Environmental Medicine, Vol. 46 No.1, 2, 2002 (page 1 ~ 6) **Review Article**

Search for Complex Disease Genes: Achievements and Failures

Tatiana I. AXENOVICH[1,2] and Pavel M. BORODIN[1,2,3]
[1] Institute of Cytology and Genetics, Novosibirsk 6330090, Russia
[2] University of Novosibirsk, Novosibirsk, 630090 Russia
[3] Space Medicine Research Center, Research Institute of Environmental Medicine
Nagoya University, Nagoya 464-8601, Japan

Abstract: In this paper we review current methods of mapping complex disease genes. We outline the rationale of these methods, fields of their application, their strong and weak points. The progress in genetic mapping of human diseases has been mainly concerned with the Mendelian diseases. Achievements in identification and mapping of complex disease genes have been less impressive. A substantial progress has been reached in development of statistical methods of complex disease mapping. However, most of them are still based on the implicit assumption that there is a gene with a significant effect. They are effective in cases when a disease is due to mutations in structural genes or in their genetic neighborhood. Statistical methods are capable of isolating the effect of the major gene from the effects of genetic background and environmental noise, and localizing it. However, these methods fail when several genes of equal effect are involved. What hinders the progress in mapping of complex disease gene is not the drawbacks of the methods themselves, but underdevelopment of the general concept of genetic anatomy of complex traits.

Key words: complex diseases, linkage analysis, QTL mapping, allelic association, statistical genetics

Introduction

The first step to identification of complex disease genes is their genetic mapping. The main principles of genetic mapping have been established at the beginning of the twentieth century.[1)] However, it took a long time to make them work on the full scale. It was necessary to accumulate and process a large body of information about the markers and wait for emergence of powerful computer technology that made possible statistical analysis of recombination. An impressive progress in mapping of the genes responsible for human inherited diseases has been reached due to advances in molecular genetics and computer science.[2)]

There are two main classes of inherited diseases: rare Mendelian diseases and common complex diseases. The progress in genetic mapping has been mainly concerned with the former class of diseases, which are determined by mutations of structural genes.[3)] The latter class includes such widespread diseases as diabetes, hypertension, asthma, some forms of cancer and psychiatric disorders. These diseases are controlled by many interacting genetic and environmental factors. Achievements in identification and mapping of these genetic factors have been less impressive.

In this paper we shall discuss the situation in the mapping of genes controlling complex diseases and outline the main problems that hinder the progress in this field.

We shall consider the following questions:

What is the main difference in etiology of Mendelian and complex diseases?

What principles the mapping of complex trait genes is based upon and what peculiarities of complex human diseases make it so difficult to identify and map the genes responsible for them?

What solutions of this problem have been suggested and how effective they have been?

Genetic Etiology of Complex Diseases

Fig 1. summarizes modern views on the etiology of complex diseases. The phenotype of a patient is determined by various genetic, environmental and social factors interacting with each other in many ways.[4)] The genetic factors involve many loci of variable effects. It is possible to isolate those of them that make most pronounced contribution to phenotype and put all other genes that play smaller roles into the category of genetic background. The division of genes according to their contribution in the phenotype into the genes of significant effect (major- or oligogenes) and genes of minor effect (polygenes) has been put forward at the dawn of genetics at the beginning of the twentieth century.[5)] It was motivated by technical reasons. Two different approaches were used to analyze major genes and polygenes.

The major genes were treated as Mendelian factors. It was possible to analyze the segregation of their alleles and estimate their contribution to the phenotype. This approach has been developed mainly in human and medical genetics. It is well known now as segregation analysis. This approach plays

Reprint requests to: Pavel M. BORODIN, Institute of Cytology and Genetics, Novosibirsk 630090 Russia.

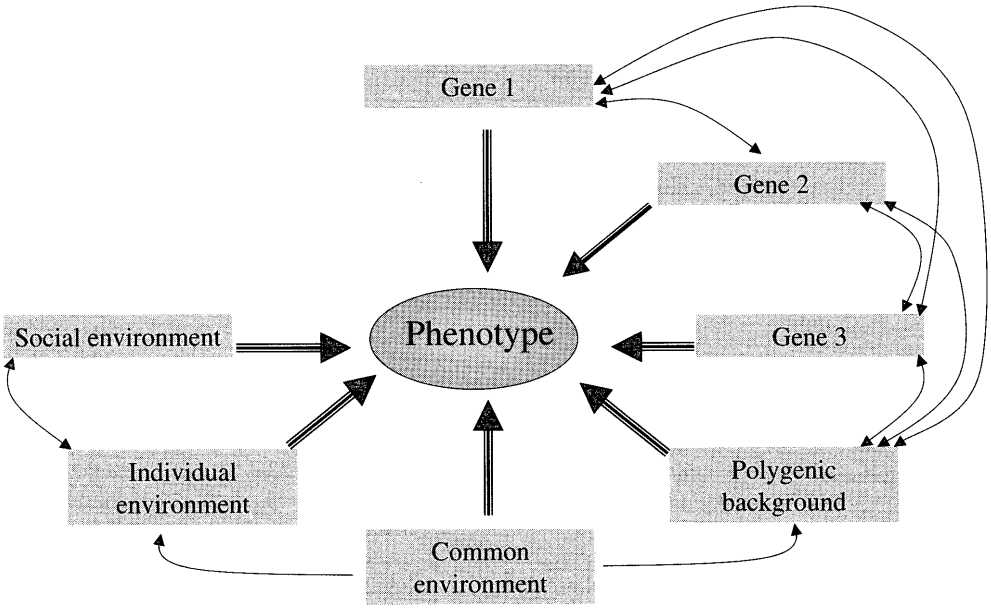

Fig. 1 Phenotype as a result of interaction of genetic and environmental factors.

an important role in the identification of Mendelian disease genes. Quantitative traits were analyzed within the framework of polygenic models. This approach ignored segregation of separate genes and operated with average genetic effects. These effects were estimated as the ratio of genetic variance to total phenotypic variance (heritability). Polygene model was mainly used in agricultural genetics.

The synthesis of these two approaches started at last quarter of the twentieth century with the models where major gene and polygene components were considered as independent of each other. These models analyzed the major gene segregation on the polygene background.[6] Further development resulted in regressive-logistic model that was taking into account the interaction between the major gene, polygene, and environmental components.[7] Thus, at least theoretically within the framework of these models we are able now to distinguish the effects of separate genes from the effects of genetic background and environmental noise.

The modern knowledge of the structure and function of human genome gives us an alternative way of categorizing the genetic factors.[8] It concerns the distinction between structural and regulatory genes, rather than relative effects of the genes (Fig. 2). The structural genes control the protein structures. Mutations of these genes usually have a significant phenotypic effect. With some caution we may draw a parallel between the structural genes and major genes. The regulatory genes can be further subdivided into *cis*-controlling elements (like enhancers) and *trans*-controlling regulators (like transcription factor genes). Mutations at the controlling elements may affect the expression of respective structural genes: the quantity of the gene product, the pattern of its splicing, its tissue and stage specificity. This is what we usually observe in case of complex diseases. They affect quantitatively rather than qualitatively a wide set of physiological, biochemical and other biological functions of an organism at specific cells and tissues and specific developmental stages. Other important features of the regulatory genes are their modular organization and functional flexibility. Usually, the structural genes have a series of *cis*-regulatory elements located at various distance from them. These elements have a greater opportunity to accumulate mutations, than the structural genes do. They do not need to maintain a reading frame. Their functions do not depend critically on their position and orientation. Therefore effects of mutations at these elements are not so strong. For this reason they serve as a rich source of phenotypic variability.[8] In this context, the regulatory genes seem to be attractive candidates to the role of polygenes.

Can we distinguish the effects of these genes and map them using the modern methods of statistical genetics?

Principles the Mapping of Complex Diseases

There are two main approaches to an identification of complex disease genes.[9]

The first approach involves a search for candidate genes. It is based on some assumptions about the possible mechanisms of the pathology. For example, a simple reasoning suggests that the genes affecting sugar metabolism can be assumed to be responsible for diabetes. Alternatively, the search for candidate genes can be based on data of comparative genomics. For example, several genes whose mutations or/and knockouts cause diabetes in the laboratory mouse have been iso-

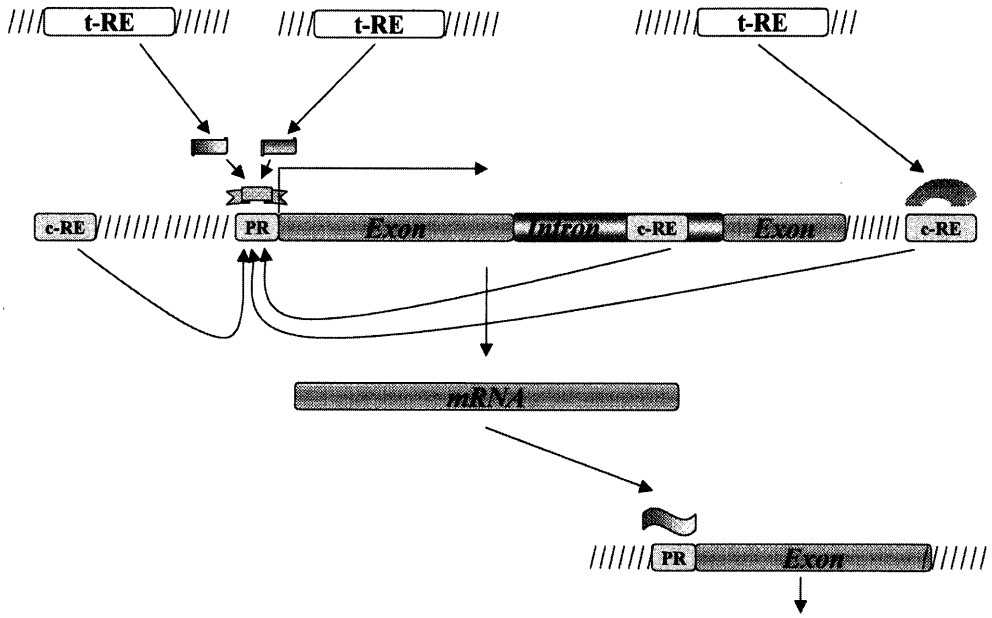

Fig. 2 A simplified scheme of organization of genetic network.
Products of *trans*-acting regulatory elements (t-RE) interact with promoters (PR) of structural genes and their *cis*-acting regulatory elements (c-RE) and modulate their transcription and splicing. The product of the structural genes may play a role in regulation of other structural genes. Altered phenotype may result from mutations in any elements involved in the genetic network.

lated.[10] Human homologues of these genes appear to be likely candidates.

The second approach is positional approach. It does not demand either knowledge of the primary cause of the disease or assumptions about candidate genes. This approach is aimed at detecting co-segregation of the disease phenotype and alleles of marker gene, whose genomic location is already known. Once the genomic position of the gene responsible for a disease is established, we may isolate it and analyze its sequence, its product and its function.

Both of these approaches start with genetic mapping. The positional approach implies an analysis of genetic linkage between a marker and the gene determining the pathology. If they are linked we observe a deviation from independent segregation of the alleles of these two loci. However, in case of complex diseases we may only assess the disease genotypes indirectly via analysis of distribution of affected and non-affected persons in a group of genetically related individuals.

When we test candidate genes we presume that the gene controlling a disease and the candidate gene are the same or, in terms of mapping, they are located at the same locus or very near to each other. If we choose the right gene, we should observe a close correlation between the phenotypes and the candidate genotypes.

The candidate genes testing as well as positional mapping have been primarily suggested for Mendelian gene mapping. A set of special statistical methods have been developed.[11] However, in spite of similarity in approaches to mapping of Mendelian and complex traits, the methods developed for statistical analysis of Mendelian traits cannot be directly applied to quantitative trait loci (QTL) mapping due to genotypic uncertainties of complex phenotypes. In case of complex diseases there is no direct and clear correspondence between genotypes and phenotypes. We can only estimate this correspondence statistically. For many years the statistical problems hampered the progress in QTL mapping. Recently many of these problems have been solved, and medical genetics acquired a rich set of statistical methods for QTL mapping. There are four groups of them: recombination analysis, analysis of components of variation, analysis of alleles identical by descent and analysis of associations.

Current Statistical Methods for QTL Mapping

Recombination analysis

Recombination analysis (or model based linkage analysis) is the most ancient method. It has been developed at the dawn of science genetics.[1] In its initial form it presumed that the parental genotypes for loci controlling the analyzed trait and marker loci and their linkage phase were known and recombinant and non-recombinant offspring were visually distinguishable. In this case we are able to estimate the recombination fraction directly. A significant difference of this estimate from 0.5 is interpreted as a proof of linkage between the marker and the gene controlling the analyzed trait.

The same idea is used in QTL mapping. However, in this case we cannot see the genotypes and therefore need to infer them using the principle of maximal likelihood. The likeli-

hood of empirical data is represented as a function of recombination fraction and the model of inheritance.[12] This is why recombination analysis is sometimes called model-based linkage analysis. The linkage test is performed using different variants of likelihood ratio test. The most widely used is *lod score* criterion.[13] To utilize this approach we need to give the most full and correct description of the model of inheritance of the trait. The problem is that this is rarely possible. However, this problem can be solved if we use model-free mapping methods.

Model-free mapping methods

In its classical form this method is based on the analysis of descent of alleles of the marker gene in a pair of affected sibs.[1] A term "alleles identical by descent (IBD)" has been introduced to describe the alleles present in the sibs that were the copies of a certain parental allele. Each pair of sibs may have 0, 1 or 2 IBD. If the marker is not linked to the gene that we are analyzing, then the distribution of the number of IBD in a pair of affected sibs should be 0.25:0.5:0.25. A significant deviation from this ratio is interpreted as an evidence in favor of linkage between the gene and the marker. Several modifications of this method have been developed recently. They allow to analyze not only qualitative but also quantitative traits and to involve into analysis not only sibs but also any other relatives.[14]

Analysis of components of variation

This is a new approach, although it has originated from traditional methods of quantitative genetics. For many years the only way of genetic analysis of quantitative traits was calculation of heritability. This parameter estimates the ratio of genotypic variance to general phenotypic variance. To get this estimate we need to assess the components of variation determined by genotype and environment. We can organize the data as a variance-covariance matrix, where diagonal elements represent general variance of the trait in a population and those out of the diagonal describe resemblance between pairs of relatives. This resemblance is a function of their kinship and genotypic component of general phenotypic variance.

If we know the distribution of the marker alleles among the individuals involved in the analysis we can use this information to detect linkage between the marker and QTL.[15] To do this we must split the genetic component of the variance into two elements, one being determined by the locus tested, and another determined by all the rest of QTL. The contribution of the former component can be determined via the identity by descent of marker alleles in a pair of relatives. Linkage analysis in this case consists of the test for significance of the contribution of the tested locus to variation of the trait.

Analysis of associations

This approach involves an analysis of the associations between the alleles of the marker locus and phenotypic traits of their carriers.[16] It is based on the assumption that the disease is caused by a mutation of a gene that is closely linked with the marker and there is a linkage disequilibrium in the population studied. If this assumption is true, then the frequency of the disease among randomly chosen members of the population depends on their marker genotypes, or, if we are dealing with rare diseases the frequency of a certain marker allele should be significantly higher in affected members of population than in non-affected individuals. For this reason we may utilize population data. This is an important advantage of the analysis of associations over the other mapping methods listed above which can only be realized on pedigree data. However, the analysis of population data may give a false-positive result due to a population heterogeneity. The use of parental, rather than population control allows overcoming this problem. In this case we compare the alleles that have been transmitted with those that have not been so to the affected offspring. Transmission disequilibrium test (TDT) is used for this purpose.[17] Thus, analysis of associations allows detecting a linkage between the markers and presumed genes of diseases. Due to its relative simplicity this method gains much popularity.

Multipoint mapping

The rate of human genome sequencing is much higher than the rate of analysis of its meaning. Therefore at present we have so many markers in human genome and relatively few genes of known function. The abundance of markers makes it possible to pass from the estimation of linkages between genes and markers to precise mapping of the genes at any given point of the genome. The multipoint mapping involves plotting of the genes according to their linkage against a series of markers distributed throughout all chromosomes. The loci that demonstrate a high *lod score* are considered as locations of QTL.[18] The multipoint mapping has a series of advantages over two-point mapping. The main advantage is that it substantially increases the information value of the data available.

What Hinders the Progress in Mapping of Complex diseases?

Thus, at present we have a series of powerful methods for QTL mapping. Every month dozens of papers describing linkage or associations between markers and various diseases appear in many journals. However the progress in identification of the genes controlling complex diseases is not very impressive. Moreover, there are many contradictions in mapping itself. The linkage found in one population is often absent in another. For example, as many as twenty loci have been suggested for putative genes for diabetes type II. Less than half of them were confirmed at two or more independent samples.[19] In many cases none of the structural genes was detected at the genomic region where putative QTL have been located. Finally, many candidate genes chosen on the basis of their biological function were shown to be irrelevant to the diseases

affecting these functions.[20]

Our own study of inheritance of idiopathic scoliosis provides a good example of this failure. With the help of complex segregation analysis we demonstrated that this pathology is controlled by a major gene with incomplete gender- and age-dependent penetrance.[21] It is known that this disease is concerned with defects in the structure of the cartilage in the growth plates of the vertebrae, which causes column deformities. Fiber proteins and proteoglycanes are main constituents of the growth plates. For this reason the structural genes for collagen, elastin and fibrillin have been examined as candidate genes. None of them passed the test.[22,23] We also examined aggrecan gene that displayed a polymorphism in the number of tandem repeats affecting the number of chondroitin sulfate chains. This gene appeared to be a very likely candidate gene for idiopathic scoliosis because the number of chondroitin sulfate chains determines the physical and chemical properties of cartilage. However we failed to detect any association between this gene and idiopathic scoliosis.[24]

There are several reasons for these disappointing failures and discrepancies. One of them is the insufficient statistical power of the methods of QTL mapping that we have at hand now. They demand further improvement. It has been suggested to combine analysis of associations with linkage testing based on the analysis of the components of variance. This combination increases the power of each method involved.[25] Although recombination analysis remains the most powerful method, its power depends crucially on the suggested model of inheritance. A misspecification of the disease model not only decreases the power, but it may also shift the region of QTL location in the case of multipoint mapping.[14] This can lead to false rejection of some candidate genes. To make things even more difficult, the same complex diseases may have different genetic structures in different populations.[15] A relative contribution of each of QTL to variation of the trait depends upon the degree of its polymorphism in the given population. If it is low, this QTL drops out of the model of inheritance and therefore changes the model. Therefore a development of new models of inheritance of complex traits is necessary in order to improve the efficiency of the statistical methods of linkage analysis.

Other important matters are the analysis of information value of samples and the sampling strategy. Most data on mapping studies come from ascertainment via proband affected by the diseases analyzed. These may be pedigrees of arbitrary structure, pairs of affected sibs, trios involving the proband and his/her parent, or case-control samples. In all these cases the ascertainment itself may lead to a loss of information. The source of the loss is the mere complexness of the complex diseases. Many if not all of them involve changes in many biochemical, physiological, neural and psychological processes. For example, diabetes affects sugar metabolism that may or may not lead to heart, kidney, eye and many other diseases. When we restrict our analysis by proven cases of any particular disease, we impoverish the information processed and lose important results. A much more effective strategy is to analyze a wide range of relevant biochemical, physiological, and other parameters in large randomly selected pedigrees. Since many complex diseases are common diseases, in any reasonably large pedigree we can find a wide set of cases of various complex diseases.[15] It is very important to involve in this analysis a possibly large number of relevant quantitative variables together with clinical diagnoses. This approach makes the search for the genes of common disease more effective. The problem is that nowadays we do not have a proper statistical method to process all these information.

These are the technical problems, however important they might be. More serious reason of slow progress in deciphering of the genetic nature of the complex diseases is the conceptual one. Consciously or unconsciously we tend to consider complex diseases in terms of Mendelian paradigm. We treat them as just more complex variants of Mendelian traits. As a consequence we face at least three problems.

i) All the current mapping methods are aimed at a single gene controlling the disease. All other genes are treated as the background, which this presumed gene is expressed on. These methods are able to detect statistically significant relationships between the marker genotype and the disease only in the case when there is one gene which makes the main contribution to polymorphism for the disease. This is why the present mapping methods work well in case of major genes, but they fail when the disease is under the control of many interacting genes and various environmental factors. Although complex diseases show familial aggregation, they do not segregate as Mendelian traits. Most alleles affecting susceptibility are neither necessary nor sufficient for the disease onset. Each of them just modifies the risk. Therefore, the "Mendelian" mapping strategy of statistical analysis is ineffective in complex trait mapping.[26]

ii) Another problem is that we usually try to apply Mendelian diallele structure of genes to complex diseases. This problem appears to be most serious when we search for associations with markers. In this case we presume that there is a single allele that causes the disease. This presumption has been proven to be true in the case of rare recessive diseases, but it is certainly false in case of complex and common diseases. The mere commonness of each of these diseases indicates that all affected individuals cannot inherit the same mutations from the same common ancestors. Multiple simultaneous origin of the same mutation seems also unconceivable.

When the analysis of association was developed, it was met with a great enthusiasm. Discovery of single nucleotide polymorphism (SNP) increased this enthusiasm even more. It was considered as a panacea for, or master key to, all prob-

lems of mapping of complex diseases.[27] However, it becomes more and more clear that the expectations were far too great. The application of this method is rather restricted. It works well in case of Mendelian diseases. It may give good results in mapping of some complex diseases when the sample comes from small isolated populations where the disease can have been inherited from a common ancestor. However, there is no reason to expect that it can solve all the problems.[28] The linkage analysis appears more promising. It detects co-segregation of markers and disease alleles in families regardless of the number and type of these alleles.

iii) What is the nature of these alleles is another and very important matter. Following the Mendelian paradigm, we tend to look for structural genes as candidates, presuming that their mutations cause the diseases. We tend to ignore the sequences that do not code proteins. This is a dangerous fallacy. A very important role of non-coding sequences in the regulation of gene functions is crystal-clear now. Some of them have already been detected in QTL screens. For example, one of the QTL associated with diabetes type II is located at the intron of calpain-10 gene and takes part in the regulation of its expression.[29] Fortunately, in this case the cis-regulating element was located inside the structural gene. However, there are many cases when these elements are located far from the structural genes they regulate. It is not clear whether it will be possible to localize them using the methods of gene mapping that are available now.

Thus, the genetics of complex traits is more complex than genetics of several genes.[30] The whole paradigm of genotypes-phenotype interaction needs to be reconsidered if we do want to get insight into the genetic nature of complex diseases.

Acknowledgements

The authors acknowledge financial support from Russian Foundation of Basis Research (01-04-49518, 01-04-48875). P.M.B. thanks Monbusho for funding a Visiting Professorships at the Space Medicine Research Center Research Institute of Environmental Medicine, Nagoya University, and Professors H. Seo, K. Koga, Y. Murata and Dr. Y. Takagishi for their kind support and valuable discussions.

References

1) Penrose L. Outline of Human Genetics. New York: Wiley, 1959.
2) Liu B. Statistical Genomics: Linkage, Mapping, and QTL Analysis. Boka Raton, New York: CRC Press, 1998.
3) McKusick V. Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders. 12 ed. Baltimore, MD: John Hopkins University Press, 1998.
4) Weiss K, Terwilliger J. How many diseases does it take to map a gene with SNPs? Nat Genet 2000; 26: 151–157.
5) Vogel F, Motulsky A. Human Genetics: Problems and Approaches. 3 ed. Berlin, Heidelberg, New York: Springer-Verlag, 1997.
6) Morton N, MacLean C. Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 1974; 26: 489–503.
7) Bonney G. Regressive logistic models for familial disease and other binary traits. Biometrics 1986; 42: 611–625.
8) Carroll S. Endless forms: the evolution of gene regulation and morphological diversity. Cell 2000; 101: 577–580.
9) Thomson G, Esposito M. The genetics of complex diseases. Trends Cell Biol 1999; 9: M17–M20.
10) Mauvais-Jarvis F, Kahn C. Understanding the pathogenesis and treatment of insulin resistance and type 2 diabetes mellitus: what can we learn from transgenic and knockout mice? Diabetes Metab 2000; 26: 433–448.
11) Ott J. Analysis of human genetic linkage. 3 ed. Baltimore, MD: John Hopkins University Press, 1999.
12) Terwilliger J. Linkage analysis, model-based. Encyclopedia of Biostatistics. New York: Wiley, 1998: 2279–2291.
13) Morton N. Sequential tests for the detecting of linkage. Am J Hum Genet, 1955; 7: 277–318.
14) Sham P, Zhao J. The power of genome-wide sib-pair linkage scan for quantitative trait loci using the new Haseman-Elston regression method. Gene Screen 2000; 1: 103–106.
15) Blangero J, Williams J, Almazy L. Quantitative trait locus mapping using human pedigrees. Hum Biol 2000; 72: 35–62.
16) Clayton D. Population association. In: Balding DJ, editor. Handbook of Statistical Genetics. London-New York: John Willey & Sons, Ldt, 2001.
17) Spielman R, McGinnis R, Ewens W. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. Am J Hum Genet 1993; 52: 506–516.
18) Greenberg D, Abreu P. Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. Genet Epidemiol 2001; 21: 299–314.
19) McCarthy M. The genetic of type 2 diabetes: the consequences of complexity. Gene Screen 2000; 1: 81–84.
20) Deng H, Li J, Recker R. LOD score exclusion analyses for candidate genes using random population samples. Ann Hum Genet 2001; 65: 313–329.
21) Axenovich T, Zaidman A, Zorkoltseva I, et al. Segregation analysis of idiopathic scoliosis: demonstration of major gene effect. Am J Med Genet 1999; 86: 389–394.
22) Carr A, Ogilvie D, Wordsworth B, et al. Segregation of structural collagen genes in adolescent idiopathic scoliosis. Clin Orthop 1992; 274: 305–310.
23) Miller N, Mims B, Child A, et al. Genetic analysis of structural elastic fiber and collagen genes in familial adolescent idiopathic scoliosis. J Orthop Res 1996; 14: 994–999.
24) Axenovich T, Zorkoltseva I, Zaidman A, et al. An association between an aggrecan gene polymorphism and idiopathic scoliosis. Am J Hum Genet, 2001; 69 (Supplement): 496.
25) Allison D, Neale M. Joint tests of linkage and association for quantitative traits. Theor Popul Biol 2001; 60: 239–251.
26) Johnson G, Todd J. Strategies in complex disease mapping. Curr Opin Genet Dev 2000; 10: 330–334.
27) Chakravarti A. It's raining SNPs, hallelujah. Nat Genet 1998; 19: 216–217.
28) Terwilliger J, Weiss K. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 1998; 9: 578–594.
29) Horikawa Y, Oda N, Cox N, et al. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 2000; 26: 163–175.
30) Altshuler D, Daly M, Kruglyak L. Guilt by association. Nat Genet 2000; 26: 135–137.

Received April 11, 2002; accepted June 21, 2002