of these correlations within a new patient group. To put it another way, the first two cases represented a multistep process in which the gene discovery steps had already been done. In this third case, the main goal of the project is to accomplish the gene discovery steps.

Given that the very specific rare phenotype LCA is caused by at least 20 genes (estimated from the fact that the six currently known genes cause about one third of the disease), how many genes would we expect to be involved in AMD, glaucoma, or PDG? It seems reasonable to me that the number could be in the hundreds, especially if one includes low penetrance "predisposition" or "modifier" genes. If this is true, then an average disease-causing gene might be responsible for less than 1% of all cases of one of these broad phenotypes, while some will undoubtedly be significantly more and less common than this.

This problem is very analogous to the problem that one faces when sequencing a cDNA library for the purpose of cataloguing all the genes that are expressed in a single tissue. In this situation, highly expressed genes are present thousands of times in a library, while genes that are expressed at low levels might be present once or not at all. Random sequencing of this library would find the most common genes very readily but would require a huge expenditure of energy to find the less common ones. To avoid this problem, one "normalizes" the library with a series of hybridizations and column separations so that in the end, all the different cDNA clones have nearly equal abundance (and an equal likelihood to be sequenced)[23]. The way that this normalization idea applies to the PDG gene discovery experiment is that when we are screening any putative gene for disease-causing variations in a group of patient samples, we would like to have an equal chance of detecting many different genotype/phenotype correlations, rather than an unnecessarily high chance of detecting only one or two.

To accomplish this, one must have very detailed clinical information about the group of patients who are going to be screened so that they can be "phenotypically normalized" before screening. That is, rather than screen a set of patients according to the frequency with which their phenotype occurs in the general population, one intentionally selects patients to widely represent all possible combinations of phenotypes. Having done this, one will have the same likelihood of detecting a rare genotype/phenotype correlation after screening only 400 samples as one would have had screening 4,000 nonnormalized ones. Of course, with this strategy, the number of "positives" one might detect with any given gene will likely be too low to prove that gene's involvement in disease. To confirm an observation made with a set of phenotypically normalized samples, one simply selects a larger number of patients with the same phenotypic characteristics as the "positives" from the first step and screens them for variations in the same gene. This focused "second step" can easily detect a statistically significant genotype/phenotype correlation. For a phenotype that is tenfold less common than average in a broad phenotypic group like PDG, this two-step normalization process will increase the likelihood of detecting a disease gene tenfold, while simultaneously decreasing the amount of work necessary to prove the genotype/phenotype correlation tenfold. This method is not just a theoretical one: it was successfully used to discover that the very rare disease known as the "enhanced S cone syndrome" is caused by mutations in the NR2E3 gene.[24]

Returning to the large experiment for the "third constituency," the phenotypic normalization has in effect reduced the patient axis by about tenfold from about 4,000 to 400. If we believe that there are over 100 genes responsible for each of these broad phenotypic groups (AMD, glaucoma, and PDG) we will need to plan to screen at least 100 to make meaningful headway in these disorders. As a result, the remaining axis, the number of amplimers per gene, will need to be severely constrained, perhaps to as few as two. Is there any way that this can be done that would allow us to find the majority of all there is to find with only 10% of the work? The answer lies in the evolutionary conservation of the gene.

One of the most valuable products of the Human Genome Project and the genome projects of other model organisms is a wealth of information about the way genes have evolved. As discussed more fully below, the sequences of large numbers of genes can be compared bioinformatically and regions of functional importance discerned. An algorithm can be devised that will synthesize a variety of annotation data from the literature to

assign what is in essence a functional importance score to every residue of a protein. Contiguous regions of high scores are then hypothesized to represent regions that are very important to the function of the protein. Extending this idea, one might expect that sequence variations that occurred in such regions would be more likely to result in an observable phenotype than variations that occurred in regions with low functional importance scores. The method we are currently using for this calculation is known as the prioritization of annotated regions, or PAR, algorithm.

To test the usefulness of this algorithm for gene discovery experiments, we retrospectively analyzed 15 disease genes that we have extensively genotyped over the past 10 years and used the PAR algorithm to select a single amplimer for a hypothetical screening experiment. We then asked how many of the 15 genes we would have detected as "possibly disease-causing" and worthy of additional screening. We found that the 15 amplimers that the PAR algorithm selected (which represented only 6% of the total coding sequence of these 15 genes) were capable of identifying 13 of the 15 genes (87%) as worthy of further screening (Braun, Shankar, Sheffield, Casavant and Stone, unpublished data, 2002). Thus, applying the PAR algorithm to the amplimer axis, we might expect to reduce the number of amplimers that need to be screened by more than 90% while losing less than 15% of the findings that would have been detected in a full screen. The combined use of phenotypic normalization for patient reduction and the PAR algorithm for the reduction of amplimers would be expected to reduce the size of the sample space by over 160-fold ($625,000 instead of $100,000,000 per disease) while resulting in about 85% of the gene discoveries that would have occurred if the entire sample space had been screened.

ESTIMATING PATHOGENIC PROBABILITY

The first half of this thesis discussed approaches for dealing with one of the major obstacles to the practical utilization of genomic information for the diagnosis and treatment of human beings: the sheer size of the genome. The second half will concentrate on the other major obstacle: the large amount of non-disease-causing normal variation that exists in the genome and that behaves as noise in most genetic tests. It may seem that looking for a single disease-causing variation in the middle of 3 billion nucleotides (the size of the haploid human genome) is the ultimate "needle in a haystack." However, since any two humans differ from one another at at least 1 million nucleotide positions, the more apt analogy is trying to find a silver needle in a haystack full of a million steel needles. In this portion of the thesis, I will discuss ways for estimating the "silver content" of any "needles" one may find in a human genome.

There are two ways that a sequence variation in the genome can be related to an alteration in the structure or function of the organism that harbors it. The most obvious way is that it can *cause* the altered phenotype by affecting the function of one or more genes in a significant way. The other is that it can be tightly *linked* (ie, on the same chromosome and so close that it is unlikely to be separated by a recombination event) to some other sequence variation that is causally related to the phenotype. For some clinically relevant purposes (such as carrier testing in a family affected with an X-linked disease), it does not matter whether a variant causes the change in the phenotype or is very tightly linked to one that does. In other cases (such as trying to deduce the function of a specific domain of a protein by characterizing the effect of a variation within that domain), it does matter quite a bit. Some kinds of experimental data will support the idea that a variant does alter the function of a gene, while other kinds of data speak only to its physical association with a gene whose function is altered.

Of course, most sequence variations will have no relationship to a patient's phenotype at all. For clinicians trying to use genomic data to help care for their patients, it is helpful to have a system for estimating (and communicating) in a standardized fashion the likelihood that a sequence variation is related to a patient's disease, especially if both functional and association information can be combined in a readily understandable way.

The system we use in our laboratory combines all readily available functional and association information into a score known as the estimate of pathogenic probability (EPP). This system is applicable only to variations in genes that have already been statistically proven to be associated with a given phenotype. But it is useful for helping us (and our collaborators) decide whether a variation is likely to be responsible for a disease that an individual already manifests. It is also useful for ranking members of a group of individuals who harbor sequence variations in a given gene according to the likelihood that their disease is caused by that gene. This may in turn be useful for selecting individuals who would be most likely to benefit from gene-replacement therapy or for selecting which individuals' clinical data to sum when trying to determine the "natural history" of a given disease. The EPP system provides an objective set of rules for communicating all that is known about the pathogenic probability of a given variant. It does not require hours of deliberation among highly trained people (which could introduce all sorts of unpredictable personal bias), and it can be easily revised as new data become available. Both of these latter features are highly desirable when one considers the volume of genotype/phenotype data that is accumulating

in even a single field like ophthalmology (as evidenced, for example by the tables appended to this thesis).

It would be ideal if a system for estimating pathogenic potential could be totally mathematical and have every term in the calculation rigorously supported by well-established statistical theory and large data sets. I must admit that the EPP system that we currently use does not meet this standard and is more of a first approximation that I hope will serve as a precursor to a more sophisticated system in the future. However, as I hope that the reader will see in the sections that follow, we have given quite a bit of thought to objective methods for capturing as much information as possible from the structure of the gene itself and from the way that the gene's alleles are distributed among patient and control groups. However, some of the decisions regarding the weights that various factors are given in the final EPP score were not derived mathematically but were determined empirically based upon our experience in analyzing real families with mutations in the genes that are tabulated in the appendix.

There are currently two sets of rules and two sets of interpretations for EPP values: one for autosomal dominant conditions and one for autosomal recessive ones. The methods are applicable to X-linked diseases as well, but we have not yet devised the specific empirical strategy for weighting the various factors for this inheritance pattern, nor are our data for any X-linked conditions included in the Appendix. In all cases, the EPP has four possible values: 0, 1, 2, and 3. An EPP of 0 means that a variation has very little probability of causing or being meaningfully associated with a disease, while an EPP of 3 means that it is extremely likely that a variation is responsible for the disease. Values of 1 and 2 indicate intermediate likelihoods that a variant is responsible for a patient's disease and have slightly different interpretations, depending on whether the disorder in question is autosomal dominant or recessive. The dominant case is the simplest: the higher numbers simply reflect higher pathogenic potential. With recessive disease, one has to consider the possibility that all alleles do not contribute equally to a recessive phenotype. That is, there may be "low penetrance" alleles that are too common in the population for them to be involved in classic recessive inheritance. This subject is discussed in detail in our paper on allelic variation in the ABCA4 gene,[4] and to date, it is the ABCA4 gene that benefits the most from this additional nuance in the EPP interpretation. For recessive diseases like Stargardt disease, an EPP of 1 indicates a "possible low penetrance allele," while an EPP of 2 indicates a "possible highly penetrant allele" and an EPP of 3 indicates a "probable highly penetrant allele."

The EPP is calculated using all readily available information about the function of the variant allele and its previous association with disease. Before considering the details of the calculation of EPP, it may be helpful to consider the types of functional and association information that might be used for such a calculation and the practical limitations of each. The most obvious way that one could assess the functional effect of a given sequence alteration would be to sample (eg, biopsy) tissue expressing the variant protein and measure the function of the protein directly. When available, this type of information is the most reliable and would obviate the need for an EPP-type system. This approach has been used widely in medicine, especially in the pregenomic era. The demonstration of the functional defect in beta globin in patients with sickle cell disease would be an example of this approach. Unfortunately, this is rarely possible in ophthalmology for two reasons. First, most affected tissues of the eye are not amenable to biopsy from living individuals, and second, the function of most newly discovered genes is so poorly understood that a meaningful assay of their function does not exist even if the tissues were available.

A second, related approach for investigating the function of an altered allele is to create an animal or in vitro model of a disease by artificially expressing (or inhibiting) the gene experimentally. In general, this is a powerful approach but is not without limitations. The main limitation is that the more closely the experiment matches the human situation, the more expensive it is and hence the less practical for assessing hundreds of sequence variations. The less the experiment matches the human situation, the more one has to be concerned that other factors in the experiment (the presence or absence of some other element in the pathogenic process) are more likely to be responsible for the different behaviors of different variants than the variants themselves. This method is also subject to the limitation that it is difficult to measure the function of a protein variant when the function of the normal protein is largely or completely unknown.

A third approach is to use extensive prior knowledge of a protein's structure and function to predict the functional effect of a mutation on the protein. For example, the structure and function of a few proteins that are important to vision (eg, rhodopsin) are exceedingly well worked out.[25-28] High-resolution x-ray crystallographic data[29,30] coupled with functional data from many experiments in several model organisms[31-33] allow one to infer a pathogenic effect of certain mutations. For example, any alteration of the residue (lysine 296) at which 11-cis retinal covalently attaches to the protein could be reasonably predicted to alter its function. This type of information can be used to contribute to the EPP value in an unbiased way by selecting a set of critical residues whose alteration would seem likely to affect the protein's function without

first looking at a set of sequence variation data from humans. When so collected, the predicted functional information provides an independent piece of information about the possible effect of a sequence variation on the protein. The main limitation to this third approach at the present time is that this type of structural and functional information is available for only a small subset of potentially disease-causing genes, and usually only a small subset of residues within these genes. It is also prone to bias and to circular arguments (for example, if one predicts a certain variant to cause disease *after* observing it in an affected patient).

The final approach, and the one that is relied on most heavily by the EPP system, is to use evolutionary evidence gathered from thousands of proteins to assess the functional affect of a given change. This method will be discussed in detail below, but for the present purpose it is sufficient to say that every possible amino acid variation is assigned a value (B) from −4 to +3 in a table known as the blosum 62 substitution matrix.[34] B values below zero indicate amino acid changes that are more likely to have a functional effect than values of zero and above.

In addition to these four kinds of functional data, the EPP system takes three kinds of association data into account. The first is simply the difference in allele frequencies between patients and controls. That is, any variation that is proposed to cause a disease should be more common in patients than controls, but as will be shown below, such a skew by itself does not always reliably infer pathogenic potential. How much rarer does one expect a highly penetrant disease-causing variation to be in unaffected control individuals than in affected patients? The answer depends upon the prevalence and the presumed mode of inheritance of the disease in question. If a heterozygous sequence variation causes detectable disease at an early age in 95% of people who harbor it (ie, is highly penetrant), one would expect the variation to be 19-fold less common in the general population than in the affected population. So, for a rare disease like retinitis pigmentosa (which occurs in about one in 4,000 people), one would expect to randomly encounter an unaffected person with a true disease-causing mutation only once in 76,000 control samples—that is to say, almost never. In contrast, for an autosomal recessive condition in which two different disease alleles must be inherited in order for a person to manifest the disease, true disease-causing alleles are surprisingly common in the general population. For a 1-in-10,000 condition like Stargardt disease, one would expect true disease alleles to be present in about 1 in 50 people. If there are multiple different disease-causing mutations in a single disease gene, then the sum of these will be present in 1 in 50 people. Most people affected with X-linked disease are males with only one X chromosome, and as a result, the relationship between the disease prevalence and the allele frequency in the unaffected male population is very similar to the situation for autosomal dominant disease. That is, highly penetrant alleles that cause very rare X-linked diseases will be extremely rare (for all practical purposes, zero) in the normal male population. The difference in the way that disease allele frequencies are related to disease prevalence is the factor that gives rise to the need for a different EPP calculation for dominant and recessive disease. For a rare dominant disease, any presence of a putative allele in the control group (barring diagnostic error or a sample swap) leads to an estimated pathogenic potential of zero. In contrast, a true highly penetrant allele for a rare recessive disease could easily be observed in a control group of 200 individuals and a low penetrance allele might be present in as many as a few percent of the general population (see Webster and associates[4] for additional discussion).

A second kind of association data that can be used in the calculation of EPP is formal lod score and haplotype analyses of large families. In this type of analysis, data from many genetic markers are considered, which allows one to detect things like ancestral relationships among families[35-37] and linkage disequilibrium among different variants in the same gene. These kinds of data should definitely be used on the "association side" of the EPP calculation when they exist. However, as with some of the types of functional data, this type of formal association information is available for only a small fraction of all sequence variants that are observed and cannot be relied upon for determining most EPP values. In practice, haplotype data are more likely to call a variant into question (eg, a variant that is in clear disequilibrium with a more believable one) than it is to strengthen the argument for its pathogenic probability.

For autosomal dominant disease, the approach that we use most commonly for tabulating association information is a simplification of the lod score method that simply counts the number of times a sequence variant has been observed to properly segregate with disease. This "M number" method will be described more fully below. However, for the purpose of understanding the EPP calculation, it is sufficient to know that an M number of 7 or higher is indicative of a less than 1% chance that the sequence variation and the disease phenotype are cosegregating by chance.

For rare recessive disease, when data are sufficient to suggest a statistically significant likelihood that the allele is more than 100-fold rarer in the control group than the disease group, the variant gets a supportive "point" toward being considered a highly penetrant allele. In contrast, when the data indicate that an allele is too common (as

predicted by the Hardy-Weinberg equation—see glossary in Appendix) to be a highly penetrant allele, it loses "points" and is placed into the "possible low penetrance allele" (EPP = 1) category.

In summary, the EPP for autosomal dominant disease is calculated in the following way:

0. Nucleotide sequence variants that do not alter an amino acid (ie, synonymous codon changes) and/or are present in approximately equal numbers in patients and controls. This category also contains rare variants that have been experimentally shown to have no effect on the function of the protein, variants whose disease association has been convincingly proven to be secondary to linkage disequilibrium with a more convincing mutation, and variants that have failed to segregate in affected individuals with rare diseases (eg, Leu45Phe—see below).
1. Variants that alter an amino acid residue and are more common in patients than controls (small presence in controls is tolerable only for common disorders like AMD or glaucoma—that is, it is related to prevalence) but have no additional functional or association data to contribute to an argument that they are pathogenic
2. Variants that alter an amino acid residue, are more common in patients than controls, and that exhibit *either* functional (B<0) *or* association (M>7) evidence for above average pathogenic potential
3. Variants that alter an amino acid residue, are more common in patients than controls, and that exhibit *both* functional (B<0) *and* association (M>7) evidence for above-average pathogenic potential

For categories 2 and 3, specific functional data can be used instead of the B number when available for awarding the functional "point." Also, frameshift mutations, nonsense mutations (stops), multi-residue insertions or deletions, and mutations involving canonical splice sites are all awarded the functional "point" (blosum 62 calculations are only relevant to single amino acid changes).

The EPP for autosomal recessive disease is calculated somewhat differently so that the numbers 0 to 3 will have a similar meaning to clinicians regardless of the inheritance pattern.

0. Nucleotide sequence variants that do not alter an amino acid (ie, synonymous codon changes) and/or are present in approximately equal numbers in patients and controls. This category also contains rare variants that have been experimentally shown to have no effect on the function of the protein and variants whose disease association has been convincingly proven to be secondary to linkage disequilibrium with a more convincing mutation.
1. Variants that alter an amino acid residue and are more common in patients than controls but that are too common according to the Hardy-Weinberg equation to represent highly penetrant alleles
2. Variants that alter an amino acid residue but are so rare that there is insufficient data to judge that they are either too common to be highly penetrant alleles or 100 times more common in patients than controls—*and* which exhibit no functional evidence for above-average pathogenic potential (ie, B>1)
3. Variants that alter an amino acid residue, are compatible with high penetrance from a Hardy-Weinberg perspective—*and* exhibit one or more of the following characteristics: more than 100-fold more common in patients than controls, exhibit functional (B<0) evidence for above-average pathogenic potential, or results in a frameshift, missplicing, or premature termination of translation

When two affected siblings fail to share genotypes at a locus, it is concluded that that locus cannot be involved in their disease in a recessive mendelian way, and any observations of putative disease alleles are considered to have been made in control individuals. Such an observation could result in the demotion of a variant that was previously considered to be an EPP = 3 because of a 100-fold or greater concentration in patients versus controls.

Leu45Phe

The following example illustrates the value of segregation information in evaluating the pathogenic potential of a new variant in a previously identified disease gene. In the early 1990s, the rhodopsin and *RDS* genes had both been convincingly shown to cause autosomal dominant retinitis pigmentosa,[38] and our laboratory was actively involved in screening these genes in patients with retinitis pigmentosa in search of novel mutations. At the time of this example, we had obtained samples from approximately 450 retinitis pigmentosa patients (diagnosed by clinicians throughout the United States) as well as 200 unaffected individuals living in Iowa (to serve as controls). We screened these patients and controls for sequence variations in the *RDS* gene using SSCP analysis (see methods in the Appendix). Whenever an individual exhibited an aberrant migration pattern on the SSCP gel, that sample was subjected to automated DNA sequencing.

Among the many sequence variations that we observed during this experiment, there was one that appeared especially promising: a mutation that changed the leucine at codon 45 into a phenylalanine. This change was observed in five unrelated probands affected with

retinitis pigmentosa and none of the controls. Since other groups[39,40] had already convincingly proven that the variations in the *RDS* gene were capable of causing retinitis pigmentosa, we felt that the presence of this change in 5 of 450 retinitis pigmentosa patients and its absence from 200 controls was convincing evidence that this sequence variation also was disease-causing. We began preparing these data for publication but also attempted to study as many relatives of the probands as we could to strengthen the manuscript. Most of the five families were small and had few, if any, affected relatives available for study.

However, one family, contributed by Dr Samuel Jacobson, consisted of a total of 14 living individuals, six of whom were affected (Figure 5). These relatives lived in a number of different cities and so their samples had to be obtained by mail. The first family member we received was an affected brother, and as we expected, he also harbored the Leu45Phe sequence variation. Next, we received a sample from the proband's youngest brother, who by history was felt to be unaffected. We were surprised to find that he also harbored the Leu45Phe sequence variation. However, we reasoned that because of his age and the known variable expressivity of the *RDS* gene, he simply did not yet manifest outward signs of the disease. The next sample we received was from a definitely affected brother of the proband, and this individual was found to lack the Leu45Phe sequence variation. There was no doubt of the patient's affected status (Figure 6). A second blood sample was obtained to rule out a sample error, and this was also found to lack the Leu45Phe change. Given the fact that the retinitis pigmentosa phenotype occurs in less than 1 in 4,000 people in the general population, we were forced to accept the fact that the Leu45Phe variation in the *RDS* gene could not possibly be causing the retinitis pigmentosa in this family.

If Leu45Phe is a non-disease-causing polymorphism, how is it possible that more than 1% of a large collection of retinitis pigmentosa samples would harbor this sequence change and that it would be absent from a large control sample set? We contacted all of the clinicians who had contributed the five probands with the Leu45Phe changes and questioned them extensively about the details of the affected families. As previously noted, most of the patients had few, if any, affected relatives, which would be unusual for an autosomal dominant disease. Even in the large family, only siblings were affected. During this questioning, a surprising fact came to light: all five of the affected probands were African American. Since less than 5% of the control population from Iowa were African American, we then suspected that the Leu45Phe change was a non-disease-causing polymorphism that was restricted to the African American population. To test this hypothesis we screened approximately 200 African American individuals from New York City whose samples had been previously contributed as part of another project. Surprisingly, none of these individuals harbored the Leu45Phe change either.

An additional round of phone calls revealed that three of the five families had at least one recent ancestor who had been born on a Caribbean island. Fortuitously, my colleague Val Sheffield had previously collected a series of samples from the Cayman Islands,[41] and a screen of these samples revealed that one of 39 individuals from the Cayman Islands did harbor the Leu45Phe change. A second sample set consisting of 21 individuals from Barbados was contributed by Dr Fielding Hejtmancik, and a screen of these samples revealed that one of 21 individuals harbored the Leu45Phe change. We now suspect that this change is common in some, as yet unidentified, population in West Africa, and that it was carried into the Caribbean hundreds of years ago.

Although in retrospect it seems almost ludicrous that we would have tested a series of samples collected from large tertiary care centers in urban areas such as Miami, Philadelphia, Chicago, Portland, and New York, and then used a set of controls that were collected from the predominantly Caucasian population of Eastern Iowa, at the time we simply did not recognize the relatively high likelihood of observing an amino-acid-changing (but non-disease-causing) sequence variation in a gene that had already been convincingly shown to cause a rare retinal disease. It might also seem that the trivial way to avoid this problem in the future would be to query patients about their ethnicity when samples are being obtained. However, in this very example, such a query would probably not have helped us because the ethnicity information we would most likely have collected would have been "African American" and not "African American of Caribbean ancestry." How would one practically distinguish individuals who had one grandparent who was born in Sweden from individuals who had one grandparent born in Ireland? There could easily be (and probably are) non-disease-causing amino-acid-changing sequence variants that are common in either Ireland or Sweden but rare or absent in the other country. The safest procedure is to obtain control samples from the same clinic population that one uses to obtain the patients with the disease under scrutiny. Perhaps the most important thing that can help an investigator avoid this type of "Leu45Phe artifact" is the recognition that such ethnic-specific non-disease-causing polymorphisms are actually reasonably common and that one must be concerned that any variation that one observes in a candidate gene-screening experiment may indeed be one of these.

Apart from their ethnicity, which we have already

discussed, what other feature of these Leu45Phe families could have raised a red flag that this change was not truly disease-causing? The answer is the pedigree structure. Since the Leu45Phe change that we observed was heterozygous, and since the vast majority of *RDS* changes that had been reported up to that time behaved in an autosomal dominant fashion, we should have expected the Leu45Phe change to behave in an autosomal dominant fashion as well. Of course, it would be understandable for any given family who harbored a heterozygous sequence change to fail to manifest an obvious autosomal dominant pattern. Most autosomal dominant diseases exhibit some degree of incomplete penetrance, and even without invoking incomplete penetrance, one can imagine an affected parent who died relatively early or who simply failed to complain of visual symptoms during his or her lifetime. However, it would be extremely unusual for a truly autosomal dominant disease to appear in only a single generation in five different families.

The Leu45Phe experience taught me that arguments for pathogenicity that are based on differences in allele frequencies between patients and controls are prone to errors caused by unsuspected differences in ethnicity between these groups. In contrast, arguments based on cosegregation of a disease allele with the phenotype are more robust, and indeed it was the nonsegregation in the larger family that saved us from publishing a non-disease-causing polymorphism as a disease-causing mutation.

M NUMBER

After the "near miss" of Leu45Phe, I became interested in whether we could devise some type of simple rule for analyzing our candidate gene screening data that would help alert us to the possibility of a non-disease-causing polymorphism. I decided to start counting and recording all affected individuals (whether seen in our lab or published in the literature) who correctly segregated the putative disease-causing variation. We call this count of correctly segregating meioses (minus the proband who would have the change "by definition" in a candidate gene screening experiment) the "M number." With each correctly segregating meiosis that we can observe or find in the literature, the likelihood that the sequence variant and the disease are associated by chance becomes less and less. By the time the M number reaches 7, the likelihood of that happening by chance has become less than 1 in 100 (Figure 7). Of course, this is quite analogous to the lod score method of evaluating the statistical significance of the segregation of genetic markers near a disease locus in a family. Like lod scores, M numbers are additive among families. M numbers are easy to calculate and do not require any information other than the affection status of the patient and their molecular status. These data are often available in manuscripts that report a disease-causing mutation, and this allows data from the literature to be objectively combined for evaluation of sequence variations.

The M number can be used to evaluate sequence variations in at least two important ways. First, for diseases that are suspected to be autosomal dominant, if one counts all of the families that one has available for study (F) and divides this number by the cumulative M number across these families, a ratio of less than or equal to 1 is suggestive that this is either a non-disease-causing polymorphism (as was the case for the Leu45Phe change) or that the families have been incompletely studied. In contrast, a well-studied, true autosomal dominant mutation such as the Pro23His variant in the rhodopsin gene will have an M number a few to many-fold greater than the number of families, because once three or four families have been identified, at least one of them will usually have a moderate to large number of affected individuals correctly segregating the variation.

The other (related) use of the M number is to show in a numerical way how much or how little one actually knows about a given sequence variation. That is, when a sequence variation has been observed in only a single affected individual in the entire world's literature, its cumulative M number will be zero, while the M number for a well-characterized variation such as Pro23His will be well over 35. When data from many different mutations are tabulated (eg, the tables that accompany this thesis) and an M number is calculated for each one, it becomes evident that certain sequence variations have a lot of segregation data to support their autosomal dominant nature, while other variations have little if any. When affected individuals are identified who do not harbor the sequence variant that is present in the proband, a value of $0.1/P$ (where P is the prevalence of the disease) is subtracted from the M number. For retinitis pigmentosa, this would be –400 for every occurrence, which means that even a single observation of a nonsegregating individual would mean that a variant was not likely to be disease-causing. However, for more common diseases like AMD and glaucoma, these segregation failures do occur quite often and should have a much milder negative effect on M.

In Appendices B and C we have compiled all data available to us from the world's literature (as well as the unpublished data from our laboratory) for families that harbor heterozygous sequence changes in the rhodopsin and *RDS* genes, and we have calculated the M numbers for all of these sequence variations. In Figures 8 and 9, these sequence variations have been ranked according to M number and then displayed as a decreasing histogram. What is interesting about these figures is that although well over 100 different amino-acid-changing sequence

variations have been identified in the rhodopsin gene (and over 75 such changes in the *RDS* gene), only about 18% of these changes have sufficient segregation information to support their autosomal dominant nature in a statistically significant way. Equally interesting is that more than half of these variations have no segregation information whatsoever (M = 0), raising the very real possibility that some of these are actually non-disease-causing polymorphisms like Leu45Phe. If family members of these "M = 0" probands could be sought and studied, some segregation failures would be identified and would make the EPP fall to zero, while for other families, correctly segregating meioses would be identified which would have the possibility to increase the EPP for that variant. This underscores the fact that our understanding of the pathogenic potential is not fixed, but can improve with time if we are diligent in gathering as much clinical information as possible from families that we identify.

I first graphed a ranked list of the world's M numbers for rhodopsin and *RDS* in 1996. At that time, the number of variants for which M = 0 was only 33% for each gene. Nearly all of the more recently published variations have no family information associated with them, and this limits the clinical utility of these variants. However, this limitation can be overcome if practical genetic testing for these genes can be deployed and if the results of this testing (with segregation information) can be reliably contributed to a curated database.

No simple rule like the M number calculation will allow a correct estimate of pathogenic potential in every case. One could, for example, observe a high M number for a non-disease-causing change if there was a true disease-causing mutation elsewhere in the same gene (perhaps undetectable with current methods). Similarly, a true disease-causing mutation could have a very low M number simply because it is extremely rare and had only been observed in a very small family. Still, despite these caveats, we have found the M number to be a simple and reliable way to summarize all the information that is known about the segregation of a heterozygous sequence variation. Moreover, when depicted cumulatively for all known sequence variations, it provides a clear reminder that the amount of information that we have for different putative disease-causing variants is quite variable, a fact that is important to remember as we counsel our patients about the clinical meaning of one of these variations.

M numbers can also be calculated for autosomal recessive and X-linked pedigrees, but the rules for these cases are beyond the scope of this thesis. Can M numbers also be calculated for variations in mitochondrial DNA that cause rare diseases such as Leber hereditary optic neuropathy (LHON)? The answer is no, but perhaps this deserves a bit of explanation. Since the molecular nature of mitochondrial disease has been known for only about 15 years, most clinicians are much less familiar with the practical consequences of the inheritance of mitochondrial disorders than they are for diseases caused by genes in the nuclear genome. The entire mitochondrial genome consists of only about 16,000 base pairs of DNA and encodes 2 rRNA subunits, 22 tRNA molecules, and 13 polypeptides.[42] The major difference between the mitochondrial genome and the nuclear genome is that except for relatively recent mutations which may be present in a "heteroplasmic" state (meaning that both normal and abnormal molecules are present in the same cell) in most individuals, the DNA sequence of their mitochondrial DNA is uniform "homoplasmic" and the same in every cell in the body (ie, nonmosaic).[43] The second major difference is that with very rare exception, all of an individual's mitochondrial DNA is derived from the mitochondria that were present in the oocyte at the time of conception and hence is derived entirely from the individual's mother.[44-47] The third major difference in mitochondrial DNA is that it does not participate in recombination. This circular mitochondrial DNA molecule is actually an evolutionary remnant of an ancient prokaryotic organism whose fusion with another prokaryote gave rise to the first eukaryotes.[48] As a result, the mitochondrial DNA replicates like a bacterial genome, and the mitochondria themselves reproduce by simple binary fission. There is no "sexual reproduction" of these mitochondria and therefore no exchange of DNA segments as there is in meiosis for the human nuclear genome. As a result, there is not a 50% chance of getting one or the other mitochondrial allele transmitted from parent to child in a given generation. There is for all practical purposes a zero percent chance of inheriting the paternal mtDNA molecule and a 100% chance of inheriting the maternal one.

Why, then, isn't every sibling of a patient with LHON affected? Why are males much more likely to be affected with this disease than females? Answers to these questions are not known at the present time, but it seems likely that both nuclear genes and environmental factors contribute to the pathogenesis of LHON. This "incomplete penetrance" of a maternal mutation coupled with a peculiar predilection for males can make a mitochondrial disease like LHON look very similar to an X-linked disease by pedigree analysis. However, on careful inspection, there is a detectable difference. In mitochondrial disease, there is never an example of transmission from an affected male to any offspring, while for X-linked disease, the limitation is only that males cannot transmit the disease to their sons. This difference was actually recognized decades ago by astute clinicians who termed this maternal inheritance pattern "cytoplasmic inheritance." Later it was realized that the cytoplasmic factor responsi-

ble for the disease was the mitochondrial DNA carried in the cytoplasm of the mother's oocyte.[49]

The importance of all this to the M number discussion is that there can be no M number for mitochondrial mutations because there is no meiosis—no "M." This actually makes it much more challenging to prove that a given sequence variation in mitochondrial DNA is responsible for a disease. For mitochondrial disease, the "association argument" needs to be based upon an extreme difference in frequency of a sequence variant in patients compared with controls (ie, essentially equal to the difference in frequency between the disease state and the normal state). As with any sequence variation, information about the conservation of that particular residue throughout phylogeny can be used, and since the mitochondrial genome is derived from a common ancestral molecule, this argument can be more effective for mitochondrial mutations than for many nuclear mutations. Some investigators have claimed that modest differences in the frequencies of mitochondrial variants are responsible (at least in part) for extremely rare diseases like LHON.[50] In my opinion, it is hard to imagine how a variant that is present in a few percent of the population could have any significantly "causative" role in a disease that occurs in less than 1 in 100,000 people per year. I believe that modest differences in the frequency of mitochondrial sequence variants are more easily explained by a combination of linkage disequilibrium and ethnic variations.

BLOSUM

In the 1970s, when scientists first began grappling in earnest with protein sequence information from many different organisms, it was recognized that there were ancestral relationships between entire proteins as well as between certain domains of proteins. As investigators considered how they might recognize meaningful relationships among proteins, they recognized that variation of certain residues (averaged over a large number of different proteins) was more tolerated by evolution than variation of other residues. They were able to use this observation when searching for meaningful homologies by assigning a greater weight to alignment of residues that were less likely to vary. Dayhoff and colleagues[51] developed this idea extensively and devised the first substitution matrices for predicting the likelihood that certain substitutions would occur after defined amounts of evolutionary time. The Dayhoff matrices were based upon comparisons of entire proteins (global alignments) that were highly homologous to each other (>85%). The parent matrix was called PAM-1 and showed the relative likelihood of specific amino acid changes being tolerated by evolution (PAM = percent accepted mutation) after the protein sequences had drifted enough to result in 1 change per 100 residues. Other matrices for predicting greater degrees of divergence were then created by multiplying the PAM-1 matrix by itself.

In 1992, Henikoff and Henikoff[34] proposed a substitution matrix that, while broadly similar to that of Dayhoff and colleagues, had some important differences. First, more than 2,000 different "blocks" of protein sequence were compared instead of entire proteins. Second, the matrices that were designed to evaluate distantly related proteins were calculated from actual observations of protein blocks exhibiting that degree of divergence, rather than by observing the values for 1% divergence and multiplying those by themselves to simulate greater divergence. Henikoff and Henikoff called their matrices "blosum" for "blocks substitution matrix," and the most widely used version is the blosum 62, which is calculated from blocks with 62% or less homology. In practice, both the PAM series of matrices and the blosum matrices perform very well in recognizing distant protein relationships.

With respect to estimating the pathogenic potential of sequence changes in human disease genes, the relevant idea embodied by these substitution matrices is that certain amino acid residues are physically and chemically more similar to one another than others. As evolution proceeds, random mutations will occur and mutations that result in the substitution of an amino acid by a very similar one will be more likely to be "accepted" by evolution because it is not deleterious to the function of the protein. In contrast, substitution of an amino acid by a very different one will more often result in an unfavorable effect on the protein and hence will not be accepted (ie, will not be present among the proteins that still exist for us to study).

Consider the blosum 62 matrix shown in Figure 10.[34] Every possible amino acid substitution is represented by a cell in the table. Positive numbers indicate a greater degree of evolutionary tolerance (and by extension functional similarity), while negative numbers suggest that when averaged over thousands of proteins, a certain amino acid change is strongly disfavored (and by extension would be expected to have a higher-than-average probability of causing dysfunction in a human disease gene). In Figure 10, the matrix is arranged to place functionally similar residues next to one another. Note that the colored boxes near the diagonal contain values for all the substitutions within these functional groups, and on the whole, these values are above 0. As one moves away from the diagonal, the numbers on the whole become more negative, suggesting greater functional differences between the amino acid pairs represented by those cells.

In the past, our laboratory employed a crude version of the idea of ranking the degree of functional impact of a specific amino acid change when we gave greater weight to amino acid substitutions that altered the charge or

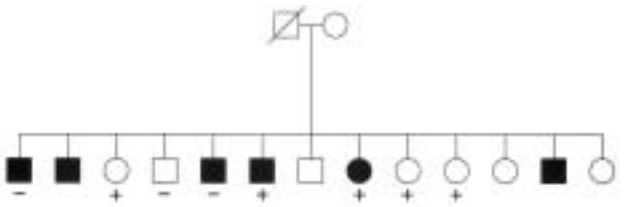

FIGURE 5

Pedigree of the Leu45Phe family. Individuals affected with retinitis pigmentosa are shown as closed symbols. A plus (+) indicates that the patient was screened for RDS variations and a heterozygous Leu45Phe change was found. A minus (-) indicates that the patient was screened and no sequence variations were observed. Note the lack of agreement between the Leu45Phe genotype and the retinitis pigmentosa phenotype.

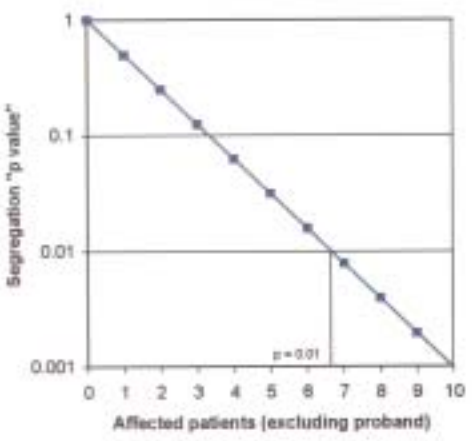

FIGURE 7

Relationship between M number and the probability of chance cosegregation of a variant and a disease. This figure illustrates that for an autosomal dominant disease, the likelihood that affected relatives will share a certain heterozygous sequence change by chance decreases with each affected family member that is studied.

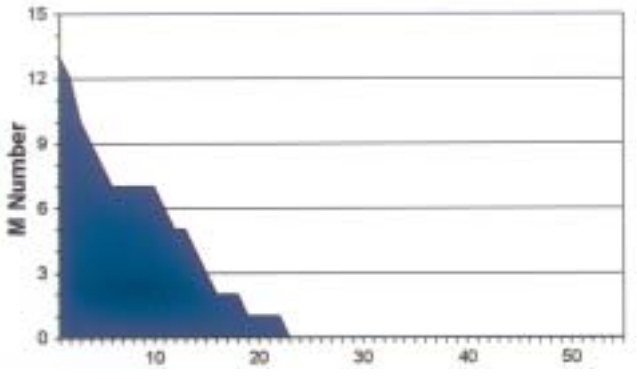

FIGURE 9

M numbers for the RDS gene. Cumulative M number values were calculated utilizing all segregation data available in our laboratory as well as segregation data available in the published literature (Appendix C). As in Figure 8, about 18% have an M>7. 60% have no segregation information at all.

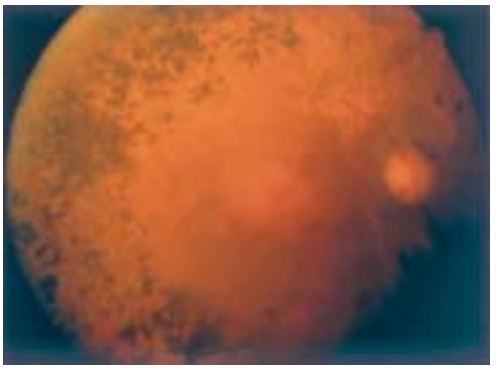

FIGURE 6

Fundus photograph of a patient from a family that harbors a Leu45Phe variant in the RDS gene. This patient did not exhibit the Leu45Phe change, and this observation was the key to discovering the non-disease-causing nature of this variation. (Photograph courtesy of Dr Sam Jacobson.)

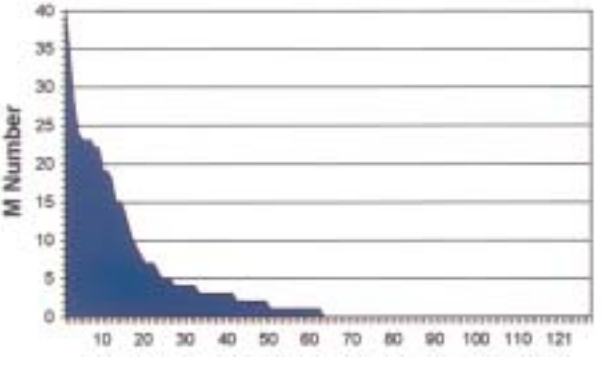

FIGURE 8

M numbers for the rhodopsin gene. Cumulative M number values were calculated utilizing all segregation data available in our laboratory as well as segregation data available in the published literature (Appendix B). The variants were ranked according to decreasing M number. About 18% have an M>7 (which corresponds to P,<.01 for chance association), while more than 50% have no segregation information at all.

FIGURE 10

Blosum 62 matrix. Henikoff and Henikoff's block substitution matrix34 quantifies the likelihood that natural selection will accept a change from one amino acid to another. More positive values signify a greater tolerance, calculated by observing "blocks" of proteins with 62% or less homology. The colored groups indicate amino acids with functional similarities. As one might expect, variation within these groups (along the diagonal of the matrix) is more well tolerated by evolution than variation among them.

polarity of the protein than we did to substitutions that did not alter these parameters.[5] We suspected that the use of the blosum 62 matrix would be superior to the charge-polarity criterion because it is based upon the actual evolutionary tolerance of substitutions rather than a somewhat arbitrary grouping of residues in a biochemistry textbook. The idea of using a blosum substitution matrix to help predict the functional effect of sequence variations is not entirely new. Ng and Henikoff[52] explored this idea for three nonhuman genes for which functional assays were available.

We performed an experiment to see whether patients with retinitis pigmentosa and amino-acid-changing point mutations in either of two well-characterized genes (rhodopsin and *RDS*) would exhibit negative blosum numbers more often than would be expected by chance. To do this, we first calculated blosum numbers for all possible point mutations in the rhodopsin and *RDS* genes. Since different proteins have different amino acid compositions and different codon usages, they will have different probabilities of harboring each specific amino acid substitution. Any codon can be changed into nine other codons by single base substitution, and because of the degeneracy of the genetic code, some of these will not change an amino acid, while others will. For rhodopsin, there are 3,132 (348 × 9) blosum values. Of these, approximately one third are synonymous (ie, they do not change the encoded amino acid). Figure 11 shows the frequency distribution of the remaining (ie, nonsynonymous) values in the rhodopsin gene. Figure 12 shows the same distribution for the *RDS* gene. Of these remaining sequence variations (those that would change an amino acid), approximately half have a blosum value of –1 or less.

Appendix B summarizes all rhodopsin variations that have ever been observed in patients with retinitis pigmentosa, and Appendix C summarizes all of the retinitis pigmentosa–associated variations in the *RDS* gene. When one compares the distributions of blosum numbers for the disease-associated variations from these two tables, one finds an average value that is significantly more negative than one would expect by chance ($P = .001$ for both). In fact, approximately 71% to 78% of the disease-associated changes in both tables exhibited a blosum value of –1 or less (which is the threshold for gaining a "function point" in the EPP system).

We were also interested in comparing the use of the blosum matrix to our old criterion of evaluating charge and polarity. With the latter system, one can only receive a score of –1 or 0. Figures 13 and 14 show the distributions of these scores for all possible point mutations in the rhodopsin and *RDS* genes. When we compared the "charge-polarity" scores for all the mutations in Appendix B and C with those that would be expected by chance, there was a noticeable difference, but the *P* value was at least tenfold larger than that for the blosum data, suggesting a greater discriminative power for the blosum method.

We realize that there will be limitations to the predictive power of the blosum matrix and that there will be true disease-causing variations with positive blosum 62 numbers as well as non-disease-causing variations with negative ones. The reason that one should expect this is that the blosum numbers reflect the overall structural and functional similarity of two residues, while in the local context of a specific protein almost any change could be tolerable or disastrous. However, the fact that the blosum values we have chosen as indicative of disease are highly associated with a large number of disease-associated variations in two different genes suggests that this is a valid method for using evolutionary data to help estimate the pathogenic potential of individual variations.

DISCUSSION

The scientific method by its very nature is an iterative process. One makes some observations, then develops a hypothesis based upon those observations, and conducts experiments to test those hypotheses. Finally, one interprets the results of the experiment and these "interpretations" become the "observations" for the next round in the process. The type of research described in this thesis occurs at the interface between basic science and clinical science, and the iterative steps of the scientific method often alternate between these two worlds. A clinician identifies a constellation of unusual findings and characterizes them well enough that a few additional families with the same disorder can be recognized. Basic scientists, screening a phenotypically normalized group of patients, may find only two or three out of hundreds to harbor changes in a novel gene, but the precise (and identical) phenotypic information provided by the clinicians allows the hypothesis to be focused 100-fold. Additional clinical work results in the identification of 30 families with the same phenotype, and additional molecular screening provides overwhelming statistical evidence that this gene is responsible for this rare phenotype. Finally, armed with a confirmed and detailed correlation between genotype and phenotype, a clinician can predict which gene is likely to be responsible for a patient's disease and order a very focused molecular study with a high likelihood of discovering clinically relevant information in a short period of time and with a modest cost.

Two of the most striking examples of this process in my own personal experience were the recognition of the tendency for N-terminal rhodopsin mutations such as Pro23His to selectively affect the inferior retina, resulting in a very recognizable Goldmann visual field (Figure 15A

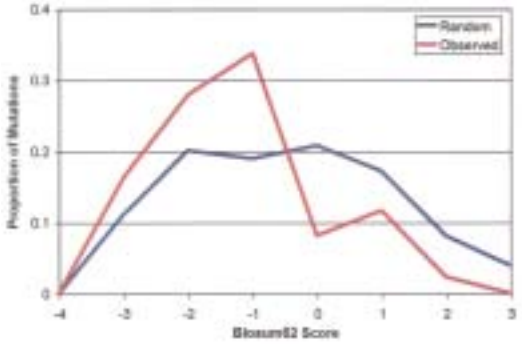

FIGURE 11

Blosum 62 scores for rhodopsin. The "random" curve illustrates the distribution of blosum 62 scores for all possible single-nucleotide substitutions in the sequence for human rhodopsin. The "observed" curve shows the distribution of scores among the single-nucleotide mutations of RHO that are observed in patients with retinitis pigmentosa. The disease-causing mutations' average is significantly more negative than that of the whole population of possible mutations (P = .001).

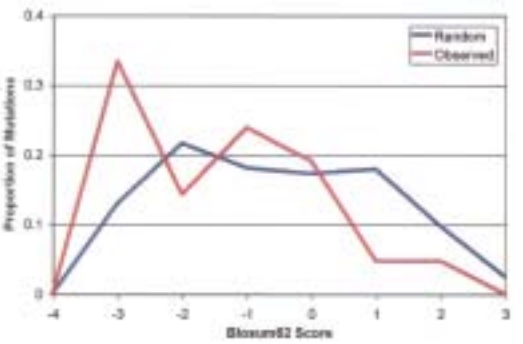

FIGURE 12

Blosum 62 scores for RDS. The "random" curve illustrates the distribution of blosum 62 scores for all possible single-nucleotide substitutions in the sequence for human RDS. The "observed" curve shows the distribution of scores among the single-nucleotide mutations of RDS that are observed in patients with retinitis pigmentosa. The disease-causing mutations' average is significantly more negative than that of the whole population of possible mutations (P = .001).

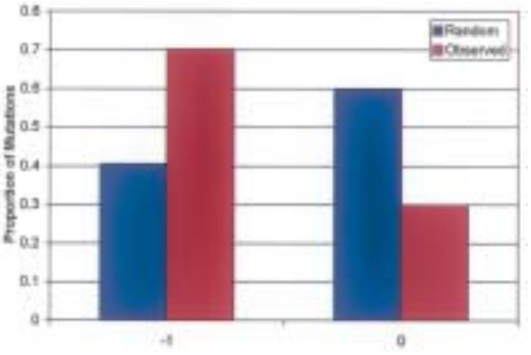

FIGURE 13

Rhodopsin charge-polarity scores. A similar experiment to the one shown in Figure 11, except that here the scores are assigned based on a charge-polarity criterion instead of the blosum 62 matrix. If the nucleotide change results in an amino acid with a different charge or polarity from the wild type, the score is –1. Otherwise, the score is 0. Again, the disease-causing mutations are, on average, more negative, but the P value is one log unit larger than with the blosum 62 matrix.

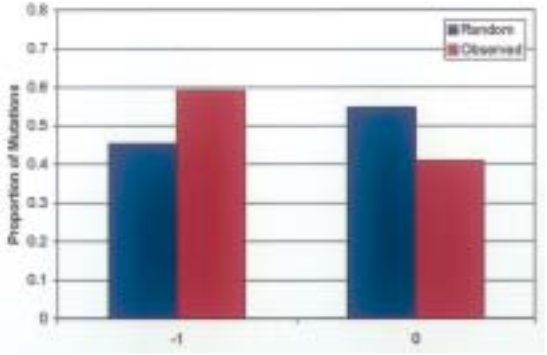

FIGURE 14

RDS charge-polarity scores. A similar experiment to the one shown in Figure 12, except that here the scores are assigned based on a charge-polarity criterion instead of the blosum 62 matrix. If the nucleotide change results in an amino acid with a different charge or polarity from the wild type, the score is –1. Otherwise, the score is 0. Again, the disease-causing mutations are, on average, more negative, but the P value is one log unit larger than with the blosum 62 matrix.

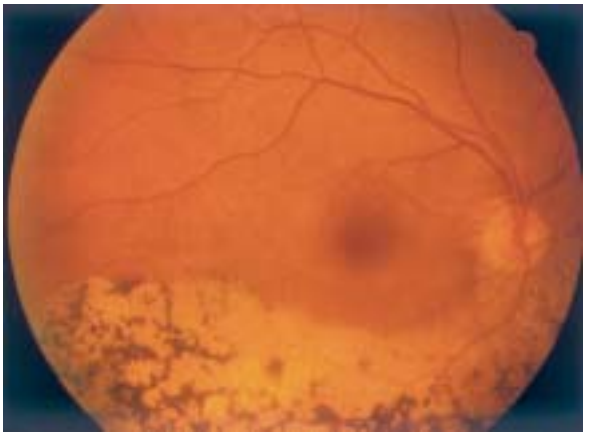

FIGURE 15A

Regional retinitis pigmentosa associated with a mutation in the rhodopsin gene. Fundus photograph from a patient with a rhodopsin mutation (Gly106Trp) demonstrating the peculiar inferior predilection often seen in this specific form of the disease.

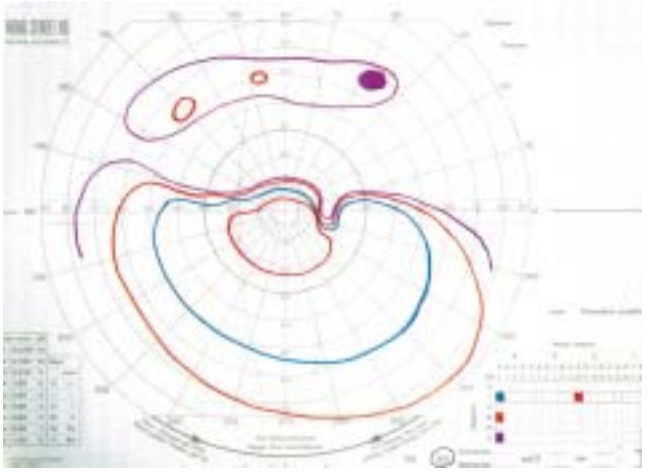

FIGURE 15B

Goldmann visual field from patient in Figure 15A, which mirrors the regional effect of the disease.

and B).[53, 54] Before this genotype-phenotype correlation was established in the early 1990s, I certainly would not have suspected a patient to have autosomal dominant disease solely on the basis of this visual field finding. However, since then, I have predicted the existence of heterozygous changes in the rhodopsin gene several times on clinical grounds alone and then confirmed it with a single molecular test. Similarly, I had never personally made the diagnosis of the enhanced S cone syndrome before Val Sheffield and I found changes in the NR2E3 gene in two of Sam Jacobson's patients.[24] However, after having the opportunity to personally examine molecularly confirmed individuals (Figure 16A) with this characteristic fundus appearance, I was recently able to correctly predict this rare diagnosis on clinical grounds in a 7-year-old girl (Figure 16B).

In the section of the thesis that discussed the importance of goals, I asked the reader to consider how he or she would prioritize the various goals of molecular ophthalmology. My view is that the various goals that I presented are interrelated parts of a long-term process that will eventually lead to a cure for many of the inherited eye diseases that we struggle with today. I like to refer to this process as gene-directed therapy (Table IV) to emphasize the critical requirement for genetic information in many of the steps and also to acknowledge that many of the treatments that result will likely be directed by genomic information but not necessarily involve gene replacement or other "true" gene therapy. As far as prioritization is concerned, I think that the key is to focus on the throughput of the entire process rather than on any specific step. I think that we should be constantly on the lookout for bottlenecks in the process and direct more intellectual energy and other resources toward these bottlenecks—something that is unlikely to occur without consciously maintaining a perspective of the "big picture." After all, bottlenecks exist because there is some real economic, technological, political, or legal barrier, and there will be little impetus to try to remove such a barrier without a perspective of the process as a whole.

For example, when one considers the steps of gene-directed therapy listed in Table IV, one can see that two of the essential steps will involve genotyping of large numbers of individuals. For step 2, we will need to gather sufficient data about new genes that we can know which 10% of the sample space contains the majority of the useful information. We also need to know which sequence variations have clinical and functional relevance and which ones are simply reflections of Caribbean ancestry. In step 4B, we will need to identify moderately large groups of people with specific sequence variations and then carefully characterize their associated clinical phenotypes.

How is this genotyping likely to be accomplished? A subset of it is a legitimate, fundable research enterprise that is closely connected to the gene discovery effort. That is, one might expect a laboratory that is trying to characterize a novel disease gene to screen between 400 and 2,000 people[4,16,36] to generate the kind of data shown in Figures 3 and 4 as part of the discovery and initial characterization of a disease gene. This would be analogous to a period of development of a new type of medical imaging during which physicians try to understand the strengths and weaknesses of the new device. But what will happen after this initial period of gene characterization? Are the data that result from a prospective genotyping of 400 to 2,000 people the highest-resolution information that we can ever hope to have from this gene?

In other branches of medicine, this amount of data would be just the beginning. Returning to the hypothetical new imaging device, if it showed promise in the laboratory (and if it were commercially viable), it would be deployed in the clinical realm, and this would allow practicing clinicians to begin making their own observations

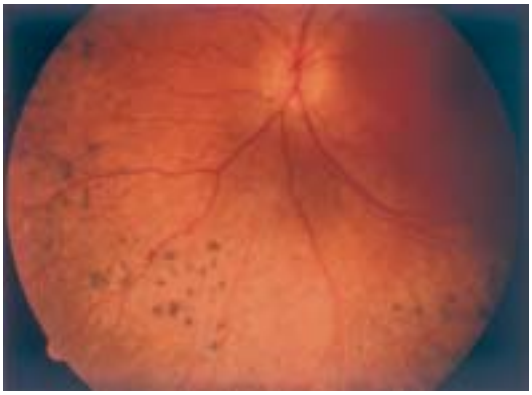

FIGURE 16A
Enhanced S cone syndrome. Fundus of a patient with the enhanced S cone syndrome that was discovered during the initial mutation screen of the NR2E3 gene.

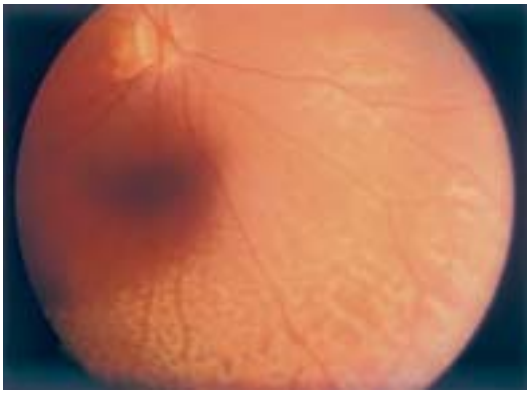

FIGURE 16B
Fundus of a 7-year-old child whose NR2E3 mutation was predicted based upon the similarity of the fundus appearance to that of the patient shown in Figure 16A.

TABLE IV: GENE-DIRECTED THERAPY

1. Gene discovery
2. Characterize the variations that occur in each gene to discover which ones are likely to cause disease and thoroughly explore the phenotypic range and natural history of these variations
3. Develop or identify in vitro or animal models of the disease that mimic the human disease sufficiently that they can be used for evaluation of potential treatments.

4A. Develop treatments based upon the knowledge of the disease mechanism gained in steps 1-3 and test them in the in vitro or animal models

4B. While the animal models are being developed and tested, screen appropriate human populations for mutations so that mechanistically homogeneous groups with defined natural histories will be available for clinical trials.

5. Once a treatment looks promising in a model, conduct a clinical trial in a human population that is as genetically and mechanistically similar to the successfully treated model as possible.
6. If the clinical trial in step 5 is successful, try to cautiously generalize the results to other related disorders.

about the way that images from this new device correlate with images from older devices or with the disease process itself in the context of the daily care of patients. Such an ongoing process of employing the new device to actually care for patients on a large scale is capable of ultimately gathering orders-of-magnitude more new information than the modest experimental data set that was gathered initially for the purpose of making the device usable in the first place.

Where does this analogy break down in the context of rare inherited eye diseases? For example, why isn't our understanding of the more than 40 genes currently known to cause mendelian forms of retinal disease being refined by daily clinical use by ophthalmologists all around the world? Why hasn't genetic testing for these disorders become commonplace and, as a result, yielded an order of magnitude of additional data about sequence variations in the human population as a whole? There is no one answer to this question, but there are some recognizable factors that are worth considering.

In medical imaging, one might have a single test, such as a magnetic resonance image (MRI) of the head, that would be equally applicable to a patient with a rare developmental abnormality, a patient who had fallen off of his bicycle, or a patient with a cerebrovascular accident. This fact has two important implications to the present discussion. First, the general applicability of the method in essence spreads the cost of its development and use over a large number of individuals and increases its chance of commercial viability (the corollary to this is that valuable methods with very limited applicability may be never be "commercially viable").

The second implication of the broad applicability of an MRI that is relevant to this discussion is the way such a method could be used by a clinician to broadly sample the state of a patient's health. This is in essence what the physician with the "run the chromosomes" request was thinking. He was asking for an "MRI of the head" of the genome. However, as I have discussed throughout this thesis, the human genomic sample space dwarfs anything that we have dealt with before in clinical medicine by several orders of magnitude. It is safe to say that we are not even remotely close to being able to simultaneously assess and understand the clinical import of millions of genetic loci that would be the true equivalent of the MRI of the brain in this analogy.

Another factor that has seriously hampered the translation of genomic knowledge into daily clinical decision making is the whole concept of the human genome as "intellectual property." Imagine the cost of an MRI if someone "owned" the normal anatomy of the head of the caudate, someone else the substantia nigra, and a third group the visual radiations. Imagine an intellectual property officer of a large university, facing the financial challenges that large universities always seem to face. Would that person be willing to relinquish the rights to the anatomy of both lateral ventricles (if the US patent office had indeed granted them to the university) if these rights had the potential to generate some revenue for the university from every MRI performed in the country? The answer is that they probably wouldn't relinquish those rights and that it might be impossible to collectively satisfy all of the different intellectual property interests and still keep the MRI commercially viable.

Obviously, the issue of the genome as intellectual property is complex, and there are certainly many situations in which intellectual property protection will allow a company to develop a diagnostic procedure or molecular treatment that would never have been developed without such protection. However, I think that in the case of very rare diseases (those that occur in 1 in 10,000 people or less, for example), the prospect for sustainable commercial genetic testing is slim in most cases. I think that a better strategy for achieving sustainable genetic testing for rare diseases is "nonprofit fee-for-service testing" in academic institutions. In this approach, CLIA-certified academic laboratories that have an interest in specific classes of disease—inherited eye diseases in my case—would offer tests to individual patients on a nonprofit basis. The motivation for doing this would be to provide a benefit to society—a primary mission item for most universities but usually, at best, a secondary one for large corporations. Academic fee-for-service testing would represent a significant improvement over performing such testing on a "research" basis in academic centers. Tests could be performed and reported in a timely fashion because doing so would no longer result in the loss of

resources for "real" research projects. In my opinion, the only local barrier to this strategy is convincing university officials about the value of such a program as a "mission item" instead of a "revenue stream." The more significant obstacle may be gaining sufficient access to existing genomic intellectual property that the tests can be legally offered. I believe that there is a great degree of overlap between this societal need and the need for development of commercially nonviable pharmaceutical treatments for patients with rare "orphan" diseases. Legislation has helped patients in the latter situation, and I am hopeful that similar legislation will help advance the academic fee-for-service strategy as well.

A final barrier to the optimal use of genomic information in clinical ophthalmology is the relative lack of experience that the current generation of physicians has with this diagnostic modality. Unfortunately, medical students and house officers currently have little opportunity to learn about the practical aspects of molecular medicine during their training. One doesn't really learn about fluorescein angiography by reading about the chemistry of fluorescein or the physics of barrier filters. One learns about fluorescein angiography by interpreting fluorescein angiograms for the care of patients during one's clinical training. Similarly, I believe that clinicians will learn how to best employ molecular approaches by actually using these approaches to care for patients. If an academic fee-for-service strategy (or something similar to it) can be made to work, even for rare diseases, it will allow practicing physicians to become more knowledgeable about the benefits and limitations of genomic approaches. Widespread practical knowledge of the medical benefits of genomic information will obviously be essential for the full societal potential of this branch of science to be realized.

I am optimistic that in the coming decade, we will overcome many of the barriers to the development for practical genetic tests; and I am hopeful that as these tests get into the hands of the majority of clinicians, our knowledge of genetic variation among the world's populations will grow well beyond the boundaries that are inherent in focused prospective studies funded by research organizations. I am also optimistic that this improved knowledge will foster the progress of the entire gene-directed therapy process to the degree that we will soon be able to prevent the loss of vision in many of our patients.

ACKNOWLEDGMENTS

The data discussed in this thesis were collected over more than 10 years, and it is no exaggeration to say that hundreds of people were involved in collecting and analyzing them. I greatly appreciate having had the opportunity to collaborate with so many generous, hard-working, and gifted people throughout my career. I am also very grateful to the more than 40,000 people who have contributed blood samples directly or indirectly to my laboratory. I must also name a few people who contributed specifically to this thesis. Since 1990, Val Sheffield has taught me how to map genes, to find mutations, and to "think big." Over 90% of my CV has been coauthored with Val, which is the most succinct metric I can think of to describe the incalculable value he has been to my career. Sam Jacobson, Jerry Fishman, and Dick Weleber have collectively taught me the majority of what I know about inherited retinal diseases and electrophysiology and, in addition to helping me get started on this career path, have been constant active collaborators for the past 12 years. Tom Casavant and Terry Braun introduced me to bioinformatics and high-performance computing, which are essential if one wants to convert "thinking big" into "doing big and fast." Jean Andorf, Luan Streb, Louisa Affatigato, Trish Duffel, Matt Rauen, Rhett Sutphin, Becky Johnston, Linda Koser, Rob Mullins, Michael Grassi, and John Fingert created and checked tables and figures, found references, read the text, and, in sum, made this manuscript possible.

REFERENCES

1. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;291(5507):1304-1351.
2. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860-921.
3. Sunyaev S, Ramensky V, Koch I, et al. Prediction of deleterious human alleles. *Hum Mol Genet* 2001;10(6):591-597.
4. Webster AR, Heon E, Lotery AJ, et al. An analysis of allelic variation in the ABCA4 gene. *Invest Ophthalmol Vis Sci* 2001;42(6):1179-1189.
5. Alward WL, Fingert JH, Coote MA, et al. Clinical features associated with mutations in the chromosome 1 open- angle glaucoma gene (GLC1A) [see comments]. *N Engl J Med* 1998;338(15):1022-1027.
6. Braude P, Pickering S, Flinter F, et al. Preimplantation genetic diagnosis. *Nat Rev Genet* 2002;3(12):941-953.
7. Rivolta C, Sharon D, DeAngelis MM, et al. Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns. *Hum Mol Genet* 2002;11(10):1219-1227.
8. Stone EM, Sheffield VC, Hageman GS. Molecular genetics of age-related macular degeneration. *Hum Mol Genet* 2001;10(20):2285-2292.
9. Johnson AT, Alward WLM, Sheffield VC, et al. Genetics and glaucoma. In: Ritch R, Shields MB, Krupin T, eds. *The Glaucomas.* Vol II. 2nd ed. Chicago: Mosby, 1996:39-54.
10. Marlhens F, Bareil, C, Griffoin, J, et al. Mutations in RPE65 cause Leber's congential amaurosis. *Nat Genet* 1997;17(Oct):139-140.

11. Sohocki MM, Bowne SJ, Sullivan LS, et al. Mutations in a new photoreceptor-pineal gene on 17p cause Leber congenital amaurosis. *Nat Genet* 2000;24(1):79-83.
12. den Hollander AI, ten Brink JB, de Kok YJ, et al. Mutations in a human homologue of *Drosophila* crumbs cause retinitis pigmentosa (RP12). *Nat Genet* 1999;23(2):217-221.
13. Freund CL, Wang QL, Chen S, et al. De novo mutations in the CRX homeobox gene associated with Leber congenital amaurosis. *Nat Genet* 1998;18(4):311-312.
14. Perrault I, Rozet JM, Calvas P, et al. Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis. *Nat Genet* 1996;14(4):461-464.
15. Gerber S, Perrault I, Hanein S, et al. Complete exon-intron structure of the RPGR-interacting protein (RPGRIP1) gene allows the identification of mutations underlying Leber congenital amaurosis. *Eur J Hum Genet* 2001;9(8):561-571.
16. Lotery AJ, Namperumalsamy P, Jacobson SG, et al. Mutation analysis of 3 genes in patients with Leber congenital amaurosis. *Arch Ophthalmol* 2000;118(4):538-543.
17. Acland GM, Aguirre GD, Ray J, et al. Gene therapy restores vision in a canine model of childhood blindness. *Nat Genet* 2001;28(1):92-95.
18. Kahn HA, Moorhead HB. *Statistics on Blindness in the Model Reporting Area 1969-70*. Washington, DC: US Department of Health, Education, and Welfare; 1973:73-427.
19. Kahn HA, Leibowitz HM, Ganley JP, et al. The Framingham Eye Study I. Outline and major prevalence findings. *Am J Epidemiol* 1977;106:17-32.
20. Klein BE, Klein R. Cataracts and macular degeneration in older Americans. *Arch Ophthalmol* 1982;100(April):571-573.
21. Leibowitz HM, Krueger DE, Maunder LR, et al. The Framingham Eye Study monograph: an ophthalmological and epidemiological study of cataract, glaucoma, diabetic retinopathy, macular degeneration, and visual acuity in a general population of 2631 adults, 1973-1975. *Surv Ophthalmol* 1980;24(Suppl):335-610.
22. Leske MC. The epidemiology of open-angle glaucoma: a review. *Am J Epidemiol* 1983;118:166-191.
23. Bonaldo MF, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 1996;6:791-806.
24. Haider NB, Jacobson SG, Cideciyan AV, et al. Mutation of a nuclear receptor gene, NR2E3, causes enhanced S cone syndrome, a disorder of retinal cell fate. *Nat Genet* 2000;24(2):127-131.
25. Stojanovic A, Hwa J. Rhodopsin and retinitis pigmentosa: shedding light on structure and function. *Receptors Channels* 2002;8(1):33-50.
26. Reeves PJ, Kim JM, Khorana HG. Structure and function in rhodopsin: a tetracycline-inducible system in stable mammalian cell lines for high-level expression of opsin mutants. *Proc Natl Acad Sci U S A.* 2002;99(21):13413-13418.
27. Garriga P, Manyosa J. The eye photoreceptor protein rhodopsin. Structural implications for retinal disease. *FEBS Lett* 2002;528(1-3):17-22.
28. Hwa J, Klein-Seetharaman J, Khorana HG. Structure and function in rhodopsin: mass spectrometric identification of the abnormal intradiscal disulfide bond in misfolded retinitis pigmentosa mutants. *Proc Natl Acad Sci U S A* 2001;98(9):4872-4876.
29. Palczewski K, Kumasaka T, Hori T, et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 2000;289(5480):739-745.
30. Teller DC, Okada T, Behnke CA, et al. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry* 2001;40(26):7761-7772.
31. Olsson JE, Gordon JW, Pawlyk BS, et al. Transgenic mice with a rhodopsin mutation (Pro23His): a mouse model of autosomal dominant retinitis pigmentosa. *Neuron* 1992;9:815-830.
32. Naash MI, Hollyfield JG, al-Ubaidi MR, et al. Simulation of human autosomal dominant retinitis pigmentosa in transgenic mice expressing a mutated murine opsin gene. *Proc Natl Acad Sci U S A* 1993;90(12):5499-5503.
33. Petters RM, Alexander CA, Wells KD, et al. Genetically engineered large animal model for studying cone photoreceptor survival and degeneration in retinitis pigmentosa. *Nat Biotechnol* 1997;15(10):965-970.
34. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;89:10915-10919.
35. Stone EM, Lotery AJ, Munier FL, et al. A single EFEMP1 mutation associated with both Malattia Leventinese and Doyne honeycomb retinal dystrophy. *Nat Genet* 1999;Vol. 22(June):199-202.
36. Fingert JH, Heon E, Liebmann JM, et al. Analysis of myocilin mutations in 1703 glaucoma patients from five different populations. *Hum Mol Genet* 1999;8(5):899-905.
37. Donoso LA, Edwards AO, Frost A, et al. Autosomal dominant Stargardt-like macular dystrophy. *Surv Ophthalmol* 2001;46(2):149-163.
38. Dryja TP, McGee TL, Reichel E, et al. A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature* 1990;343:364-366.
39. Kajiwara K, Hahn LB, Mukai S, et al. Mutations in the human retinal degeneration slow gene in autosomal dominant retinitis pigmentosa. *Nature* 1991;354:480-483.
40. Farrar GJ, Kenna P, Siobhan, et al. A three base pair deletion in the peripherin RDS gene in one form of retinitis pigmentosa. *Nature* 1991;354(12):478-483.
41. Nystuen A, Benke PJ, Merren J, et al. A cerebellar ataxia locus identified by DNA pooling to search for linkage disequilibrium in an isolated population from the Cayman Islands. *Hum Mol Genet* 1996;5(4):525-531.
42. Taanman JW. The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta* 1999;1410(2):103-123.
43. Lightowlers RN, Chinnery PF, Turnbull DM, et al. Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. *Trends Genet* 1997;13(11):450-455.

44. Birky CW Jr. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc Natl Acad Sci U S A* 1995;92(25):11331-11338.
45. Birky CW Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu Rev Genet* 2001;35:125-148.
46. Cummins JM, Wakayama T, Yanagimachi R. Fate of microinjected sperm components in the mouse oocyte and embryo. *Zygote* 1997;5(4):301-308.
47. Wallace D. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. 1994 William Allan Award Address. *Am J Hum Genet* 1995;57:201.
48. Lang BF, Gray MW, Burger G. Mitochondrial genome evolution and the origin of eukaryotes. *Annu Rev Genet* 1999;33:351-397.
49. Giles RE, Blanc H, Cann HM, et al. Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* 1980;77(11):6715-6719.
50. Newman NJ. Leber's hereditary optic neuropathy. New genetic considerations. *Arch Neurol* 1993;50(5):540-548.
51. Dayhoff MO, Schwartz R, Orcutt BC. *Atlas of Protein Sequence and Structure*. Silver Spring, MD: National Biomedical Research Foundation; 1978: 5.
52. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11(5):863-874.
53. Heckenlively JR, Rodriguez JA, Daiger SP. Autosomal dominant sectoral retinitis pigmentosa. Two families with transversion mutation in codon 23 of rhodopsin. *Arch Ophthalmol* 1991;109:84-91.
54. Stone EM, Kimura AE, Nichols BE, et al. Regional distribution of retinal degeneration in patients with the proline to histidine mutation in codon 23 of the rhodopsin gene. *Ophthalmology* 1991;98:1806-1813.
55. Kajiwara K, Sandberg MA, Berson EL, et al. A null mutation in the human peripherin/RDS gene in a family with autosomal dominant retinitis punctata albescens. *Nature Genet* 1993;3:208-212.
56. Meins M, Gruning G, Blankenagel A, et al. Heterozygous 'null allele' mutation in the human peripherin/RDS gene. *Hum Mol Genet* 1993;2(12):2181-2182.
57. Apfelstedt-Sylla E, Theischen M, Ruther K, et al. Extensive intrafamilial and interfamilial phenotypic variation among patients with autosomal dominant retinal dystrophy and mutations in the human RDS/peripherin gene. *Br J Ophthalmol* 1995;79:28-34.
58. Dryja TP, Hahn LB, Kajiwara K, et al. Dominant and digenic mutations in the peripherin/RDS and ROM1 genes in the retinitis pigmentosa. *Invest Ophthalmol Vis Sci* 1997;38(10):1972-1982.
59. Wroblewski JJ, Wells JA, Eckstein A, et al. Ocular findings associated with a 3 base pair deletion in the peripherin-RDS gene in autosomal dominant retinitis pigmentosa. *Br J Ophthalmol* 1994;78:831-836.
60. Wells J, Wroblewski J, Keen J, et al. Mutations in the human retinal degeneration slow (RDS) gene can cause either retinitis pigmentosa or macular dystrophy. *Nat Genet* 1993;3:213-218.
61. Souied EH, Rozet JM, Gerber S, et al. Two novel missense mutations in the peripherin/RDS gene in two unrelated French patients with autosomal dominant retinitis pigmentosa. *Eur J Ophthalmol* 1998;8:98-101.
62. Ekstrom U, Andreasson S, Ponjavic V, et al. A Swedish family with a mutation in the peripherin/RDS gene (Arg-172-Trp) associated with a progressive retinal degeneration. *Ophthalmic Genet* 1998;19:149-156.
63. Gruning G, Millan JM, Meins M, et al. Mutations in the human peripherin/RDS gene associated with autosomal dominant retinitis pigmentosa. *Hum Mutat* 1994;3(3):321-323.
64. Vilela C, Beneyto M, Bosch R, et al. Clinical and genetic aspects of two Spanish families with autosomal dominant retinitis pigmentosa(ADRP). *Ophthalmic Genet* 1996;17(1):29-33.
65. Kajiwara K, Berson EL, Dryja TP. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* 1994;264:1604-1608.
66. Budu HS, Matsumoto M, Yamada T, et al. Peripherin/RDS gene mutation (Pro210Leu) and polymorphisms in Japanese patients with retinal dystrophies. *Jpn J Ophthalmol* 2001;45(4):355-358.
67. Ekstrom U, Ponjavic V, Abrahamson M, et al. Phenotypic expression of autosomal dominant retinitis pigmentosa in a Swedish family expressing a Phe-211-Leu variant of peripherin/RDS. *Ophthalmic Genet* 1998;19:27-37.
68. Farrar GJ, Kenna P, Jordan SA, et al. Autosomal dominant retinitis pigmentosa: a novel mutation at the peripherin/RDS locus in the original 6p-linked pedigree. *Genomics* 1992;14(3):805-807.
69. Saga M, Mashima Y, Akeo K, et al. A novel Cys-214-Ser mutation in the peripherin/RDS gene in a Japanese family with autosomal dominant retinitis pigmentosa. *Hum Genet* 1993;92(5):519-521.
70. Yang H, Lou C, Zhou J, et al. The study of RDS gene mutation and clinical phenotype in a family with primary retinitis pigmentosa. *Chung Hua Yen Ko Tsa Chih* 2000;36(1):52-55.
71. Richards SC, Creel DJ. Pattern dystrophy and retinitis pigmentosa caused by a peripherin/RDS mutation. *Retina* 1995;15(1):68-72.
72. Bareil C, Hamel C, Arnaud B, et al. A complex allele (106delTC and IVS2+22ins7) in the peripherin/RDS gene in retinitis pigmentosa with macular dystrophy. *Ophthalmic Genet* 1997;18(3):129-138.
73. Nakazawa M, Kikawa E, Kamio K, et al. Ocular findings in patients with autosomal dominant retinits pigmetosa and transversion mutation in codon 244 (Asn244Lys) of the peripherin/RDS gene. *Arch Ophthalmol* 1994;112:1567-1573.
74. Kikawa F, Nakazawa M, Chida Y, et al. A novel mutation (Asn244Lys) in the peripherin/RDS gene causing autosomal dominant retinitis pigmentosa associated with bull's eye maculopathy detected by nonradioisotopic SSCP. *Genomics* 1994;20:137-139.
75. Bunge S, Wedemann H, David D, et al. Molecular analysis and genetic mapping of the rhodopsin gene in families with autosomal dominant retinitis pigmentosa. *Genomics* 1993;17:230-233.

76. van den Born LI, van Schooneveld MJ, de Jong LA, et al. Thr4Lys rhodopsin mutation is associated with autosomal dominant retinitis pigmentosa of the cone-rod type in a small Dutch family. *Ophthalmic Genet* 1994;15(2):51-60.
77. Kranich H, Bartkowski S, Denton MJ, et al. Autosomal dominant 'sector' retinitis pigmentosa due to a point mutation predicting an Asn-15-Ser substitution of rhodopsin. *Hum Mol Genet* 1993;2(6):813-814.
78. Fujiki K, Hotta Y, Murakami A, et al. Missense mutation of rhodopsin gene codon 15 found in Japanese autosomal dominant retinitis pigmentosa. *Jpn J Hum Genet* 1995;40(3):271-277.
79. Sullivan LJ, Makris GS, Dickinson P, et al. A new codon 15 rhodopsin gene mutation in autosomal dominant retinitis pigmentosa is associated with sectorial disease. *Arch Ophthalmol* 1993;111:1512-1517.
80. Fujiki K, Hotta Y, Hayakawa M, et al. Point mutations of rhodopsin gene found in Japanese families with autosomal dominant retinitis pigmentosa (ADRP). *Jpn J Hum Genet* 1992;37(2):125-132.
81. Hayakawa M, Hotta Y, Imai Y, et al. Clinical features of autosomal dominant retinitis pigmentosa with rhodopsin gene codon 17 mutation and retinal neovascularization in a Japanese patient. *Am J Ophthalmol* 1993;115:168-173.
82. Li ZY JS, Milam AH. Autosomal dominant retinitis pigmentosa caused by the threonine-17-methionine rhodopsin mutation:retinal histopathology and immunocytochemistry. *Exp Eye Res* 1994;58(4):397-408.
83. Dryja TP, Hahn LB, Cowley GS, et al. Mutation spectrum of the rhodopsin gene among patients with autosomal dominant retinitis pigmentosa. *Proc Natl Acad Sci U S A* 1991;88(Oct):9370-9374.
84. Sung C-H, Davenport CM, Hennessey JC, et al. Rhodopsin mutations in autosomal dominant retinitis pigmentosa. *Proc Natl Acad Sci U S A* 1991;88(Aug):6481-6485.
85. Jacobson SG, Kemp CM, Sung C, et al. Retinal function and rhodopsin levels in autosomal dominant retinitis pigmentosa with rhodopsin mutations. *Am J Ophthalmol* 1991;112:256-271.
86. Sandberg MA, Pawlyk BS, Berson EL. Acuity recovery and cone pigment regeneration after a bleach in patients with retinitis pigmentosa and rhodopsin mutations. *Invest Ophthalmol Vis Sci* 1999;40(10):2457-2461.
87. Berson EL, Rosner B, Weigel-DiFranco C, et al. Disease progression in patients with dominant retinitis pigmentosa and rhodopsin mutations. *Invest Ophthal Vis Sci* 2002;43(9):3027-3036.
88. Kemp CM, Jacobson SG, Roman AJ, et al. Abnormal rod dark adaptation in autosomal dominant retinitis pigmentosa with proline-23-histidine rhodopsin mutation. *Am J Ophthalmol* 1992;113:165-174.
89. To K, Adamian M, Dryja TP, et al. Histopathologic study of variation in severity of retinitis pigmentosa due to the dominant rhodopsin mutation Pro23His. *Am J Ophthalmol* 2002;134(2):290-293.
90. Berson EL, Rosner B, Sandberg MA, et al. Ocular findings in patients with autosomal dominant retinitis pigmentosa and a rhodopsin gene defect (pro-23-his). *Arch Ophthalmol* 1991;109:92-101.
91. To K, Adamian M, Dryja TP, et al. Retinal histopathology of an autopsy eye with advanced retinitis pigmentosa in a family with rhodopsin Glu181Lys. *Am J Ophthalmol* 2000;130(6):790-792.
92. Birch DG, Hood DC, Nusinowitz S, et al. Abnormal activation and inactivation mechanisms of rod transduction in patients with autosomal dominant retinitis pigmentosa and the pro-23-his mutation. *Invest Ophthal Vis Sci* 1995;36(8):1603-1614.
93. al-Maghtheh M, Inglehearn C, Lunt P, et al. Two new rhodopsin transversion mutations (L40R; M216K) in families with autosomal dominant retinitis pigmentosa. *Hum Mutat* 1994;3(4):409-410.
94. Kim RY, Al-Maghtheh M, Fitzke FW, et al. Dominant retinitis pigmentosa associated with two rhodopsin gene mutations. Leu-40-Arg and an insertion disrupting the 5-splice junction of exon 5. *Arch Ophthalmol* 1993;111:1518-1524.
95. Reig C, antich J, Gean E, et al. Identification of a novel rhodopsin mutation (Met-44-Thr) in a simplex case of retinitis pigmentosa. *Hum Genet* 1994;94(3):283-286.
96. Macke JP, Davenport CM, Jacobson SG, et al. Identification of novel rhodopsin mutation responsible for retinitis pigmentosa: implications for the structure and function of rhodopsin. *Am J Hum Genet* 1993;53:80-89.
97. Vaithinathan R, Berson EL, Dryja TP. Further screening of the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa. *Genomics* 1994;21:461-463.
98. Inglehearn CF, Keen TJ, Bashir R, et al. A completed screen for mutations of the rhodopsin gene in a panel of patients with autosomal dominant retinitis pigmentosa. *Hum Mol Genet* 1992;1:41-45.
99. Moore AT, Fitzke FW, Kemp CM, et al. Abnormal dark adaptation kinetics in autosomal dominant sector retinitis pigmentosa due to rod opsin mutation. *Br J Ophthalmol* 1992;76:465-469.
100. Dryja TP, McGee TL, Hahn LB, et al. Mutations within the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa. *N Engl J Med* 1990;323(19):1302-1307.
101. Jacobson SG, Kemp CM, Cideciyan AV, et al. Phenotypes of stop codon and splice site rhodopsin mutations causing retinitis pigmentosa. *Invest Ophthalmol Vis Sci* 1994;35:2521-2534.
102. Milam AH, Li Z-Y, Cideciyan AV, et al. Clinicopathologic effects of the Q64ter rhodopsin mutation in retinitis pigmentosa. *Invest Ophthalmol Vis Sci* 1996;37:753-765.
103. Keen TJ, Inglehearn CF, Lester DH, et al. Autosomal dominant retinitis pigmentosa: four new mutations in rhodopsin, one of them in the retinal attachment site. *Genomics* 1991;11:199-205.
104. al-Jandal N, Farrar GJ, Kiang AS, et al. A novel mutation within the rhodopsin gene (Thr-94-Ile) causing autosomal dominant congenital stationary night blindness. *Hum Mutat* 1999;13:75-81.
105. Budu MM, Hayasaka S, Yamada T, et al. Rhodopsin gene codon 106 mutation (Gly-to-Arg) in a Japanese family with autosomal dominant retinitis pigmentosa. *Jpn J Ophthalmol* 2000;44(6):610-614.

106. Ayuso C, Reig C, Garcia-Sandoval B, et al. G106R rhodopsin mutation is also present in Spanish ADRP patients. *Ophthalmic Genet* 1996;17:95-101.
107. Fuchs S, Kranich H, Denton MJ, et al. Three novel rhodopsin nutations (C110F, L131P, A164V) in patients with autosomal dominant retinitis pigmentosa. *Hum Mol Genet* 1994;3:1203.
108. Millan JM, Fuchs S, Paricio N, et al. Gly114Asp mutation of rhodopsin in autosomal dominant retinitis pigmentosa. *Mol Cell Probes* 1995;9(1):67-69.
109. Dryja TP, McEvoy JA, McGee TL, et al. Novel rhodopsin mutations Gly114Val and Gln184Pro in dominant retinitis pigmentosa. *Invest Ophthalmol Vis Sci* 2000;41(10):3124-3127.
110. Souied E, Gerber S, Rozet J-M, et al. Five novel missense mutations of the rhodopsin gene in autosomal dominant retinitis pigmentosa. *Hum Molec Gen* 1994;3(8):1433-1434.
111. Andreasson S, Ehinger B, Abrahamson M, et al. A six-generation famliy with autosomal dominant retinitis pigmentosa and a rhodopsin gene mutation (arginine-135-leucine). *Ophthalmic Paediatr Genet* 1992;13:145-153.
112. Souied E, Soubrane G, Benlian P, et al. Retinitis puntata albescens associated with the Arg135Trp mutation in the rhodopsin gene. *Am J Ophthalmol* 1996;121:19-25.
113. Reig C, Antich J, Gean E, et al. Identification of Arg-135-Leu mutation in the rhodopsin gene in a family with autosomal dominant retinitis pigmentosa. *Med Clin (Barc)* 1996;106(6):219-221.
114. Pannarale MR, Grammatico B, Iannaccone A, et al. Autosomal-dominant retinitis pigmentosa associated with an Arg-135-Trp point mutation of the rhodopsin gene. Clinical features and longitudinal observations. *Ophthalmology* 1996;103(9):1443-1452.
115. Kumaramanickavel G, Maw M, Denton MJ, et al. Missense rhodopsin mutation in a family with recessive RP. *Nat Genet* 1994;8:10-11.
116. Simonelli F, Rinaldi M, Nesti A, et al. Ocular signs associated with a rhodopsin mutation Cys167Arg in a family with autosomal dominant retinitis pigmentosa. *Br J Ophthalmol* 1998;82:709.
117. Antinolo G, Sanchez B, Borrego S, et al. Identification of a new mutation at codon 171 of rhodopson gene causing autosomal dominant retinitis pigmentosa. *Hum Molec Genet* 1994;3(8):1421.
118. Farrar GJ, Kenna P, Redmond R, et al. Autosomal dominant retinitis pigmentosa: a mutation in codon 178 of the rhodopsin gene in two families of Celtic origin. *Genomics* 1991;11:1170-1171.
119. Richards JE, Scott KM, Sieving PA. Disruption of conserved rhodopsin disulfide bond by Cys187Tyr mutation causes early and severe autosomal dominant retinitis pigmentosa. *Ophthalmology* 1995;102(4):669-677.
120. Farrar GJ, Findlay JBC, Kumar-Singh R, et al. Autosomal dominant retinitis pigmentosa: a novel mutation in the rhodopsin gene in the original 3q linked family. *Hum Mol Genet* 1992;1:769-771.
121. Reig C, Llecha N, Antich J, et al. A missense mutation (^{211}His->Arg) and a silent (^{160}Thr) mutation within the rhodopsin gene in a Spanish autosomal dominant retintis pigmentosa famliy. *Hum Mol Genet* 1994;3:195-196.
122. Haim M, Grundmann K, Gal A., et al. Novel rhodopsin mutation (M216R) in a Danish family with autosomal dominant retinitis pigmentosa. *Ophthalmic Genet* 1996;17:193-197.
123. Kucinskas V, Payne AM, Ambrasiene D, et al. Molecular genetic study of autosomal dominant retinitis pigmentosa in Lithuanian patients. *Hum Hered* 1999;49:71-74.
124. Rosenfeld PJ, Cowley GS, McGee TL, et al. A null mutation in the rhodopsin gene causes rod photoreceptors dysfunction and autosomal recessive retinitis pigmentosa. *Nature Genet* 1992;1:209-213.
125. Inglehearn CF, Bashir R, Lester DH, et al. A 3-bp deletion in the rhodopsin gene in a family with autosomal dominant retinitis pigmentosa. *Am J Hum Genet* 1991;48:26-30.
126. Ponjavic V, Abrahamson M, Andreasson S, et al. A mild phenotype of autosomal dominant retinitis pigmentosa is associated with the rhodopsin mutation Pro-267-Leu. *Ophthalmic Genet* 1997;18(2):63-70.
127. Owens SL, Fitzke FW, Inglehearn CF, et al. Ocular manifestations in autosomal dominant retinitis pigmentosa with a Lys-296-Glu rhodopsin mutation at the retinal binding site. *Br J Ophthalmol* 1994;78:353-358.
128. Chan WM, Yeung KY, Pang CP, et al. Rhodopsin mutations in Chinese patients with retinitis pigmentosa. *Br J Ophthalmol* 2001;85(9):1046-1048.
129. Sohocki MM, Daiger SP, Bowne SJ, et al. Prevalence of mutations causing retintitis pigmentosa and other inherited retinopathies. *Hum Mutat* 2001;17(1):42-51.
130. Horn M, Humphries P, Kunisch M, Marchese C, et al. Related deletions in exon 5 of the human rhodopsin gene causing a shift in the reading frame and autosomal dominant retintis pigmentosa. *Hum Genet* 1992;90(3):255-257.
131. Zhao K, Xiong S, Wang L, et al. Novel rhodopsin mutation in a Chinese family with autosomal dominant retinitis pigmentosa. *Ophthalmic Genet* 2001;22(3):155-162.
132. Rosas DJ, Roman AJ, Weissbrod P, et al. Autosomal dominant retinitis pigmentosa in a large family: A clinical and molecular genetic study. *Invest Ophthalmol Vis Sci* 1994;35:3134-3144.
133. Berson EL, Sandberg MA, Dryja TP. Autosomal dominant retinitis pigmentosa with rhodopsin, valine-345-methionine. *Trans Am Ophthmol Soc* 1991;89:117-130.
134. Dikshit MAR. Mutation analysis of codons 345 and 347 of rhodopsin gene in Indian retinitis pigmentosa patients. *J Genet* 2001;80(2):111-116.
135. Macke JP, Hennessey JC, Nathans J. Rhodopsin mutation proline347-to-alanine in a family with autosomal dominant retinitis pigmentosa indicates an important role for proline at position 347. *Hum Mol Genet* 1995;4(4):775-776.
136. Gal A, Artlich A., Ludwig M, et al. Pro-347-Arg mutation of the rhodopsin gene in autosomal dominant retinitis pigmentosa. *Genomics* 1991;11:468-470.

137. Zhang X, Fu W, Pang CP, et al. [Screening for point mutations in rhodopsin gene among one hundred Chinese patients with retinitis pigmentosa]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 2002;19(6):463-466.
138. Berson EL, Rosner B, Sandberg MA, et al. Ocular findings in patients with autosomal dominant retinitis pigmentosa and rhodopsin, proline-347-leucine. *Am J Ophthalmol* 1991;111:614-623.
139. Trujillo MJ, del Rio T, Reig C, et al. The Pro347Leu mutation of the rhodopsin gene in a Spanish family with autosomal dominant retinosis. *Med Clin (Barc)* 1998;110:501-504.
140. Rosenfeld PJ, Hahn LB, Sandberg MA, et al. Low incidence of retinitis pigmentosa among heterozygous carriers of a specific rhodopsin splice site mutation. *Invest Ophthamol Vis Sci* 1995;36(11):2186-2192.

APPENDIX A: MUTATION DETECTION METHODS

The 11 tables in appendices B through L summarize the data my laboratory has collected on the Rhodopsin, *RDS, RP1, AIPL1, CRB1, CRX, GUCY2D, RPE65, RPGRIP1, GLC1A* and *ABCAY* genes. These data are illustrative of the wide range of sequence variations one observes in large human populations and are also illustrative of the value of the EPP method for estimating the pathogenic potential of a large number of variations. These data were collected over a 12-year period, and the methods that were employed varied as the technology available to us improved. However, the majority of the data represented in these tables was collected using a combination of the following PCR and electrophoretic detection methods.

DENATURING GRADIENT GEL ELECTROPHORESIS PCR

Approximately 500 ng of the DNA sample was used in a polymerase chain reaction (PCR) amplification reaction. The coding sequences of the genes were PCR amplified in separate reactions of approximately 200 to 300 base pairs in size. Each pair of primers included one primer with a 5′ 40bp GC-clamp. This sequence was incorporated into the amplified fragment during the PCR amplification. Oligonucleotide primers were synthesized using phosphoramidite chemistry and an Applied Biosystems model 391 DNA synthesizer. Each 100 µL amplification reaction contained: 10 µL of 10 × PCR buffer (67mM Tris pH8.8, 6.7mM $MgCl_2$, 16 mM ammonium sulfate, 10 mM 2-mercaptoethanol), 10% DMSO, dNTPs (final concentration 1.25mM each dNTP), 500 ng of genomic DNA, 50 picomoles of each primer, and 1.5 units of Taq DNA polymerase. The polymerase chain reaction was performed for 40 cycles in a Perkin-Elmer thermocycler.

DENATURING GRADIENT GEL ELECTROPHORESIS

Successful PCR amplification was checked by electrophoresing 10 µL of each sample on a 1.5% agarose gel. Ten to 20 µL of each PCR-amplified product was analyzed on a denaturing gradient gel (8% polyacrylamide/50%–75% denaturant). Samples were electrophoresed at 150 V for 8 hours at 60°C constant temperature. The gel was stained with ethidium bromide and photographed. Base changes were identified by the presence of one or more new bands or a shift in position of a band compared to control samples.

SINGLE-STRAND CONFORMATION POLYMORPHISM ANALYSIS PCR

Twelve and one-half nanograms of each patient's DNA were used as template in a 8.35 µL PCR containing: 1.25 µL 10× buffer (100 mM Tris-HCL pH 8.3, 500 mM KCl, 15 mM $MgCl_2$); 300 µm of each dCTP, dATP, dGTP, and dTTP; 1 pmol of each primer; and 0.25 units Biolase polymerase (Biolase). Samples were denatured for 5 minutes at 94°C and incubated for 35 cycles under the following conditions: 94°C for 30 sec, 55°C for 30 sec, 72°C for 30 sec in a DNA thermocycler (Omnigene).

SINGLE-STRAND CONFORMATION POLYMORPHISM ANALYSIS

After amplification, 5 µL of stop solution (95% formamide, 10 mM NaOH, 0.05% Bromophenol Blue, 0.05% Xylene Cyanol) was added to each sample. Amplification products were denatured for 3 minutes at 94°C and electrophoresed on 6% polyacrylamide, 5% glycerol gels at 25 W for approximately 3 hours at room temperature. Following electrophoresis, gels were stained with silver nitrate.

AUTORADIOGRAPHIC SEQUENCING

DNA product from a 100-mL PCR synthesis was electrophoresed on a 1.5% preparative agarose gel. The gel was stained with ethidium bromide, and the desired band was cut from the agarose gel. The gel fragment was frozen at either –70°C for 15 min or at –20°C overnight and was centrifuged in a Costar 0.22-mm cellulose acetate filter unit. The DNA was ethanol precipitated and resuspended in 15 mL of dH_2O. Sequencing was then performed with 7 mL of the sample by using a USB sequencing kit according to the manufacturer's instructions, with the modification that 10 pmol of primer was used. Sequencing reactions were electrophoresed on 8% polyacrylamide sequencing gels containing 7M urea. Gels were dried and autoradiographed overnight with Kodak X-OMAT film.

AUTOMATED DNA SEQUENCING

Abnormal PCR products identified by single-strand conformation polymorphism analysis were sequenced using fluorescent dideoxynucleotides on an Applied Biosystems (ABI) model 377 automated sequencer. Mutations were identified by the approximately equal peak intensity of two fluorescent dyes at the mutant base. All sequencing was bidirectional.

APPENDIX B: SUMMARY OF VARIATIONS OBSERVED IN THE RHODOPSIN (RHO) GENE IN RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

		UNIVERSITY OF IOWA DATA			WORLD DATA		COMBINED DATA			
AMPLIMER	VARIATION	RP PROBANDS	CONTROLS	M#	RP PROBANDS	M#	TOTAL M#	BLOSUM	EPP	REFERENCE†
1A	Thr4Ala	1	0	0	0	0	0	0	1	
1A	Thr4Lys	0	0	0	3	1	1	-1	2	75, 76
1A	Ans15Ser	0	0	0	18	15	15	1	2	77,78,79
1A	Thr17Met	8	0	6	29	17	23	-1	3	75, 80, 81, 82, 83, 84, 85, 86, 87, 88
1A	Arg21His	1	0	2	0	0	2	0	1	
1A	Pro23Leu	0	0	0	3	1	1	-3	2	83, 89
1A	Pro23Ala	6	0	5	19	17	22	-1	3	84
1A	Pro23His	35	0	15	139	25	40	-2	3	38, 53, 84, 87, 88, 89, 90, 91, 92, 93
1A	Gln28Arg	2	0	1	0	0	1	1	1	
1A	Gln28His	0	0	0	1	0	0	0	1	75
1A	Leu40Arg	0	0	0	4	2	2	-2	2	88, 94, 95
1A	Tyr43Cys	2	0	0	0	0	0	-2	2	
1A	Met44Thr	0	0	0	1	0	0	-1	2	96
1A	Phe45Leu	2	0	0	6	4	4	0	1	84, 87
1A	Leu46Arg	0	0	0	0	0	0	-1	2	
1A	Gly51Ala	5	0	0	3	2	2	0	1	97
1A	Gly51Arg	1	0	0	6	2	2	-2	2	88, 94, 98
1A	Gly51Val	0	0	0	11	7	7	-3	3	83, 88
1A	Pro53Arg	0	0	0	1	0	0	-2	2	99
1A	Thr58Arg	5	0	10	21	12	22	-1	3	75, 88, 99, 100, 101
1B	Thr62Thr	1	0	0	0	0	0	5	0	
1B	Gln64Stop	2	0	0	9	7	7	+	3	97, 101, 103
1B	12bp del at codon 68	0	0	0	1	0	0	+	2	104
1B	Arg69His	1	0	0	0	0	0	0	1	
1B	Thr70Thr	1	0	0	0	0	0	5	0	
1B	Val81Ala	1	0	0	0	0	0	0	1	
1C	Ala82Ala	1	0	0	0	0	0	4	0	
1C	Val87Asp	1	0	2	2	1	3	-3	2	84
1C	Gly89Asp	0	0	0	37	23	23	-1	3	83, 84, 87, 88
1C	Thr94Ile	0	0	0	5	4	4	-1	2	105
1C	3bp del at codon 99	1	0	0	0	0	0	+	2	
1C	Val104Ile	1	0	0	0	0	0	3	1	
1C	Gly106Arg	6	0	7	12	8	15	-2	3	97, 99, 106, 107
1C	Gly106Trp	2	0	2	3	2	4	-2	2	84
1C	Cys110Arg	2	0	4	0	0	4	-3	2	
1C	Cys110Phe	0	0	0	4	3	3	-2	2	108
1C	Cys110Tyr	3	0	0	4	0	0	-2	2	87, 88, 94
1C	Gly114Asp	0	0	0	15	9	9	-1	3	64, 87, 88, 98, 109
1C	Gly120Gly	27	1	0	0	0	0	6	0	
2A	Leu125Arg	1	0	0	4	2	2	-2	2	83, 88
2A	Ser127Phe	0	0	0	2	1	1	-2	2	110
2A	Ser127Ser	1	0	0	0	0	0	4	0	
2A	Leu131Pro	0	0	0	5	3	3	-3	2	108, 111
2A	Glu134Glu	1	0	0	0	0	0	5	0	
2A	Arg135Leu	8	0	0	32	28	28	-2	3	84, 85, 112
2A	Arg135Trp	14	0	3	34	21	24	-3	3	85, 86, 88, 97, 113, 114, 115
2A	Arg135Gly	0	0	0	1	0	0	-2	2	75
2A	Tyr136Stop	1	0	0	0	0	0	+	2	
2A	Cys140Ser	0	0	0	2	1	1	-1	2	97
2A	Arg147Cys	1	0	1	0	0	1	-3	2	
2A	Phe148Phe	1	0	0	0	0	0	6	0	
2A	Glu150Lys	0	0	0	3	2	2	1	1	111
2B	Thr160Thr	4	0	0	0	0	0	5	0	

APPENDIX B: (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE RHODOPSIN (RHO) GENE IN RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

		UNIVERSITY OF IOWA DATA			WORLD DATA		COMBINED DATA			
AMPLIMER	VARIATION	RP PROBANDS	CONTROLS	M#	RP PROBANDS	M#	TOTAL M#	BLOSUM	EPP	REFERENCE†
2B	Ala164Glu	0	0	0	2	0	0	-1	2	98
2B	Ala164Val	0	0	0	3	2	2	0	1	108
2B	Cys167Arg	0	0	0	6	3	3	-3	2	83, 88, 117
2B	Pro170Arg	1	0	0	0	0	0	-2	2	
2B	Pro171Leu	0	0	0	10	4	4	-3	2	83, 87, 88
2B	Pro171Gln	2	0	3	3	2	5	-1	2	118
2B	Pro171Ser	1	0	0	2	1	1	-1	2	98
2B	20bp 3' G->A	1	0	0	0	0	0	-	0	
3	Tyr178Asn	0	0	0	2	1	1	-2	2	111
3	Tyr178Cys	1	0	2	20	17	19	-2	3	84, 86, 119
3	Glu181Glu	1	0	0	0	0	0	5	0	
3	Glu181Lys	4	0	1	9	2	3	1	1	75, 83, 88, 92
3	Gly182Ser	1	0	1	0	0	1	0	1	
3	Gln184Pro	0	0	0	1	2	2	-1	2	110
3	Cys185Arg	1	0	0	0	0	0	-3	2	
3	Ser186Leu	1	0	0	0	0	0	-2	2	
3	Ser186Pro	1	0	0	2	0	0	-1	2	83, 88
3	Ser186Ser	2	0	0	0	0	0	4	0	
3	Cys187Tyr	2	0	0	13	12	12	-2	3	120
3	Gly188Glu	0	0	0	2	0	0	-2	2	88, 97
3	Gly188Arg	0	0	0	5	3	3	-2	2	75, 83
3	Asp190Asn	4	0	3	24	16	19	1	2	83, 88, 104
3	Asp190Tyr	1	0	0	0	0	0	-3	2	
3	Asp190Gly	1	0	0	12	6	6	-1	2	75, 83, 84, 88
3	Thr193Met	1	0	0	0	0	0	-1	2	
3	Met207Arg	0	0	0	8	7	7	-1	3	121
3	Val209Met	0	0	0	1	0	0	1	1	97
3	His211Arg	0	0	0	5	3	3	0	1	97, 122
3	His211Pro	0	0	0	1	0	0	-2	2	104
3	Met216Arg	0	0	0	1	0	0	-1	2	123
3	Met216Lys	1	0	0	0	0	0	-1	2	
3	Phe220Cys	1	0	0	1	0	0	-1	2	75
3	Pro221His	1	0	0	0	0	0	-2	2	
3	Cys222Arg	0	0	0	1	0	0	-3	2	75
3	4bp 3' C->T	37‡	3‡	0	0	0	0	-	0	
4A	Lys248Lys	1	0	0	0	0	0	5	0	
4A	Lys248Arg	0	0	0	1	0	0	2	1	124
4A	Glu249Stop	0	0	0	1	0	0	+	2	125
4A	Arg252Pro	1	0	0	0	0	0	-2	2	
4A	Met253Ile	2	0	0	0	0	0	1	1	
4A	3bp del at codon 255	0	0	0	2	1	1	+	2	126
4A	Ser260Arg	1	0	0	0	0	0	-1	2	
4A	3bp del at codon 264	0	0	0	4	3	3	+	2	98
4A	Pro267Arg	0	0	0	4	3	3	-3	2	111
4A	Pro267Leu	2	0	0	2	1	1	-3	2	127
4B	Lys296Glu	0	0	0	28	23	23	1	2	75, 88, 104, 128
4B	Ser297Arg	1	0	0	3	1	1	-1	2	88, 111
4B	Ala299Ser	0	0	0	1	0	0	1	1	129
4B	Gln312Stop	1	0	0	0	0	0	+	2	
4B	1bp 3' G-> T splice site	0	0	0	2	0	0	+	2	141
5	1bp 5' G-> A splice site	0	0	0	1	2	2	+	2	86
5	3bp del at codons 318 & 319	1	0	0	1	0	0	+	2	130

APPENDIX B: (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE RHODOPSIN (RHO) GENE IN RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

		UNIVERSITY OF IOWA DATA			WORLD DATA		COMBINED DATA			
AMPLIMER	VARIATION	RP PROBANDS	CONTROLS	M#	RP PROBANDS	M#	TOTAL M#	BLOSUM	EPP	REFERENCE†
5	7bp del at codons 318-319	1	0	0	0	0	0	+	2	
5	Cys323Cys	1	0	0	0	0	0	9	0	
5	Cys323Ser	1	0	0	0	0	0	-1	2	
5	Cys323Stop	3	0	1	0	0	1	+	2	
5	17 bp del at codon 332	0	0	0	1	0	0	+	2	130
5	1bp del at codon 335	1	0	0	0	0	0	+	2	
5	1 bp del at codon 340	0	0	0	1	0	0	+	2	131
5	Thr340Met	2	0	0	0	0	0	-1	2	
5	Thr340Thr	2	0	0	0	0	0	5	0	
5	8 bp del at codon 341	0	0	0	1	0	0	+	2	131
5	Glu341Lys	1	0	0	0	0	0	1	1	
5	Glu341Stop	0	0	0	11	10	10	+	3	132
5	Thr342Met	1	0	0	0	0	0	-1	2	
5	Gln344Stop	0	0	0	7	5	5	+	2	84, 85
5	Val345Leu	0	0	0	20	14	14	1	2	88, 98, 133, 134
5	Val345Met	1	0	0	12	8	8	1	2	75, 83, 88, 135
5	Pro347Ala	5	0	14	5	4	18	-1	3	136
5	Pro347Arg	1	0	0	7	5	5	-2	2	75, 137
5	Pro347Cys	1	0	0	0	0	0	-3	2	
5	Pro347Gln	1	0	1	4	3	4	-1	2	98
5	Pro347Leu	10	0	0	75	34	34	-3	3	75, 78, 83, 87, 88, 99, 101, 124, 129, 137, 138, 139, 140
5	Pro347Ser	0	0	0	5	4	4	-1	2	75, 88, 101
5	Pro347Thr	2	0	0	0	0	0	-1	2	
5	Ala348Ser	1	0	0	0	0	0	1	1	

*This table summarizes the sequence variations that we have observed at the University of Iowa as well as those we have found in the published literature. The data in this table are derived from approximately 35,000 polymerase chain reactions. A total of 2,782 probands with RP were screened for variations in the entire coding sequence of the rhodopsin gene. An additional 388 probands were screened only for mutations in amplimers 1A-3 & 5 and an additional 156 RP probands were screened only for mutations in amplimers 1A, 2A & 5. A total of 113 normal control subjects were screened for variations in the entire coding sequence. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

†References were obtained by conducting PubMed and Human Gene Mutation Database searches online. When possible, M numbers were calculated using the cited references.

‡The actual number of occurrences of these variants cannot be accurately determined because they are so common that not every instance was recorded.

APPENDIX C: SUMMARY OF VARIATIONS OBSERVED IN THE PERIPHERIN (RDS) GENE IN RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

		UNIVERSITY OF IOWA DATA			WORLD DATA		COMBINED DATA			
AMPLIMER	VARIATION	RP PROBANDS	CONTROLS	M#	RP PROBANDS	M#	TOTAL M#	BLOSUM	EPP	REFERENCE†
1A	2bp del at codon 25	0	0	0	1	1	1	+	2	54
1A	1bp ins at codon 32	1	0	2	0	0	2	+	2	
1A	Leu45Phe	6	2	-8,00 (2SF)	0	0	-800	0	0	
1B	Arg46Stop	1	0	0	2	1	1	+	2	55, 56
1B	Asn53Ile	1	0	0	0	0	0	-3	2	
1B	Gly68Arg	0	0	0	1	2	2	-2	2	57
1B	Cys72Cys	1	0	0	0	0	0	9	0	
1B	Val106Val	14	0	0	0	0	0	4	0	
1C	3bp del at codons 119-120	1	0	0	2	12	12	+	3	59, 60
1C	Arg123Trp	1	0	0	0	0	0	-3	2	
1C	Leu126Arg	0	0	0	1	4	4	-2	2	57
1C	Gly133Gly	1	0	0	0	0	0	6	0	
1C	Gly137Asp	1	0	0	0	0	0	-1	2	
1C	Asp145Asn	1	0	-4,00 (1SF)	0	0	-400	1	0	
1C	Lys153Arg	1	0	3	0	0	3	2	1	
1C	Ile156Met	1	0	0	0	0	0	1	1	
1C	Ile156Ile	1	0	0	0	0	0	4	0	
1C	Gly51Ala	2	0	0	0	0	0	1	1	
1C	Ile161Ile	4	0	0	0	0	0	4	0	
1C	Cys165Tyr	0	0	0	1	1	1	-2	2	61
1C	Arg172Trp	3	0	0	1	6	6	-3	2	62
1C	Arg172Gln	1	0	0	0	0	0	1	1	
1C	Asp173Val	0	0	0	1	7	7	-3	3	63, 64
1C	Leu185Pro	0	0	0	4	10	10	-3	3	39, 57, 65
2A	5bp del at codons 195-196	1	0	0	0	0	0	+	2	
2A	Gly206Asp	1	0	0	0	0	0	-1	2	
2A	Pro210Arg	1	0	0	0	0	0	-2	2	
2A	Pro210Leu	2	0	0	1	0	0	-3	2	66
2A	Pro210Ser	1	0	1	0	0	1	-1	2	
2A	Phe211Leu	0	0	0	2	9	9	0	2	61, 67
2A	Ser212Gly	0	0	0	1	13	13	0	2	68
2A	Cys214Ser	0	0	0	1	2	2	-1	2	69
2A	Pro216Leu	1	0	0	2	8	8	-3	3	39, 70
2A	Pro216Ser	2	0	2	2	5	7	-1	3	71, 72
2A	3bp del at codon 219	0	0	0	1	7	7	+	3	39
2A	Pro221His	1	0	0	0	0	0	-2	2	
2A	Ser231Stop	1	0	0	0	0	0	+	2	
2B	Tyr236Cys	1	0	0	0	0	0	-2	2	
2B	Tyr236Tyr	1	0	0	0	0	0	7	0	
2B	Gln239Stop	1	0	2	0	0	0	+	2	
2B	Asn244Lys	0	0	0	2	7	7	0	2	73, 74
2B	Gly266Asp	0	0	0	1	5	5	-1	2	57
2B	Thr269Arg	1	0	0	0	0	0	-1	2	
2B	3bp 3' A->T	1	0	0	0	0	0	-	0	
3A	Ser289Leu	1	0	0	0	0	0	-2	2	
3A	Ser303Ser	1	0	0	0	0	0	4	0	
3A	Glu304Gln	40‡	11‡	0	0	0	0	2	0	
3A	1bp del at codon 307	0	0	0	1	7	7	+	3	56

APPENDIX C: (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE PERIPHERIN (RDS) GENE IN RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

		UNIVERSITY OF IOWA DATA			WORLD DATA		COMBINED DATA			
AMPLIMER	VARIATION	RP PROBANDS	CONTROLS	M#	RP PROBANDS	M#	TOTAL M#	BLOSUM	EPP	REFERENCE†
3A	3bp del at codon 309	1	0	0	0	0	0	+	2	
3A	Lys310Arg	5	0	0	0	0	0	2	1	
3A	Pro313Leu	1	0	0	0	0	0	-3	2	
3A	2bp del at codon 319	1	0	5	0	0	5	+	2	
3B	Gly338Asp	14‡	6‡	0	0	0	0	-1	0	
3B	Pro352Ser	7‡	2‡	0	0	0	0	-1	0	
3B	13bp 3' c->T	5	0	0	0	0	0	-	0	

*This table summarizes the sequence variations that we have observed at the University of Iowa as well as those we have found in the published literature. The data in this table are derived from approximately 25,000 polymerase chain reactions. A total of 2,743 probands with RP were screened for variations in the entire coding sequence of the RDS gene. An additional 427 probands were screened only for mutations in amplimers 1A, 2A, and 2B. A total of 113 normal control subjects were screened for variations in the entire coding sequence. M# = The number of correctly segregating affected meioses (see text). SF = Segregation Failure. Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

†References were obtained by conducting PubMed and Human Gene Mutation Database searches online. When possible, M numbers were calculated using the cited references.

‡The actual number of occurrences of these variants cannot be accurately determined because they are so common that not every instance was recorded.

APPENDIX D: SUMMARY OF VARIATIONS OBSERVED IN THE RP1 GENE IN RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

EXON	VARIATION	RP	CONTROLS	BLOSUM	M#	EPP
3	6bp 5' T->C	7	1	-	0	0
2B	Leu76Leu	1	0	4	0	0
2B	Thr93Thr	1	0	5	0	0
4H	Arg677stop	10	0	+	17	3
4H	1 bp ins at codon 658	1	0	+	0	2
4H	Gln689Stop	1	0	+	0	2
4I	Leu749Phe	1	0	0	0	1
4I	1bp del at codon 747	1	0	+	2	2
4J	5bp del at codons 762-763	1	0	+	0	2
4K	35bp del codon 829	1	0	+	0	2
4M	Asn985Tyr heterozygous	72	19	-2	0	0
4M	Asn985Tyr homozygous	33	6	-2	0	0
4N	1bp del at codon 1053	2	0	+	0	2
4U	Leu1417 Val	1	0	1	0	1
4U	Cys1402Phe	1	0	-2	0	2
4X	Arg1595Gln	1	0	-2	0	2
4Y	Ala1670Thr heterozygous	68	39	0	0	0
4Y	Ala1670Thr homozygous	11	1	0	0	0
4Y	Ser1691Pro heterozygous	68	39	-1	0	0
4Y	Ser1691Pro homozygous	11	1	-1	0	0
4Z	Gln1725Gln heterozygous	74	41	5	0	0
4Z	Gln1725Gln homozygous	13	2	5	0	0
4AE	Cys2033Tyr	29	29	-2	0	0

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 13,400 polymerase chain reactions. A total of 185 probands with RP were screened for variations in the entire coding sequence of the RP1 gene. An additional 182 probands were screened only for mutations in amplimers 4H-4K and an additional 2,803 RP probands were screened only for mutations in amplimer 4H only. A total of 96 normal control subjects were screened for variations in the entire coding sequence. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX E: SUMMARY OF VARIATIONS OBSERVED IN THE AIPL1 GENE IN LEBER CONGENITAL AMAUROSIS (LCA), AUTOSOMAL RECESSIVE RETINITIS PIGMENTOSA (ARRP), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	LCA (676 ALLELES)	ARRP (36 ALLELES)	CONTROLS (282 ALLELES)	BLOSUM	FREQ	EPP
1	36bp 3' 1bp del C	9	0	0	-	-	0
1	45bp 3' T->C	1	0	0	-	-	0
2	Phe37Phe	33	0	7	6	-	0
2	Phe70Phe	1	0	0	6	-	0
2	Cys89Cys	1	0	0	9	-	0
2	Asp90His	20	0	2	-1	+	2
2	Asp90Asp	2	0	0	6	-	0
3A	10bp 5' A ->C	39	0	3	-	-	0
3A	2bp 5' A->G splice site	4	0	0	+	-	3
3A	Leu100Leu	54	0	4	4	-	0
3A	Gln105Gln	1	0	0	5	-	0
3A	Thr114Ile	2	0	0	-1	-	3
3B	1bp 3' G->A splice site	1	0	0	+	-	3
3B	34bp 3' 1bp ins T	1	0	0	-	-	0
3B	Tyr134Phe	1	0	1	3	-	0
3B	Gln141His	1	0	0	0	-	2
4	Val180Ile	I	0	0	3	-	2
4	Val196Ile	1	0	0	3	-	2
4	Arg209Arg	0	0	1	5	-	0
5	18bp 5' G->A	57	0	2	-	-	0
5	Pro217Pro	57	0	3	7	-	0
6A	6bp ins at codon 282-283	1	0	0	+	-	3
6A	Trp278Stop	4	0	0	+	-	3
6A	Arg302Leu	8	0	0	-2	+	3

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 4,770 polymerase chain reactions. A total of 338 probands with LCA and 18 with ARRP were screened for variations in the entire coding sequence of the AIPL1 gene. A total of 141 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX F: SUMMARY OF VARIATIONS OBSERVED IN THE CRB1 GENE IN LEBER CONGENITAL AMAUROSIS (LCA) AUTOSOMAL RECESSIVE RETINITIS PIGMENTOSA (RP) AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	LCA (588 ALLELES)	ARRP (252 ALLELES)	CONTROLS (418 ALLELES)	BLOSUM	FREQ	EPP
2A	12bp 5' A	339	55	286	-	-	0
2A	12bp 5' T	249	77	179	-	-	0
2A	1bp ins at codon 38	1	0	0	+	-	3
2A/2B	2bp ins at codon 86-87	2	0	0	+	-	3
2B	5bp del at codon 143-144	1	0	0	+	-	3
2B	Phe144Val	1	0	0	-1	-	3
2B	Val162Met	0	0	1	1	-	0
2C	7bp del at codon 204-207	5	0	0	+	+	3
3A	48-51bp 5' 4bp del	1	0	0	-	-	0
4	31bp 5' C->T	0	1	0	-	-	0
4	32bp 5' C->T	0	0	1	-	-	0
4	35bp 5' C->T	4	0	0	-	-	0
4	35bp 3' C->T	3	0	0	-	-	0
5	54bp 5' G->T	1	1	0	-	-	0
4	Thr289Met	1	0	0	-1	-	3
5	Asn351Asn	0	1	0	6	-	0
5	Cys383Tyr	1	0	0	-2	-	3
6B	Leu470Leu	1	2	0	4	-	0
6B	Thr476Thr	1	0	0	5	-	0
6B	Cys480Arg	2	0	0	-3	-	3
6B	Cys480Gly	1	0	0	-3	-	3
6C	Ala511Ala	1	0	0	4	-	0
6C	Asn549Asn	1	0	0	6	-	0
6E	Cys681Tyr	1	0	0	-2	-	3
7A	3bp del at codon 749	1	0	0	+	-	3
7A	Leu753Pro	1	0	0	-3	-	3
7B	Arg764Cys	2	0	0	-3	-	3
7B	Arg769Arg	1	0	0	5	-	0
7B	Arg769His	0	0	1	1	-	0
7B	Lys801Stop	4	0	0	+	-	3
7C	4bp del at codons 850-851	1	0	0	+	-	3
7C	1bp ins at codon 871	0	0	0	+	-	3
8	Asn894Ser	0	1	0	1	-	2
8	Pro941Pro	0	0	1	7	-	0
9A	Cys948Tyr	10	1	0	-2	+	3
9B	Asn1057Asn	1	0	0	6	-	0
9C	Val1133Met	0	0	1	1	-	0
9D	Gly1205Arg	1	0	0	-2	-	3
9D	Cys1218Phe	0	1	0	-2	-	3
11A	Asn1317His	1	0	0	1	-	2
11B	Arg1331His	2	0	1	0	-	0
11B	Cys1332Stop	2	0	0	+	-	3
11B	10bp 3' G->C	0	0	1	-	-	0

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 20,380 polymerase chain reactions. A total of 294 probands with LCA and 126 with ARRP were screened for variations in the entire coding sequence of the CRB1 gene. A total of 209 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX G: SUMMARY OF VARIATIONS OBSERVED IN THE CRX GENE IN LEBER CONGENITAL AMAUROSIS (LCA), CONE-ROD DYSTROPHY (CRD), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	LCA (676 ALLELES)	CRD (618 ALLELES)	CONTROLS (208 ALLELES)	BLOSUM	FREQ	EPP
1	12bp 3' C->T	93	89	30	-	-	0
2	Arg41Gln	0	1	0	1	-	2
2	Arg41Trp	0	1	0	-3	-	3
2	Ala56Thr	1	0	0	0	-	2
3A	Gly122Asp	1	3	0	-1	-	3
3B	Ala158Thr	2	4	0	0	-	2
3B	2bp del at codon 168	1	0	0	+	-	3
3B	5bp ins at codon 178-180	0	1	0	+	-	3
3B	1bp del at codon 145	1	0	0	+	-	3
3B	1 bp ins codon 190	1	0	0	+	-	3
3B	1bp del at codon 168	0	1	0	+	-	3
3C	1bp del at codon 217	1	0	0	+	-	3
3C	4bp del at codon 196-197	0	1	0	+	-	3
3C	Val242Met	0	2	0	1	-	2
3D	1bp del at codon 239	0	1	0	+	-	3
3E	Thr273Met	0	1	0	-1	-	3

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 6,310 polymerase chain reactions. A total of 338 probands with LCA and 309 with CRD were screened for variations in the entire coding sequence of the CRX gene. A total of 104 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX H. SUMMARY OF VARIATIONS OBSERVED IN THE GUCY2D GENE IN LEBER CONGENITAL AMAUROSIS (LCA) AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	LCA	CONTROLS	BLOSUM (676 ALLELES)	FREQ (192 ALLELES)	EPP
2A	Trp21Arg	12	4	-3	-	0
2A	6bp del at codons 42-45	1	1	+	-	0
2A	Ala52Ser	151	63	1	-	0
2B	Glu97Gln	1	0	2	-	2
2B	Glu103Lys	1	0	1	-	2
3	His247His	16	0	8	-	0
3	Thr312Met	1	0	-1	-	3
4	Val373Val	11	2	4	-	0
4	Ser448Stop	2	0	+	+	3
4	Cys457Cys	1	0	9	-	0
6	Leu513Phe	1	0	0	-	2
6	21bp 3' G->A	1	0	-	-	0
8	Pro575Leu	2	0	-3	-	3
9	21bp 3' G->T	1	0	-	-	0
10	Arg660Gln	1	0	1	-	2
10	Arg660Stop	2	0	+	-	3
10	Arg677Arg	0	1	5	-	0
10	Pro698Ser	2	1	-1	-	0
10	1bp del at codon 701	2	0	+	-	3
10	Pro701Ser	19	1	-1	+	2
10	Ala703Ala	41	4	4	-	0
11	Tyr746Cys	4	0	-2	-	3
11	Glu750Stop	2	0	+	-	3
12	Arg768Trp	5	0	-3	-	3
12	Met773Leu	2	0	2	-	2
12	Leu782His	7	7	-3	-	0
13	Thr839Ala	1	0	0	-	2
14	31bp 5' C->T	6	1	-	-	0
14	Pro859Pro	6	1	7	-	0
15	1bp del G 1bp 3' splice site	1	0	+	-	3
16	Cys984Tyr	1	0	-2	-	3
17	7bp 5' G->T	20	4	-	-	0
17	Arg1029Ser	1	0	-1	-	3
17	Arg1040Gly	7	0	-2	+	3
18	Gly1061Ser	1	0	0	-	2
19	7bp 5' C->T	1	2	-	-	0
19	Leu1094Leu	1	0	4	-	0
19	Pro1099Pro	4	3	7	-	0
20	143bp 3' T->C	8	1	-	-	0
20	190bp 3' A->G	1	0	-	-	0

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 11,460 polymerase chain reactions. A total of 338 probands with LCA were screened for variations in the entire coding sequence of the GUCY2D gene. A total of 96 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX I: SUMMARY OF VARIATIONS OBSERVED IN THE RPE65 GENE IN LEBER CONGENITAL AMAUROSIS (LCA), AUTOSOMAL RECESSIVE RETINITIS PIGMENTOSA (RP), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	LCA (676 ALLELES)	ARRP (280 ALLELES)	CONTROLS (192 ALLELES)	BLOSUM	FREQ	EPP
1	5bp 3' G->A splice site?	5	0	0	-	-	0
3	Gly40Ser	1	0	0	0	-	2
3	1bp del at codon 46	1	0	0	+	-	3
4	Arg91Gln	1	0	0	0	-	2
4	Arg91Trp	7	0	0	-3	+	3
4	20bp del at codon 97	2	0	0	+	-	3
5	Ala132Thr	2	0	0	0	-	2
8	23bp 5' A->G	1	0	0	-	-	0
8	Tyr239Asp	2	0	0	-3	-	3
9	Val287Phe	2	0	0	-1	-	3
9	Lys294Thr	3	1	0	-1	-	3
9	1bp del at codon 297-298	1	0	0	+	-	3
9	Tyr318Asn	1	0	0	-2	-	3
9	Asn321Lys	3	0	1	-3	-	0
9	Val326 Val	2	0	0	4	-	0
10	Glu352Glu	32	8	4	5	-	0
10	1bp ins at codon 357	1	0	0	+	-	3
10	Glu364Glu	1	0	0	5	-	0
10	Tyr368His	2	0	0	2	-	2
11	Ala360Pro	0	1	0	-1	-	3
11	Thr385Thr	2	0	0	5	-	0
11	Ala393Glu	1	0	0	-1	-	3
11	Leu408Pro	2	0	0	-3	-	3
11	29bp 3' G->A	2	0	1	-	-	0
12	Glu417Gln	0	1	0	0	-	2
12	22bp del at codon 427	0	1	0	+	-	3
12	Ala434Val	2	1	0	0	-	2
12	20bp 3' A->C	0	1	0	-	-	0
14	Gly484asp	0	1	0	-1	-	3
14	Ile520Thr	0	1	0	-1	-	3

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 9,640 polymerase chain reactions. A total of 338 probands with LCA and 140 with ARRP were screened for variations in the entire coding sequence of the RPE65 gene. A total of 96 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX J: SUMMARY OF VARIATIONS OBSERVED IN THE RPGRIP1 GENE IN LEBER CONGENITAL AMAUROSIS (LCA) AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	LCA (550 ALLELES)	CONTROLS (256 ALLELES)	BLOSUM	FREQ	EPP
3B	Leu150Leu	2	0	4	-	0
4	Pro175Pro	56	21	7	-	0
4	Glu188Glu	1	0	5	-	0
4	Lys192Glu	76	41	1	-	0
5B	Asp248His	1	0	-1	-	3
8	His320Pro	1	0	-2	-	3
10	Glu400Glu	1	0	5	-	0
13	Ala547Ser	31	13	1	-	0
13	Arg580Gly	2	0	-2	-	3
13	Pro585Ser	0	1	-1	-	0
14A	Gln589His	0	2	0	-	0
14A	Arg598Gln	1	0	1	-	2
14A	Pro599Pro	0	1	7	-	0
14B	7bp 3' G->A	77	33	-	-	0
15	Leu762Leu	1	0	4	-	0
15	13bp 5' T->G	0	1	-	-	0
16A	Arg812Gln	2	0	1	-	2
16B	Asp877Gly	1	0	-1	-	3
17	13bp 5' T->G	0	1	-	-	0
18A	Ile975Thr	1	0	-1	-	3
18A	Val999Ala	0	1	0	-	0
18B	1bp 3' G->C splice site	115	27	+	-	1
19	His1057His	1	0	8	-	0
21	1bp del A at codon 1164	1	0	+	-	3
21	15bp 5' C->T	1	0	-	-	0
22	Asp1182Asp	5	5	6	-	0
23	Gly1240Glu	2	0	-2	-	3

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 12,700 polymerase chain reactions. A total of 275 probands with LCA were screened for variations in the entire coding sequence of the RPGRIP1 gene. A total of 128 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX K: SUMMARY OF VARIATIONS OBSERVED IN THE GLC1A GENE IN GLAUCOMA, JUVENILE OPEN-ANGLE GLAUCOMA (JOAG), NORMAL TENSION GLAUCOMA (NTG), OCULAR HYPERTENSION (OHT) AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	GLAUCOMA	JOAG	NTG	OHT	CONTROLS	BLOSUM	M #	EPP
1A	83bp 5' G->A	130	13	16	14	27	-	0	0
1B	Cys9Ser	0	0	0	0	1	-1	0	0
1B	Gly12Arg	1	0	0	1	2	-2	0	0
1B	Pro13Pro	0	11	0	0	13	7	0	0
1B	Val18Leu	0	0	0	0	1	1	0	0
1B	Gln19His	1	0	2	0	1	1	0	0
1B	Arg46Stop	1	0	0	0	0	+	0	2
1C	Ser69Ser	1	0	0	0	0	4	0	0
1C	Val70Val	1	0	0	0	0	4	0	0
1C	Asn73Ser	0	0	0	0	1	1	0	0
1C	Arg76Lys	46	3	37	3	25	2	1	0
1C	Arg82Cys	1	0	0	0	0	-3	2	2
1C	Arg82His	0	0	0	0	1	0	1	0
1C	Thr88Thr	0	0	0	0	1	5	0	0
1D	Glu96Glu	0	0	0	0	1	5	0	0
1D	Gly122Gly	2	1	0	1	0	0	0	0
1D	Thr123Thr	1	0	0	0	0	5	0	0
1E	Leu159Leu	0	5	0	0	0	4	0	0
1E	1bp ins at codon 162	1	0	0	1	0	+	0	2
1E	Arg189Glu	0	0	0	0	1	0	0	0
1E	14bp 3' G->A	0	1	0	0	0	-	0	0
1E	19bp 3' G->C	1	0	0	1	3	-	0	0
2	Ser203Phe	0	0	0	0	1	-2	0	0
2	Thr204Met	1	0	0	0	0	-1	0	2
2	Thr204Thr	0	0	0	0	1	5	0	0
2	Asp208Glu	2	1	1	1	0	2	0	1
2	Leu215Pro	0	0	2	0	0	-3	0	2
3A	Thr256Met	0	0	1	0	0	-1	0	2
3A	Lys266Lys	1	0	0	0	0	5	0	0
3B	Thr285Thr	7	2	0	1	1	5	0	0
3B	Trp286Arg	1	0	1	0	0	-3	0	2
3B	Thr290Ala	0	1	0	0	0	1	0	1
3B	Thr293Lys	2	1	0	0	0	-1	0	2
3B	Leu318Leu	0	0	1	0	0	4	0	0
3B	Thr325Thr	0	8	1	0	18	5	0	0
3B	Val329Val	1	0	0	0	0	4	0	0
3B	Val329Met	1	0	0	0	1	1	0	0
3C	Ser331Ser	1	0	0	0	0	4	0	0
3C	Gln337Arg	1	0	0	0	0	0	0	0
3C	Tyr347Tyr	49	3	6	28	9	7	0	0
3C	Tyr347Stop	0	0	0	1	0	+	0	2
3C	Thr351Thr	0	1	0	1	0	5	0	0
3C	Glu352Lys	1	0	0	1	0	1	0	1
3C	Thr353Ile	1	0	0	0	0	-1	0	2
3C	Pro361Ser	1	0	0	0	0	-1	0	2
3C	Gly364Val	0	1	0	0	0	-3	15	3
3C	Gln368Stop	15	1	0	4	0	+	4	2
3C	Pro370Pro	1	0	0	1	0	7	0	0
3C	Gln377Arg	1	0	0	0	0	0	0	
3D	Asp380Gly	0	1	0	0	0	-1	2	2
3D	Glu396Glu	0	0	0	5	3	5	0	0
3D	6bp ins at codon 396	0	1	0	0	0	+	0	2
3D	Lys398Arg	8	0	0	6	4	2	0	0
3D	Val402Ile	0	0	0	0	1	3	0	0
3D	Arg422His	0	0	0	0	0	0	0	0
3E	Tyr437His	0	2	0	0	0	2	39	2
3E	Thr438Thr	0	0	0	1	0	5	0	0
3E	Val439Val	1	0	0	0	0	4	0	0
3E	Ala445Val	0	0	0	1	0	0	0	0
3E	1bp del at codon 453	0	1	0	0	0	+	0	2

APPENDIX K: (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE GLC1A GENE IN GLAUCOMA, JUVENILE OPEN-ANGLE GLAUCOMA (JOAG), NORMAL TENSION GLAUCOMA (NTG), OCULAR HYPERTENSION (OHT) AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	GLAUCOMA	JOAG	NTG	OHT	CONTROLS	BLOSUM	M #	EPP
3E	Ile465Met	1	0	0	0	0	1	0	1
3E	Arg470Cys	2	0	0	0	0	-3	0	2
3F	Ile477Asn	0	1	0	0	0	-3	32	2
3F	Pro481Thr	0	1	0	0	0	-1	1	2
3F	Ala488Ala	0	0	0	1	0	4	0	0
3F	Val495Ile	1	0	0	0	0	3	1	1
3F	Lys500Arg	0	0	0	0	1	2	0	0

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table are derived from approximately 27,028 polymerase chain reactions. A total of 969 probands with glaucoma were screened for variations in the entire coding sequence of the GLC1A gene. An additional 484 probands and 104 normal controls were screened for mutations in all amplimers except 1A, and an additional 620 probands were screened only for mutations in amplimers 3C, 3D, 3E, and 3F. A total of 176 normal control subjects were screened for variations in the entire coding sequence. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX L. SUMMARY OF VARIATIONS OBSERVED IN THE ABCA4 GENE IN AGE-RELATED MACULAR DEGENERATON (AMD), STARGARDT DISEASE (STGD), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	AMD (364 ALLELES)	CONTROLS (192 ALLELES)	STGD (748 ALLELES)	BLOSUM	FREQ	EPP
2	1bp del at codon 36	0	0	1	+	-	3
2	1bp 3' G->A splice site	0	0	1	+	-	3
3	45bp 5' T->G	0	0	1	-	-	0
3	19bp 5' G->A	0	0	1	-	-	0
3	Cys54Tyr	0	0	6	-2	-	3
3	Ala60Val	0	0	2	0	-	2
3	Gly65Glu	0	0	2	-2	-	3
3	Cys75Gly	0	0	2	-3	-	3
3	Asn78Asn	0	0	1	6	-	0
3	4bp del at codon 83-84	0	0	2	+	-	3
3	Ser100Pro	0	0	1	-1	-	3
3	20bp 3' C->T	0	0	2	-	-	0
3	26bp 3' G->A	1	1	9	-	-	0
5	Arg152Stop	0	0	2	+	-	3
6	Ala192Thr	0	0	1	0	-	2
6	Ser206Arg	0	0	3	-1	-	3
6	Arg212Cys	0	0	7	-3	+	3
6	Arg212His	2	2	6	0	-	0
6	Arg220Cys	0	0	2	-3	-	3
6	1bp del at codon 221	0	0	1	+	-	3
6	13bp del at codon 222-226	0	0	1	+	-	3
6	Asp249Gly	0	0	1	-2	-	3
7	32bp 5' T->C	3	0	10	-	-	0
8	11bp 5' C->T	0	0	1	-	-	0
8	Pro291Pro	0	0	1	7	-	3
8	Thr300Asn	0	0	1	0	-	2
8	Pro327Pro	0	0	1	7	-	3
8	Glu328Stop	0	0	0	+	-	3
8	Arg333Trp	0	0	1	-3	-	3
9	14bp 5' T->C	0	0	2	-	-	0
9	Asn380Lys	0	0	1	-3	-	3
9	Arg408Stop	0	0	1	+	-	3
10	14bp 5' T->C	2	4	6	-	-	0
10	Ser416Ser	1	0	0	4	-	0
10	His423Arg	1	0	7	0	-	1
10	His423His	2	0	2	8	-	0
10	Ser445Arg	0	0	1	-1	-	3
10	1bp del at codon 448	0	0	1	+	-	3
10	1bp del 5bp 3'	18	3	37	-	-	0
10	6bp 3' G->C	0	0	1	-	-	0
11	Glu471Lys	0	0	3	-3	-	3
11	5bp del at codon 505-506	0	0	1	+	-	3
12	Leu541Pro	0	0	11	-3	+	3
12	Leu541Leu	0	0	1	4	-	0
12	Val551Val	0	0	1	4	-	0
12	22bp 3' G->T	0	0	1	-	-	0
13	64bp 5' G->A	0	0	0	-	-	0
13	54bp 5' G->A	3	9	41	-	-	0
13	50bp 5' G->A	0	0	1	-	-	0
13	37bp 5' G->A	0	0	1	-	-	0
13	Arg602Trp	0	0	3	-3	-	3
13	Arg602Gln	0	0	1	1	-	2
13	Gly607Trp	0	0	1	-2	-	3
13	Phe608Ile	0	0	1	0	-	2
13	Val643Met	0	0	1	1	-	2
14	Trp663Stop	0	0	1	+	-	3
14	2bp del at codon 669	0	0	3	+	-	3
14	Arg681Stop	0	0	2	+	-	3
14	Ser709Ser	0	0	1	4	-	0

APPENDIX L. (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE ABCA4 GENE IN AGE-RELATED MACULAR DEGENERATON (AMD), STARGARDT DISEASE (STGD), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	AMD (364 ALLELES)	CONTROLS (192 ALLELES)	STGD (748 ALLELES)	BLOSUM	FREQ	EPP
14	Thr716Met	0	0	1	-1	-	3
15	Cys764Tyr	0	0	1	-2	-	3
15	Ser765Asn	0	0	1	1	-	2
15	Val767Asp	0	0	2	-3	-	3
16	10bp 5' C->G	0	0	2	-	-	0
16	16bp del at codon 795-800	0	0	1	+	-	3
16	Gly818Glu	0	0	1	-2	-	3
16	Trp821Arg	0	0	1	-3	-	3
16	Val849Ala	0	0	4	0	-	2
16	Gly851Asp	0	0	1	-1	-	3
16	Ala854Thr	0	0	1	0	-	2
17	12bp 5' C-.G	1	0	0	-	-	0
17	Gly863Ala	2	2	28	0	-	1
17	Phe873Leu	0	0	1	0	-	2
17	60bp 3' G->C	6	0	2	-	-	0
18	48bp 5' G->C	4	0	9	-	-	0
18	47bp 5' T->C	0	0	0	-	-	0
18	21bp 5' A->T	0	0	1	-	-	0
18	Thr897Ile	0	0	1	-1	-	3
18	Thr901Ala	0	1	0	1	-	0
18	Thr901Arg	0	0	2	-1	-	3
19	56bp 5' G->A	0	0	1	-	-	0
19	Arg943Gln	20	13	37	1	-	0
19	Thr959Thr	1	1	0	5	-	0
19	1bp del at codon 961	0	0	1	+	-	3
19	Asn965Ser	0	0	3	1	-	2
19	Thr971Asn	0	0	1	0	-	2
19	Thr972Asn	0	0	1	0	-	2
20	Ser974Pro	0	0	1	-1	-	3
20	Leu988Leu	0	0	4	4	-	0
20	Val989Ala	0	0	2	0	-	2
20	8bp del 993-995	0	0	1	+	-	3
20	Leu1014Arg	0	0	1	-2	-	3
20	61bp 3' G->A	0	0	1	-	-	0
21	14bp 5' T->A	0	0	3	-	-	0
21	Thr1019Ala	0	0	1	0	-	2
21	Glu1022Lys	0	0	1	1	-	2
21	Lys1031Glu	0	0	1	1	-	2
21	Ala1038Val	1	0	17	0	+	3
21	83bp 3' A->T	0	0	6	-	-	0
22	2bp ins at codon 1068-1069	0	0	1	+	-	3
22	Glu1087Lys	0	0	2	1	-	2
22	Arg1108Cys	0	0	6	-3	+	3
22	Arg1108His	0	0	1	0	-	2
23	Glu1122Lys	0	0	1	1	-	2
23	Arg1129Leu	0	0	3	-2	-	3
23	31bp 3' C->A	0	0	0	-	-	0
24	12bp 5' C->T	0	1	0	-	-	0
24	Cys1158Stop	0	0	1	+	-	3
24	32bp 3' G->A	0	0	1	-	-	0
25	16bp 5' T->A	0	0	1	-	-	0
25	Leu1232Leu	0	0	1	4	-	0
25	Leu1250Pro	0	0	1	-3	-	3
26	6bp del at codon 1279-1280	0	0	1	+	-	3
27	Pro1314Thr	0	1	0	-1	-	0
28	36bp 5' A->T	1	0	0	-	-	0
28	Pro1380Leu	0	0	10	-3	+	3
28	Pro1401Pro	6	9	56	7	-	0
28	Trp1408Arg	0	0	2	-3	-	3

APPENDIX L. (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE ABCA4 GENE IN AGE-RELATED MACULAR DEGENERATON (AMD), STARGARDT DISEASE (STGD), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	AMD (364 ALLELES)	CONTROLS (192 ALLELES)	STGD (748 ALLELES)	BLOSUM	FREQ	EPP
28	Trp1408Leu	0	0	2	-2	-	3
28	Gln1412stop	0	0	1	+	-	3
28	4bp 3' C->T	0	0	1	-	-	0
29	47bp 5' T->C	3	0	1	-	-	0
29	38bp 5' G->A	0	0	10	-	-	0
29	Val1433Ile	1	0	0	3	-	0
29	Phe1440Ser	0	0	1	-2	-	3
29	32bp 3' A->G	0	0	4	-	-	0
29	54bp 3' G->A	0	0	0	-	-	0
30	1bp 5' G->T splice site	0	0	1	+	-	3
30	Pro1486Leu	0	0	1	-3	-	3
30	Cys1488Arg	0	0	3	-3	-	3
30	Cys1488Phe	0	0	2	-2	-	3
30	Cys1490Tyr	0	0	3	-2	-	3
30	Cys1502Cys	0	0	2	9	-	0
30	1bp ins at codon 1511	0	0	2	+	-	3
30	1bp 3' G->T splice site	0	0	1	+	-	3
30	3bp 3' G->A	0	0	2	-	-	0
30	1bp del 21bp 3'	1	0	0	-	-	0
30	35bp 3' G->C	0	0	1	-	-	0
30	40bp 3' C->T	0	0	1	-	-	0
30	Gln1513Arg	0	0	1	1	-	2
30	1bp 3' G->T splice site	0	0	1	+	-	3
31	Leu1525Pro	0	0	1	-3	-	3
33	4bp del at codon 1578-1579	0	0	1	+	-	3
33	Asp1582Asp	0	0	1	6	-	0
33	48bp 3' C->T	1	0	2	-	-	0
35	6bp del and 4bp ins at codon 1620-1621	0	0	1	+	-	3
35	Ala1637Thr	0	0	1	0	-	2
35	Arg1640Trp	0	0	1	-3	-	3
35	Arg1640Gln	0	0	1	1	-	2
35	Tyr1652Asp	0	0	1	-2	-	3
36	Val1693Ile	0	0	1	3	-	2
36	Leu1729Pro	0	0	2	-3	-	3
36	20bp 3' G->A	0	0	1	-	-	2
37	Ser1736Pro	0	0	1	-1	-	3
37	11bp del at codon 1738-1741	0	0	1	+	-	3
37	11bp del at codon 1742-1745	0	0	1	+	-	3
37	9bp del codon 17960-1762	0	0	1	+	-	3
37	1bp del at codon 1763	0	0	1	+	-	3
37	45bp 3' G->T	0	0	1	-	-	0
38	Asn1799Asp	0	0	1	1	-	2
38	Asp1817Glu	1	0	4	2	-	1
39	1bp ins 51bp 5'	36	34	85	-	-	0
39	10bp 5' T->C	0	1	18	-	-	0
39	5bp 3' G->A	0	0	6	-	-	0
40	70bp 5' T->C	20	8	79	-	-	0
40	Asn1868Ile	20	7	79	-3	-	0
40	Val1884Glu	0	0	1	-2	-	3
40	Gly1886Glu	0	0	1	-2	-	3
40	Leu1894Leu	62	30	176	4	-	0
40	Val1896Asp	0	0	1	-3	-	3
40	Arg1898His	0	0	1	0	-	2
40	5bp 3' G->A	0	0	1	-	-	0
41	24bp 5' A->C	3	1	8	-	-	0
41	Leu1938Leu	20	7	50	4	-	0
41	37bp 3' A->C	0	0	1	-	-	0
42	3bp 5' G->A	0	0	1	-	-	0

APPENDIX L. (CONT.) SUMMARY OF VARIATIONS OBSERVED IN THE ABCA4 GENE IN AGE-RELATED MACULAR DEGENERATON (AMD), STARGARDT DISEASE (STGD), AND NORMAL CONTROL SUBJECTS*

AMPLIMER	VARIATION	AMD (364 ALLELES)	CONTROLS (192 ALLELES)	STGD (748 ALLELES)	BLOSUM	FREQ	EPP
42	43bp 5' C->A	52	28	117	-	-	0
42	24bp 5' G->A	1	0	0	-	-	0
42	11bp 5' G->A	52	28	115	-	-	0
42	Pro1948Leu	11	7	28	-3	-	0
42	Pro1948Pro	41	22	92	7	-	0
42	Asp1956Asp	0	0	1	6	-	0
42	Gly1961Glu	1	0	43	-2	+	3
42	22bp 3' C->A	2	0	0	-	-	0
43	Leu1970Phe	1	0	1	0	-	2
43	1bp del at codon 1973	0	0	1	+	-	3
44	17bp 5' G->A	1	1	0	-	-	0
44	16bp 5' G->A	3	0	4	-	-	0
44	Ile2023Ile	17	12	62	4	-	0
44	Leu2027Phe	0	0	9	0	+	3
44	Arg2030Stop	0	0	2	+	-	3
44	Arg2030Gln	0	0	1	1	-	2
44	Arg2038Trp	0	0	1	-3	-	3
45	Val2050Leu	1	0	0	1	-	0
45	Tyr2071Phe	0	0	1	3	-	2
45	Arg2077Trp	0	0	2	-3	-	3
45	Ile2083Ile	22	14	55	4	-	0
45	Leu2085Leu	0	1	1	4	-	0
45	7bp 3' G->A	1	4	8	-	-	0
46	Asp2095Asp	2	1	30	5	-	0
46	Arg2107His	0	0	10	0	+	3
46	Val2114Val	0	0	1	4	-	0
46	His2128Arg	0	0	1	0	-	2
47	Arg2149Leu	0	0	1	-2	-	3
47	Cys2150Tyr	0	0	5	-2	-	3
48	Asp2177Asn	2	0	0	1	-	0
48	Leu2229Pro	0	0	1	-3	-	3
48	8bp del at codon 2236-2238	0	0	1	+	-	3
48	1bp 3' G->A splice site	0	0	1	+	-	3
48	21bp 3' C->T	0	0	13	-	-	0
49	27bp 5' C->G	0	0	2	-	-	0
49	3bp 5' T->C	16	4	52	-	-	0
49	Val2244Val	0	0	1	4	-	0
49	Ser2255Ile	16	4	54	-2	-	0
49	28bp 3' C->G	2	0	4	-	-	0
50	85bp 5' C->T	0	0	1	-	-	0
50	26bp 3' C->A	1	0	0	-	-	0

*This table summarizes the sequence variations that we have observed at the University of Iowa. The data in this table have been previously published4 and are derived from approximately 33,252 polymerase chain reactions. A total of 374 probands with Stargardt disease and 182 with AMD were screened for variations in the entire coding sequence of the ABCA4 gene. A total of 96 normal control subjects were screened for variations in the entire coding sequence. Freq = The compatibility of the data with the hypothesis that the variant exists in a ratio of greater than or equal to 100:1 in disease alleles with respect to control alleles (see Webster et al, 20014); (+) the data support a greater than or equal to 100:1 ratio in disease alleles versus control alleles; (-) the data do not support this ratio. M# = The number of correctly segregating affected meioses (see text). Blosum = the score for the amino acid change on the blosum 62 substitution matrix (see text). EPP = estimate of pathogenic probability (see text).

APPENDIX M: GLOSSARY

Mendelian trait (or disease): One that is caused by variation in a single gene.

Complex trait (or disease): One that has a nonmendelian genetic component; for example:
Polygenic trait (or disease): One that is caused by the additive effect of variation in more than one gene.
Multifactorial trait (or disease): One that is caused by the additive effects of genetic and nongenetic (eg, environmental, developmental) factors.

Genetic heterogeneity: The situation in which a single clinical phenotype (eg, Leber congenital amaurosis) is caused by variations in different genes in different individuals.

Genetic background: A term that is used to refer to all the genes in the genome except the one (or the few) under study; this term is used to imply the existence of a polygenic mechanism for expression of a trait.

Penetrance: The fraction of individuals with a certain genotype who manifest a certain trait at a specified age; the existence of "incomplete penetrance" implies a polygenic or multifactorial mechanism for expression of a trait.

Expressivity: The extent or quality of an allele's effect on the phenotype of an individual who carries it; the existence of "variable expressivity" implies a polygenic or multifactorial mechanism for expression of a trait.

Disease-causing gene: This term is usually used to refer to a normal gene that has suffered a mutation that alters the gene's function to the degree that a clinically detectable abnormal phenotype results.

Sequence variant, sequence change, or change: One or more contiguous nucleotides that differ from the most common sequence in the population.

Non-disease-causing variant (NDCV): A variant that never, under any circumstances, alters the phenotype of an individual in a way that would be judged to be clinically abnormal.

Disease-associated variant (DAV): A variant that does not itself alter an individual's phenotype in any way but which is so tightly linked to a DCV that its presence can be reliably predictive of disease.

High penetrance disease-causing variants (HPDCV): Variants that would be expected to alter the phenotype of an individual sufficiently that a clinician can detect an abnormality in a very high proportion (eg, 90%) of individuals who carry them in the appropriate configuration. The expression of an HPDCV is rarely significantly affected by the presence of other genetic or environmental factors, and thus the frequency of such variants (as a group) can be predicted from the prevalence of the disease and the Hardy Weinberg equation.

Low penetrance disease-causing variants (LPDCV): Variants that are capable of altering the phenotype of an individual sufficiently that a clinician can detect an abnormality but only do so in a rather low proportion (eg, <50%) of individuals who carry them; the expression of an LPDCV requires the presence of an additive genetic or environmental factor, and thus their frequencies are higher than would be predicted by the Hardy Weinberg equation.

Threshold-modifying variants: Variants that lie outside a disease-causing gene (perhaps even on a different chromosome) and whose presence increases or decreases the likelihood that an altered phenotype will result from the presence of a DCV within the disease-causing gene.

Hardy Weinberg equation: $p^2 + 2pq + q^2 = 1$. A relationship that can be used for predicting allele frequencies from disease prevalence in stable populations.