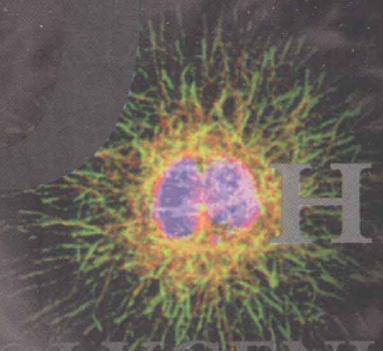
Proceedings of International Symposium for Mapping and Identification of Genes for Complex Traits

GENETICS FOR COMPLEX POLYGENIC DISEASES AND HUMAN HEALTH



Edited By

Deng Hongwen

Wu Xiushan

Yin Dazhong

Hunan Normal University Press, Changsha, Hunan, China

Proceedings of International Symposium for
Mapping and Identification of Genes for Complex Traits

GENETICS FOR COMPLEX POLYGENIC DISEASES AND HUMAN HEALTH

复杂多基因疾病的遗传学与人类健康

Edited By
Deng Hongwen
Wu Xiushan
Yin Dazhong

Hunan Normal University Press, Changsha, Hunan, China

图书在版编目(CIP)数据

复杂多基因疾病的遗传学与人类健康 / 邓红文,吴秀山,印大中著 .—长沙:湖南师范大学出版社,2002.12
ISBN7—81081—229—7/Q·002

Ⅰ.多 … Ⅱ.①邓 …②吴 …③印 … Ⅲ.多基因—人类遗传学—研究 Ⅳ.Q987

中国版本图书馆 CIP 数据核字(2002)第 094128 号

复杂多基因疾病的遗传学与人类健康

◇邓红文 吴秀山 印大中 著

◇责任编辑:李 阳 李东屏
◇责任校对:陈冠初
◇出版发行:湖南师范大学出版社
地址/长沙市岳麓山 邮编/410081
电话/0731.8853867 8872751 传真/0731.8872636
网址/www.hunnu.edu.cn
◇经销:湖南省新华书店
◇印刷:国防科技大学印刷厂印刷

◇开本:787×1092 1/16
◇印张:17.25
◇字数:464 千字
◇版次:2003 年 2 月第 1 版 2003 年 2 月第 1 次印刷
◇印数:1—1000
◇书号:ISBN7—81081—229—7/Q·002
◇定价:40.00 元

Preface

With the rapid progress of the human genome research, mapping and identifying of genes contributing to quantitative traits and complex diseases and their function studies are becoming increasingly important. Much success has been achieved in the identification of genes and their functional studies for Mendelian-inherited traits. However, most of biologically important traits vital for human welfare are complex polygenic traits, determined by both multiple genetic and multiple environmental factors. These traits include those that are important in the medical area (including most common diseases such as coronary heart disease, osteoporosis and diabetes etc), in agriculture, forestry and animal husbandry (crop and milk production etc). One of the greatest challenges in modern biological research is to map and identify these genes in genome of important organisms. The research involves multiple disciplines, such as Molecular Biology, Genetics, Biostatistics, Bioinformatics, Medicine, Forestry and Agronomy etc. Some encouraging progress has been achieved and necessary technologies and methodologies are being under constant development in gene identification for polygenic traits and complex diseases. The research is still at relatively elementary stage with a relatively small scope in China.

To promote international exchange and collaboration, particularly those between scientists in China and those in other countries, with the support from the National Science Foundation of China, Hunan Normal University and private funding from Dr. Gao Shouquan of Hunan Research Institute of Chinese Medicine, an International symposium of mapping and identification of genes for complex polygenic traits and diseases was held during May 23-26, 2001 in Changsha, China. Nine active scholars from USA and twelve nationally known Chinese scholars presented their work at the symposium, together with dozens of other selected oral and poster presentations from domestic researchers of China. The meeting was viewed by the nearly 100 meeting participants as highly beneficial in promoting and enhancing research in mapping and identification of genes for complex polygenic traits in China and successful in fostering international exchange and collaboration in this research field.

The topics of the meeting presentations covered some of the latest research progress and issues on mapping and identification of genes for polygenic complex traits and diseases; the basic experimental designs, methods, technologies, and analysis strategies in this field; relevance of progresses in Human Genome Project to the research of complex polygenic traits and diseases; and researches on biological high technologies and products and mapping and identifying of genes for polygenic trait and complex diseases. A number of the papers were submitted for publication to the meeting organizers. These papers cover a range of topics that some mirror a few presentations at the meeting and some reflect a number of aspects of our current status in this research field in China and abroad. We thought it might be useful to compile these papers in a book to share it with colleagues who could not attend the meeting. This book may serve, in some aspects, as a comparison to measure our future research progress in this research field in China. It is our sincere hope that, after not a long period of time, when we look back to the research summarized in the papers in this book, we will be able to say proudly that we indeed have been progressing fast and steady in this field in China.

We acknowledge many individuals who have aided in the whole process. We are especially grateful to the most notable patience and insights from the members of Laboratory of Molecular and Statistical Genetics in Hunan Normal University such as Liu Manyuan, Chen Xiangding, Liu Xianghua, Lei Shufeng, Li Yumei, Zhou Xiaogang, Cao Chike, Deng Feiyan, Jiang Deke, Li Miaoxin, Jian Weixia, and to many teachers and students in the other laboratories of Life Science College in Hunan Normal University who are helpful, we thank their cooperation and friendship. Thanks to our publisher, especially to the Mr. Li Yang, he did a great job.

Deng Hongwen, Ph.D.
Cheung Kong Scholar of the Ministry of Education of China
Specially Appointed Professor of Biological Sciences of Hunan Normal University, China
Associate Professor of Medicine and Associate Professor of Biomedical Sciences
Creighton University, USA

Contents

PART ONE

Invited Papers

Study Design for Genetic Epidemiology and Gene Mapping: The Korean Diaspora Project

Joseph D. Terwilliger[1,2,3], *Harald H.H. Göring*[4], *Patrik K.E. Magnusson*[1], *Joseph H. Lee*[5,6]

1. Columbia Genome Center, Columbia University, New York, NY10032, USA
2. Department of Psychiatry, Columbia University, New York, NY10032, USA
3. Division of Medical Genetics, New York State Psychiatric Institute, New York, NY10032, USA
4. Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX, USA
5. Sergievsky Center and Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY10032, USA
6. Department of Epidemiology, Columbia University, New York, NY10032, USA

Abstract: Study designs in genetic epidemiology and gene mapping have traditionally focused on single (often simple) data structures with oversimplified epidemiological models. However, when the true state of nature is even slightly more complex, these simple methods will fail almost surely. To this end, it may be important to restructure our thinking in terms of what sorts of etiological models would be consistent with what we know about evolution, secular trends in phenotypic expression, and the ephemeral nature of the genotype-phenotype relationships, among other things. Segregation analysis attempts to make unbiased estimates of the penetrance functions, or P(Phenotype | Genotype) while gene mapping strategies make full use of ascertainment bias to maximize the predictive value, P(Genotype | Phenotype; Ascertainment), such that neither can be predicted meaningfully from the other in general. Popular misconceptions about the "world-changing effects" of the genome project have lead even serious scientists to ignore inconvenient scientific facts in an effort to justify expenditures of large sums of money on mapping strategies that are critically dependent on difficult-to-justify assumptions. Combining data from different study designs will certainly increase the number of degrees of freedom available for estimation of parameters and testing of hypotheses in a given dataset. Naturally the size of the dataset also needs to be large, but such approaches are potentially much more informative about the true state of nature. Each traditional study design focuses on one specific aspect of etiological variance. Twin studies focus on relative effects of genes conditional on shared environmental factors within a household. Adoption studies focus on inter-household variation in environmental exposure while partially controlling for genetics. Migrant studies focus on intercultural variation in environmental exposures while controlling for population level genetic differences, and family studies focus on the complex sets of genetic correlations that can be measured within larger kindreds. The components of variance emphasized by each study are radically different, and it is clear that a study combining elements of each would certainly be more informative about the relative contributions of nature, nurture, and interactions between them on trait

etiology. Similarly, this information can be used to assist in mapping strategies, and genotype-phenotype correlation studies. We are currently developing such a study using populations of the Korean diaspora, composed of families of Koreans in Kazakstan, Korea, China (延边) and the USA, as well as internationally adopted Koreans in the USA and Europe, hopefully with data from their biological relatives in Korea as well.

Key words: genetic epidemiology; Korean; multifactorial disease; quantitative traits; linkage; genetic disease

There has been a dramatic increase in interest in human genetic research in recent years, spawned by the technological progress in sequencing the human genome, and the increasing accessibility and affordability of the resources required for large scale genotyping and sequencing efforts. However, before applying these tools in practice, it is important to clarify the scientific questions, and to determine whether technology is enough to solve the specific problems we are interested in[1]. If we can genotype a sufficiently dense set of markers in a large number of families, we can map the chromosomal location of a polymorphism with known genotypes with classical linkage analysis. If we wish to perform mapping studies based on linkage disequilibrium analysis, again to identify the chromosomal location of a polymorphism with known genotypes, then it is fair to ask how many markers would be needed to have high power and precise estimates of map position[2]. The forces governing the quality and quantity of linkage disequilibrium are random and chaotic, but have a certain regularity to them, which can allow mapping of a polymorphism with unknown genomic location if enough individuals are studied with enough markers[3]. Of course, how much is "enough" is a fair question to ask, but this question can be addressed within the realm of technology, especially when the day comes when we will have technology available to allow us to sequence the whole genome in any individual we desire efficiently, quickly, and cheaply[4].

If one wishes to correlate genetic variation to some observed phenotype, however, the power of the study is more directly related to how well the underlying pattern of genetic variation can be predicted from the observed phenotypes, conditional on the ascertainment scheme employed. If the phenotype predicts the genotype of some locus with high probability, the genomic position of that locus can be mapped as described above (e.g.[5]). However, if there is a weak predictive value of phenotype on the underlying risk genotype, mapping is unlikely to be successful irrespective of the technological capabilities (e.g.[6]). You cannot find a genetic risk factor that is not predicted by the measured phenotypes in your dataset[7]!

The power to map a variant with known genotypes, and the power to predict genotypes of some risk locus from observed phenotypes are orthogonal, as the power to map genes which influence some phenotypic outcome, P(successful mapping | phenotypes known) = P(successful mapping | disease locus genotypes known)P(disease locus genotypes known | phenotypes known). While numerous investigators have looked into properties of the former relationship (how to map a SNP)[1,3,8~15], very few have candidly dealt with the latter (e.g.[1~4],[7]). Unfortunately, the phenotype→genotype transformation probabilities dominate the power of our mapping studies of complex traits, and these probabilities cannot be influenced by advancement in molecular genetic technology (see[1~4],[7],[16~18], for details).

Most published investigations of the power of complex disease gene mapping studies are based on grossly oversimplified models for the phenotype→genotype relationship, which rarely model the sorts of complexity that we know exist in nature (see[19],[20]). We propose that careful attention to study design, and appropriate

choice of phenotypes to be studied, are the most promising means to improve the overall power, because the predictive value of phenotypes on the underlying genotypes is a direct function of the ascertainment scheme employed[3,16]. To this end, we propose that careful consideration be given to the myriad of factors which we know can influence the phenotype→genotype relationships. Given that most of the problems plaguing complex-trait gene mapping are more biological than technological in nature, it is fair to ask the question "what is the power of my study to identify the genetic factors underlying the phenotype if the complete sequence of every individual in my study were available?"[4] The purpose of such a question is to focus the discussion on the underlying biological complexity, rather than on technological issues. While technological problems can and will be solved in time, we will not be able to escape biological realities, and our studies, therefore, must acknowledge their existence instead of just wishing them away...

1. Architecture of complex traits

Let us start with a model of the etiology of a multifactorial trait, which includes a number of types of risk factors, as outlined in figure 1. Each type of risk factor shown has a different effect on genetic epidemiology studies, leading to different sorts of conclusions. The conventional genetic epidemiology study designs are set up in such a way that they make very strict sets of assumptions about the nature of these factors-typically ignoring the existence of several of them-leading to an overestimation of the contribution of other factors[1]. In practice, these studies are generally designed to overestimate the effects of major genes, because this is what genetic epidemiologists are trying to detect[3].

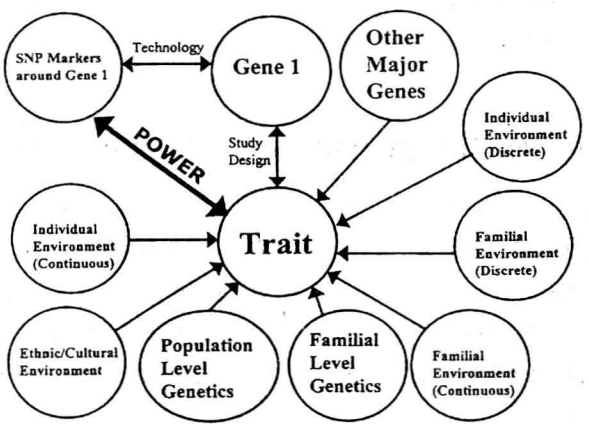


Figure 1 Etiological architecture of a multifactorial trait

For genetic factors, it is important to distinguish between major genes (meaning individual loci whose genotypes have significant individual effect on the trait distribution), and polygenic effects (meaning the cumulative effects of multiple genetic factors of individually minor effect, which in combination can have large influence on the trait distribution). Furthermore, it is important to discriminate between factors which lead to detectable phenotypic correlation among family members, and those which lead only to phenotypic correlations among individuals within a population as compared to individuals from other populations. Within families, there are continuous and discrete types of correlations. Discrete correlations may be due to major genes, such that relatives have an increased risk of sharing genetic factors which have a large impact on the trait,

potentially leading to clusters of relatives at one end of the phenotypic distribution, while other relatives-who do not carry risk genotypes at these loci-may have trait values in another portion of the trait distribution. Continuous correlations could be due to polygenes, where relatives share some proportion of the alleles that impact the trait in a minor way, leading in aggregate to quantitative correlations in phenotype, without discrete clustering.

Environmental correlations can be broken down into a similarly large number of factors, either discrete or continuous in nature. For example, individual discrete factors, like whether or not someone smokes can have an enormous individual effect on traits distributions, while continuous factors (like how much an individual smokes) can have more quantitative relationships to risk. In practice, the effects of discrete environmental factors mimic the discrete effects of major genes, while the continuous environmental factors mimic the continuous effects of polygenic background (continuous familial or population-level genetic effects). Shared environment in families, as well, can comprise both discrete and continuous risk. Relatives of smokers are more likely to smoke, though not all relatives do, and this effect can mimic a major gene effect in many of the conventional study designs for genetic epidemiology. Likewise there is a quantitative correlation among siblings in terms of how much they exercise or how much fat they eat, and this can be confounded with the polygenic background effect. On the population level there are differential effects of both discrete and continuous variation. Variation in discrete factors can lead to differences in cancer incidence around sites of nuclear disasters, while continuous variation like population distributions of fat intake can affect population distributions of serum cholesterol or triglycerides, though these are not subdivided in the drawing for lack of space. Of course genetic factors can influence the choice of environment, and vice versa, and all the factors shown in the drawing are correlated, with their effects on the trait almost never actually being independent (e.g.[5]).

2. Study designs for genetic epidemiology: Assumptions and limitations

The study designs most often employed in genetic epidemiology are varied, with the most common study designs being twin studies, adoption studies, migrant studies, and family studies. Each of these study designs makes some implicit assumptions about some of the types of factors outlined in Figure 1. These assumptions are critical to evaluating what one observes, especially when different study designs lead to (apparently) different conclusions. In order to understand the limitations of the conventional approaches, let us examine them in more detail (see also Table 1).

In twin studies, the phenotypic correlation among sets of monozygotic (MZ) twins is compared to that among dizygotic (DZ) twins. Because MZ twins are genetically identical (aside from somatic mutations, some methylation patterns, genetic rearrangements in the immune system, and other modifications), and DZ twins share only 50% of their autosomal genetic material on average, the more genetic a phenotype is, the greater the MZ twin correlation is expected to be, when compared to the DZ twin correlation and the population prevalence. The implicit assumptions of this type of analysis are that members of MZ and DZ twin pairs have identically distributed environmental risk factors, and that these factors are equally correlated within twin-pairs independent of zygosity. These assumptions are not completely realistic, however, as it is not unlikely that there are very different correlations, conditional on zygosity, in many environmental factors (ranging from in utero environment, to later lifestyle factors) among twin pairs, the effect of which is that the contributions of

genes may tend to be overestimated by design. Furthermore, for some traits, being a twin itself may be a risk factor, due to both pre-and post-natal environmental reasons, further complicating the inferential value of twins without considering other data structures as well. Since the only parameters that can be independently estimated are prevalence and twin correlation among MZ and DZ twins respectively, it is very difficult to discriminate among genetic models. The net effect of the assumptions in twin studies, on average, is likely to lead to some overestimation of the contribution of genetics to the trait.

In adoption studies, genetic and environmental contributions can be dissected by examining different sociological and biological family relationships. The phenotypic correlation between biological siblings who were raised in different families can be compared to the phenotypic correlation between biological siblings who were raised in the same family, and also to phenotypic correlations between non-biological siblings raised in the same family. The following factors would thus be estimable: distribution of the trait in adoptees and non-adoptees (to evaluate the effects of the discrete environmental risk factors of being adopted and having an adopted sibling); trait correlation among biological sibs reared together (due to shared genetic and environmental factors); trait correlation among biological sibs reared apart (due to shared genetic factors and shared in-utero environmental factors—corrected for the effects of 'being an adoptee'); trait correlation among non-biological sibs reared together (due to shared familial environmental factors—corrected for the effects of 'being an adoptee'). In these studies, the effects of in-utero environment and genetics are again confounded, and the effects of 'being an adoptee' and 'having an adoptive sibling' are somewhat confounded with the effects of familial environment in the adoptive families, complicating aspects of the analysis. It is also generally assumed that there is no effect of the type of families who choose to adopt children on the familial environment or attempts by adoption agencies to match children with families similar to their biological families in some characteristics. Further, it is generally ignored that familial correlations in environment may vary conditional on adoption status, though the latter is measurable by looking also at the correlations among pairs of adoptive siblings who are non-biologically related, but both adopted by the same family. In general, the net effect of most of the assumptions made in adoption studies is likely to overestimate the contribution of genetic factors to the trait, just as in twin studies.

In migrant studies, one studies population level differences in the distribution of a trait among persons from the same 'genetically defined' population living in different cultural and environmental conditions. In such studies, it is assumed that the migrant populations are representing the same gene pool, while environmentally and culturally different, and thus these studies are a population-level analog of the adoption studies described above. Such studies are also the human analog of the type of experimental study often done in plants (randomized block design) where an identical set of plant strains are raised in different fields under different environmental conditions to test for interactions between genetic and environmental factors. The main difference is that in the plant studies the organisms in a given strain are genetically identical, while in human studies, there is significant inter-individual genetic (and environmental) variation within the population, complicating the interpretation. In such studies, one conventionally studies two "genetically-defined" populations (A and B), comparing the phenotypic distribution in population A in its homeland with that of migrants from population A living in population B's homeland, with that of population B in its own homeland. In some cases where there is reciprocal bi-directional migration it is also possible to study population B in population A's homeland, but this is not very common given the reasons underlying most human migrations in

recent years. In these studies, there is again comparison of the same ethnic population in two environments, to measure the contribution of environment to the trait distribution, while the difference between populations A and B in the same environment provides a measure of the genetic contribution to phenotypic variation. The effect of self-selection may influence the estimates of genetic vs. environmental contribution where those who choose to migrate may differ somewhat from those who choose to stay at different levels, though at least on a genetic level, it is possible to test for genetic heterogeneity (which is the only bias of relevance to such analyses) by taking advantage of the recent advances in biotechnology. Of course this all depends on how similar or how different the populations are in both a genetic and an environmental sense-a similar objection to that raised above for adoption studies.

In family studies, one generally assumes that the environmental contribution to the trait has a common distribution for all individuals, also in the best analyses, a household effect may be included, that is assumed to be identically distributed among all members of a given nuclear (or sometimes extended) family. The residual correlation in the phenotype among relatives is presumed to be due solely to genetic factors, which are shared among relatives according to Mendelian segregation laws. The simplest family study would compare a phenotype in sibling pairs reared together, one component of the adoption design described above. In general, sibling correlations in general should be no greater than the correlation among DZ twins who tend to be more correlated in environmental exposures and are age-matched. Furthermore, they should also be no less correlated phenotypically than non-biological sibs reared together (who share only familial environment but not genetics), or biological sibs reared apart (who share genetics but not post-natal environment). Clearly they must be less correlated than MZ-twins reared together as well, as MZ twins share 100% of their genes, while sibs share only 50%. One can characterize additional restrictions on the correlation structures among different sets of relatives as well, as a function of the data from the other study designs described, but this information is almost never utilized in practice. One concern about family studies is that it is difficult to provide a good model for the environmental exposure correlations among relatives, and because the effects of familially correlated discrete environmental risk factors tend to mimic the effects of major genes, the interpretation of such segregation analysis can be extremely complicated. Discrete environmental risk factors, like smoking, owning an expensive sports car, going to medical school, and viral infection, among others, are often shared among relatives, and can mimic the effects of major genes. Assuming such factors do not exist would lead to a systematic overestimation of underlying major gene effects. This is because discrete risk factors correlated among relatives are inferred to be genetic in classical segregation analyses, as this is really the only evidence used to discriminate shared environment and polygenic background effects from major gene effects.

Given the issues discussed above, it is not surprising to find inconsistent conclusions from different types of study looking at the same trait within or between populations. When the same trait is studied using different study designs, the results very often appear inconsistent, largely because the assumptions underlying each design differ significantly (see Table 1). Of course, it is possible to combine multiple study designs in a single analysis, to better separate the effects of genetic and environmental factors and their interactions, e.g. by collecting families from a single genetic population living in different environments, in order to look at the effects of cultural environment on the familial correlations in phenotype. Adoption information can be analyzed simultaneously as well, if the ascertainment conditions are the same (e.g. random ascertainment). Joint analyses of heterogeneous data structures admit more degrees of freedom to the analysis, thereby attenuating the

Table 1 Overview of estimable etiological components from various study designs

Data Structure	Familial Genetics	Familial Environment	Individual Environment	Population Genetics	Cultural Environment	Major Genes	In-utero environment	Gene Mapping
MZ Twins	Same	"Same" + MZ Twins	Twin effect	Same	Same	Same	1 or 2 placentas	LD
DZ Twins	50% Shared	"Same" + MZ Twins	Twin effect	Same	Same	50% Shared	2 placentas	Linkage and LD
Sibs reared together	50% Shared	Same		Same	Same	50% Shared		Linkage and LD
Sibs reared apart	50% Shared	Different	Adoption effect	Same	Same, except crosscultural adoption	50% Shared		Linkage and LD
Non-sibs reared together	Independent	Same	Adoption effect	Same, except "transracial" adoption	Same	Independent	Independent	LD
Half-sibs	25% Shared	"Same" if raised in same household	Half-sib effect	Same, except if nonoshared parent from different population	Same if raised in household from same culture	25% Shared	Independent if father shared and mother not shared	Linkage and LD
Migrant Studies			"Self-selection for migration" effect possible	"Same" within genetic population, different between	"Same" within cultural population, different between			LD
Family Studies	Correlated by Mendelian laws	Correlated in less quantifiable patterns		Same	Same	Correlated by Mendelian laws		Linkage and LD
Migrant-family-adoption study	Varies	Varies	Varies	Varies	Varies	Varies	Varies	Linkage and LD
Inbred animal backcross	Same	Same	Same	Same	Same	75% shared	Same	Linkage equals LD

need for so many assumptions about the correlation structures of certain etiological parameters (*cf.*[18]). If twins are available from the same population, adding them to the analysis allows one to study additional

parameters, and if families from genetically quite different populations are available as well, one can address issues of population-level genetic effects that may be invisible to family-based analyses. The more heterogeneous the sample is, the greater the number of parameters that can be estimated, providing more information about the interrelationships among the factors. It is important to note, however, that the sample size requirements for each individual data structure remain essentially as large as in the individual study designs, because the effects that we are trying to estimate are typically quite small. As in all statistical approaches, there is a tradeoff between the number of degrees of freedom available for parameter estimation, and the sample sizes one needs to ascertain. The cost of not ascertaining heterogeneous samples, however, is that the resulting parameter estimates are tightly connected to the assumptions underlying the study design selection. This is generally undesirable, and tends to overestimate the genetic contributions, by design.

3. Study designs for gene mapping: Assumptions and limitations

In gene mapping, there have been a number of changes over time in the philosophical approaches to study design. Initially most gene mapping studies were based on collecting large families of individuals, many of whom shared some disease phenotype. The underlying assumption was that these individuals shared the same phenotype because they shared some underlying genotypic risk factors. Within a given family, they were reasonably assumed to share the risk alleles from some common source, and thus the gene involved could easily be localized to some chromosomal region by linkage analysis. Linkage analysis is based on the observation that loci, which are close together on the same chromosome, will tend to be co-inherited more often than two random loci in the genome. Thus individuals presumed to share a genotype at some putative disease locus would also share some genotype at nearby marker loci. If we have a genome-spanning map of marker loci, then it would be possible to find marker loci whose genotypes are shared by the affected individuals within a pedigree, due to linkage, and thus assign the disease gene to some chromosomal region. This approach worked rather successfully for unraveling simple Mendelian traits, for which the disease phenotype was a very predictive surrogate for the underlying disease genotype (see the arrow marked "Study design" in figure 1, which represents the correlations between phenotype and genotype of the locus to be mapped), which were correlated to the marker locus genotypes due to linkage (see the arrow marked "Technology" in figure 1, which represents correlations among genotypes of trait-and marker-loci). The ascertainment of large families increases, in general, the predictive value of phenotype for the underlying genotypes of the loci to be identified.

When the etiology is more complex, phenotypes are less reliable predictors of the underlying genotypes of any locus to be mapped, and this leads to greater difficulty in mapping them. For example, consider retinitis pigmentosa (RP)[4], for which pathogenic genotypes of any one of more than ten independent loci could each cause the same disease. In this case, ascertainment of a random person with the disease is a poor predictor of the genotype of any given locus, since he could be affected due to risk alleles of any of the more than ten disease loci. However, ascertainment of a single large family with multiple individuals sharing this rare disease increases the chance that the individuals are all affected because of genotypes of the same risk locus, and thus the loci whose variant alleles can cause RP were mapped successfully through analyses of large extended pedigrees. However, reliance on small families, such as affected sib-pairs, would have required hundreds or thousands of families for identification of any of these genes, because each sib-pair could be affected with RP