

Genetic Linkage Analysis

Stefan M. Pulst, MD

Genetic linkage analysis is a powerful tool to detect the chromosomal location of disease genes. It is based on the observation that genes that reside physically close on a chromosome remain linked during meiosis. For most neurologic diseases for which the underlying biochemical defect was not known, the identification of the chromosomal location of the disease gene was the first step in its eventual isolation. By now, genes that have been isolated in this way include examples from all types of neurologic diseases, from neurodegenerative diseases such as Alzheimer, Parkinson, or ataxias, to diseases of ion channels leading to periodic paralysis or hemiplegic migraine, to tumor syndromes such as neurofibromatosis types 1 and 2.

Arch Neurol. 1999;56:667-672

With the advent of new genetic markers and automated genotyping, genetic mapping can be conducted extremely rapidly. Genetic linkage maps have been generated for the human genome and for model organisms and have provided the basis for the construction of physical maps that permit the rapid mapping of disease traits.

As soon as a chromosomal location for a disease phenotype has been established, genetic linkage analysis helps determine whether the disease phenotype is only caused by mutation in a single gene or mutations in other genes can give rise to an identical or similar phenotype. Often it is found that similar phenotypes can be caused by mutations in very different genes. Good examples are the autosomal dominant spinocerebellar ataxias, which are caused by mutations in different genes but have very similar phenotypes.

In addition to providing novel, genotype-based classifications of neurologic diseases, genetic linkage analysis can aid in diagnosis. However, in contrast to direct mutational analysis such as detection of an expanded CAG repeat in the Hunting-

tin gene, diagnosis using flanking markers requires the analysis of several family members.

LINKAGE OF GENES

When Mendel observed an "independent assortment of traits" (Mendel's second law), he was fortunate to have chosen traits that were not localized close to one another on the same chromosome.¹ Subsequent studies revealed that many genes were indeed linked, ie, that traits did not assort or segregate independently, but that traits encoded by these linked genes were inherited together. Studying *Drosophila* genetics, T. H. Morgan showed that the degree of linkage increased with physical proximity of the genes and that the 4 genetic linkage groups actually corresponded to the presence of 4 chromosomes in *Drosophila*.

The first trait in humans linked to a chromosome was actually sex itself. This was followed by linkage of the Duffy locus to chromosome 1 after the observation that certain Duffy blood group alleles were linked to a microscopically visible chromosome 1 polymorphism. Interestingly, the Duffy locus was also the first protein polymorphism linked to a neurologic disease, the Charcot-Marie-Tooth locus on chromosome 1, now called

From the Division of Neurology and Rose Moss Laboratory for Parkinson's and Neurodegenerative Diseases, Cedars-Sinai Medical Center, University of California at Los Angeles School of Medicine, Los Angeles.

Glossary

- Allele: one of several forms of a DNA sequence at a specific chromosomal locus.
- Genotype: the 2 types of alleles at a given chromosomal locus in an individual.
- Haplotype: the string of alleles at linked (adjacent) loci on a single chromosome.
- Locus: unique chromosomal location of a DNA sequence or gene.
- Lod score: statistical measure of the likelihood of genetic linkage between two loci.
- Marker: polymorphic protein or DNA sequence at a specific chromosomal locus.
- Morgan: unit of genetic distance.
- Phenocopy: Phenotype identical to the one caused by gene under study, but caused by another gene or environmental factors.
- Phenotype: physical or biochemical characteristics of an organism or cell.
- Polymorphism: presence of 2 or more alleles at a locus.
- Recombination fraction: percentage of offspring who are recombinant.

CMT1B.² With the ability to detect DNA polymorphisms, the study of genetic linkage blossomed, because polymorphisms were not limited to the relatively rare protein polymorphisms.³

The segregation of an autosomal dominant disease trait and alleles at 3 marker loci is illustrated in **Figure 1**. The markers are perfectly informative, since individual II-1 is heterozygous at the 3-marker loci. Comparison of haplotypes of individuals I-1 and II-2 indicates that the 1-2-4 haplotype marks the chromosome with the disease mutation. The 2 allele of marker B shows perfect cosegregation with the disease trait, whereas the 1 allele cosegregates with the wild-type (normal) phenotype. Marker A shows one recombination event in the unaffected individual III-10, whereas marker C detects multiple recombination events. Thus, the disease trait shows linkage to markers A and B, but it is unlinked to marker C. It is also interesting to examine the recombination occurring between the marker loci. Three recombination events are observed between markers A and B: 2 occurred on the maternal chromosome, 1 on the paternal. Thus, the maternal recombination fraction is 20%; and the paternal, 10%, giving a sex-averaged recombination fraction of 15%. Between marker loci B and C, 8 recombination events are detected.

LOGARITHM OF THE ODDS SCORE ANALYSIS

It is intuitively obvious that the observation of linkage in 4 meioses is less significant than the observation of linkage in 20 meioses. A measure for the likelihood of linkage is the logarithm of the odds (lod) score. The lod score Z is the logarithm of the odds that the loci are linked divided by the odds that the loci are unlinked.⁴ Expression of the likelihood as a logarithm allows summation of likelihoods observed in different pedigrees.

Since the true genetic distance between 2 loci is frequently unknown, the lod score is calculated for different recombination fractions providing a maximum likelihood estimate for the recombination fraction (θ_{\max}) at which the greatest lod score (Z_{\max}) is observed. Pairwise lod scores for the 3 markers in Figure 1 and the disease trait are shown at fixed recombination fractions (**Table**). Since no recombination events occurred between marker locus B and the disease, the Z_{\max} is observed at $\theta = 0$. Due to the 1 recombination event observed in III-10, the most likely distance between marker A and the disease locus is calculated to be at a recombination fraction of 10%. For perfect families, lod score calculations can be carried out by hand. However, for real-life pedigrees with

missing family members and reduced penetrance assumptions, computer programs such as MLINK or LIPED are necessary.⁵

A lod score higher than 3.0 is generally accepted as evidence for linkage, when the human genome is screened with 100 to 300 markers. Although a lod score of 3 would translate into odds of 1000:1 favoring linkage, the corresponding significance level is closer to $P = .05$ owing to the calculation of linkage for multiple markers with the concomitant increase in observing a positive lod score by chance. A lod score lower than -2 is accepted as evidence against linkage. Lod score analysis requires the assumption of precise genetic models, including penetrance, disease gene frequency, and the clear classification of individuals as affected or unaffected. Thus, the lod scores in the Table undergo significant changes when penetrance is reduced to 0.5. Similarly, the misdiagnosis of 1 individual or the presence of phenocopies can drastically affect the lod score and the calculated location of the disease gene.

Instead of calculating lod scores between 2 loci at a time, it is also possible to calculate maximum likelihood estimates for multiple loci at a time. Thus, it is possible to order loci and to place a disease locus on a map of ordered genetic marker loci. This analysis also compensates for noninformativeness of markers in specific meioses.⁵

GENETIC MARKERS

Before the advent of the polymerase chain reaction (PCR), the most commonly used procedure to detect DNA sequence differences involved the use of DNA restriction enzymes and Southern blotting (restriction fragment length polymorphisms).³ With the discovery of the PCR, a novel class of short tandem repeat (STR) polymorphisms was discovered.⁶ The STRs are repeated dinucleotide, trinucleotide, or tetranucleotide sequences such as $(CA)_n$ or $(GATA)_n$ (**Figure 2**, top); STR loci are multiallelic, which increases the probability that a given individual is heterozygous and thus that the paternal and maternal alleles can be distinguished.

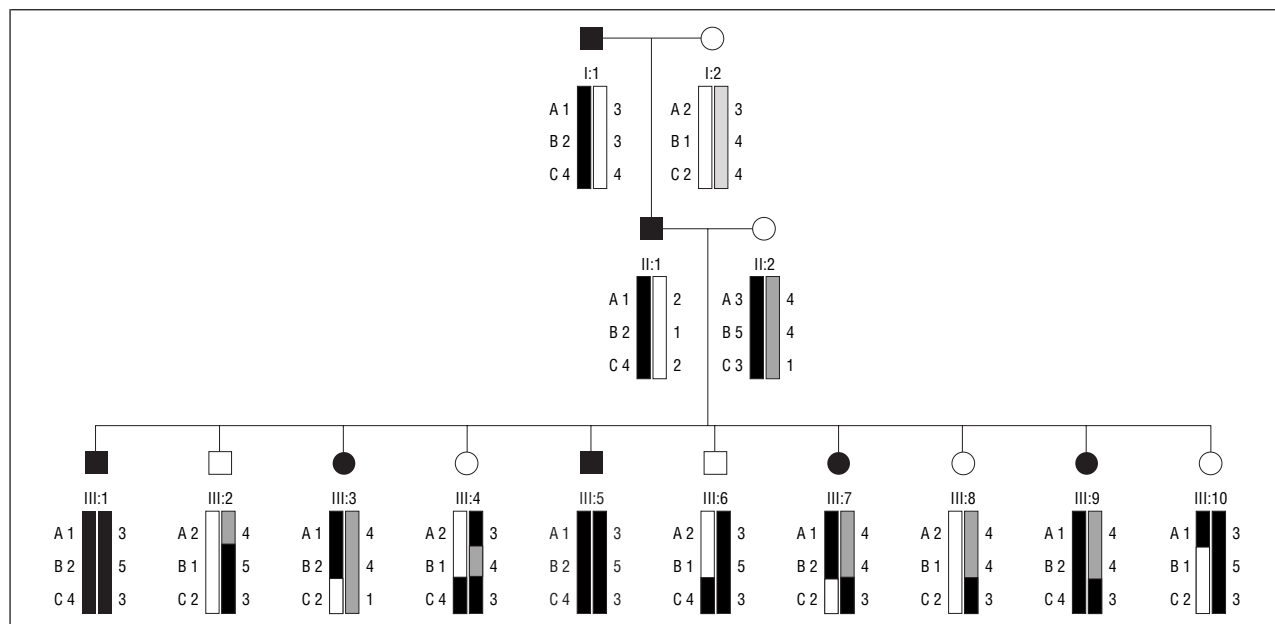


Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

Logarithm of the Odds Scores Calculated Using the MLINK Program⁵ Under the Assumption of Autosomal Dominant Inheritance and a Disease Frequency of 1 in 100 000

Locus	Penetrance	Recombination Fractions						
		0	0.01	0.05	0.1	0.2	0.3	0.4
A	1.0	−∞	1.0	1.5	1.6	1.4	1.1	0.6
	0.5	1.8	1.8	1.7	1.6	1.2	0.9	0.5
B	1.0	3.0	2.9	2.8	2.6	2.0	1.5	0.8
	0.5	2.1	2.1	1.9	1.8	1.4	1.0	0.5
C	1.0	−∞	−5.0	−2.3	−1.3	−0.4	0.0	0.1
	0.5	−4.9	−2.5	−1.1	−0.6	−0.1	0.0	0.1

DNA polymorphisms based on microsatellite repeats can be easily detected. The DNA sequence flanking the repeat is used to synthesize complementary oligonucleotide primers (Figure 2, top). After PCR amplification, the alleles of different lengths are resolved by electrophoresis (Figure 2, bottom). For genome searches several STR markers can be amplified by differential fluorescence labeling and simultaneously detected.

Single-nucleotide polymorphisms have the disadvantage of being bi-allelic and are thus not as highly polymorphic as STRs. However, they represent the most frequent type of polymorphism and their detection can be automated.⁷ Genotyping chips have

the potential to detect hundreds of single-nucleotide polymorphisms at a time.

GENETIC MAPS

The discovery of STRs made it feasible to generate a large number of markers and then map them by genetic linkage analysis. In contrast to the physical map that measures distances in base pairs or a karyotypic map that assigns markers to chromosomal bands, the genetic map gives distances in recombinational units (the centiMorgan [cM]). On average 1 cM corresponds to 1 million base pairs. Male and female recombination rates are not identical, and there are significant changes in the rate of recombination along

specific chromosomal segments. Chromosomal regions with increased recombination are referred to as “recombination hot spots.” The total human genome contains about 3700 cM.⁸ Genetic maps of the human genome have provided the backbone for the construction of physical maps of human chromosomes. Genetic maps are now available to map disease traits in model organisms such as *Caenorhabditis elegans*, zebra fish, and the mouse. The latest versions of human genetic and physical maps can be found electronically.⁹

IDENTIFICATION OF GENETIC LOCI FOR NEUROLOGIC DISEASES

The localization of mendelian disease genes on genetic maps has provided the first step toward the eventual isolation of these genes. Positive linkage also provides the formal evidence that the particular disease is indeed inherited as a mendelian trait. Even a couple of years ago it would have been possible to list the disease genes mapped by genetic linkage analysis, but in 1999, the number of disease loci is far too great and includes all classes of neurologic disease. After a disease gene is localized to a specific chromosomal region, fine mapping is greatly facilitated.

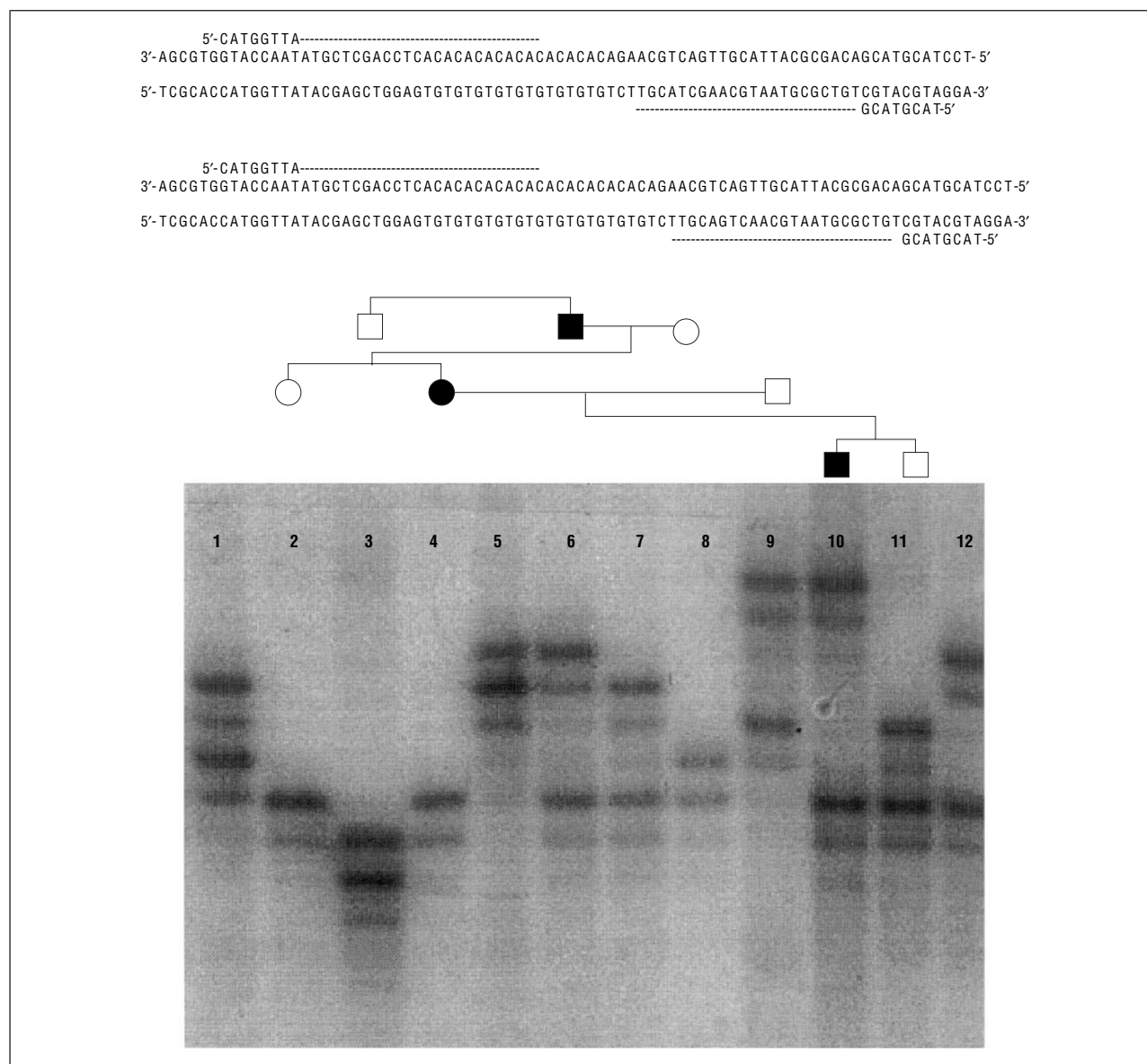


Figure 2. Top, Schematic of a microsatellite locus; 2 alleles containing different (CA/GT)_n tracts and the location of oligonucleotides used to amplify the polymorphic CA tract are shown. The oligonucleotides serve as primers for a DNA polymerase and 2 new templates are synthesized. At the beginning of the second cycle, the primers, supplied in molar excess, anneal to 2 templates each, at the third to 4 each and so on, resulting in exponential DNA amplification. The dotted line denotes the newly synthesized DNA strand. Bottom, An example of DNA analysis with a microsatellite marker: oligonucleotide primers radioactively labeled with phosphorus 32 flanking a CA repeat marker (AFM164ze3 in the D22S275 locus) are used to generate polymerase chain reaction amplicons. These are separated on a polyacrylamide gel and visualized by autoradiography. Note the great heterozygosity of this marker system. Mendelian inheritance of the alleles can be seen by comparison with the pedigree symbols provided above the autoradiograph. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

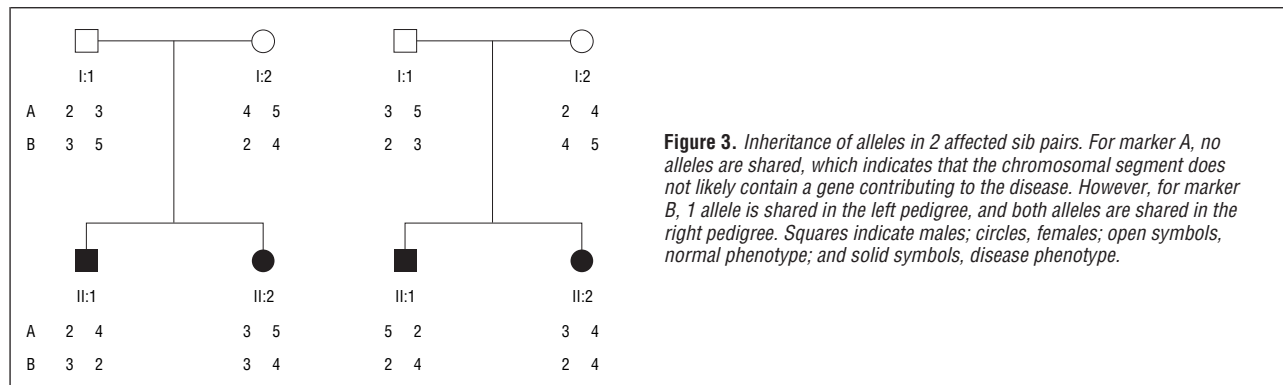
tated by the density of markers on today's genetic maps. This is in stark contrast to just a few years ago when a research team had to generate new markers, order them within the candidate region, and then use the newly generated markers to define recombinational breakpoints.

Although Figure 1 illustrates the use of genetic linkage analysis for a dominant trait, recessive traits can be mapped in a similar fashion. Recessive mutations, especially rare recessive alleles, are commonly ob-

served in consanguineous matings and in highly inbred or geographically confined populations. As a result, affected individuals likely carry identical mutant alleles, including marker alleles that tightly flank the disease locus. This implies that the candidate region can be identified by searching for marker loci that are homozygous in affected individuals. The shared homozygous allele can even be detected when DNA samples from affected individuals are pooled.¹⁰

Allelic and Nonallelic Heterogeneity

Even before a disease gene is actually isolated, knowledge of its chromosomal location allows important studies. In families of sufficient size, linkage to the same locus (locus homogeneity) may be established. Examples of different phenotypes mapping to the same locus, such as Machado Joseph disease and spinocerebellar ataxia type 3¹¹ as well as identical phenotypes mapping to dif-



ferent locations, abound.¹² Diseases with differing phenotypes but mapping to the same locus are called allelic because they are caused by different mutations in the same gene. Familial hemiplegic migraine, episodic ataxia type 2, and spinocerebellar ataxia type 6 are allelic because they are caused by different mutations in the same gene, a calcium channel gene on chromosome 19.

For the demonstration that 2 disorders are not allelic, a relatively small pedigree can be sufficient as long as 1 recombination event for 2 markers flanking the disease locus is detected. For example, exclusion of linkage to all known spinocerebellar ataxia loci in a family with ataxia provided the impetus to search for additional ataxia genes.¹³ The presence of nonallelic or locus heterogeneity has important implications for molecular genetic diagnosis using linked genetic markers.

Genetic Testing Using Linked Genetic Markers

Even before gene isolation, markers closely flanking a disease gene can be used to track the mutation in a pedigree. For indirect testing, samples from at least 2 affected first-degree relatives are required in addition to the proband's DNA. The accuracy of the molecular diagnosis is dependent on the correctness of diagnosis in all individuals used for testing. For example, if the diagnosis of neurofibromatosis (NF) type 2 is missed in affected family members and the diagnosis of NF1 is made, indirect testing would employ markers on chromosome 17 flanking the *NF1* gene instead of

markers on chromosome 22 flanking the *NF2* gene.

However, indirect testing is not limited to diseases in which the causative mutation has not yet been identified. Since tracking of the disease chromosome (or more precisely the chromosomal region that contains the gene) can be performed independently of the precise mutation causing the disease, indirect testing can be used for diseases where the disease is caused by a great number of different mutations in a large gene. Examples are NF1 or cases of Duchenne muscular dystrophy in which the mutations are not detected by multiplex exonic PCR. Predictive accuracy for indirect testing is very high (with the caveats mentioned below), when intragenic or closely flanking markers are used. For a large gene such as the *DMD* gene that covers a genetic distance of close to 10 cM, recombination between the genetic marker and the disease mutation needs to be excluded by markers that flank the gene. These markers must be informative (heterozygous in the affected or carrier parent) so that a recombination event occurring within the gene can be excluded.

As soon as markers linked to a disease have been identified, the disease can be studied in its presymptomatic stage where only subtle biochemical or morphologic abnormalities can be detected. This early study provides the opportunity to determine penetrance of a particular mutation in that asymptomatic gene carriers can be identified.¹⁴ In diseases where the identification of asymptomatic gene carriers can avoid painful or costly screening procedures, molecular genetic testing is

cost-effective because it can exclude non-gene carriers from further screening. As with all genetic testing involving asymptomatic individuals, counseling protocols should be followed.

LINKAGE DISEQUILIBRIUM

Linkage disequilibrium refers to the occurrence of specific alleles at 2 loci with a frequency greater than expected by chance. If the alleles at locus A are a1 and a2 with frequencies of 0.7 and 0.3, and alleles at locus B are b1 and b2 with frequencies of 0.6 and 0.4, the expected frequencies of haplotypes would be a1b1, 0.42; a1b2, 0.28; a2b1, 0.18; and a2b2, 0.12. Even if the 2 loci are closely linked, unrestricted recombination should result in allelic combinations in the general population that are close to the frequencies given above. When a particular combination occurs at a higher frequency, for example a2b2 at a frequency of 0.45, this is called linkage disequilibrium. Linkage disequilibrium may result from natural selection or by chance. When a disease mutation arises on a founder chromosome and not much time has elapsed since the mutational event, the disease mutation will be in linkage disequilibrium with alleles from loci close to the gene. Thus, linkage disequilibrium can be a powerful tool for genetic mapping.

ASSOCIATION

Linkage and association should not be confused with one another. Linkage refers to the relationship of *loci*, whereas association refers to the relationship of *alleles* at a frequency

greater than predicted by chance. To study association, one has to determine allele frequencies in unrelated cases and compare them with the allele frequencies found in controls. Association studies may point to genetic factors involved in the pathogenesis or susceptibility of a disease. One of the best-studied associations of a protein allele and a common disease is the association of the $\epsilon 4$ allele at the apolipoprotein E locus with age of onset in Alzheimer disease.¹² Association can also be seen with very close linkage in the general population, given that the disease mutation arose on a shared ancestral chromosome (see above). Great caution needs to be exercised to exclude population stratification when individuals with the disease belong to a genetically different subset of the population.

MODEL-FREE GENETIC ANALYSIS

For many complex diseases, a precise genetic model cannot be specified, which makes lod score analysis invalid. Although less powerful than lod score analysis, nonmendelian diseases can be analyzed by studying affected sibling pairs. For a randomly chosen marker, sibs are expected to share both parental alleles 25%, 1 allele 50%, and no allele 25% of the time (**Figure 3**). It is obvious that a large number of sib pairs must be studied before a statistically significant deviation from the expected allele distribution can be identified, especially when a large number of genes contribute to the phenotype or when genetic factors interact with environmental stressors. Although sib-pair analysis was successful in identifying susceptibility loci in multiple sclerosis, only a subset of loci was shared by 3 independent studies, which indicates the necessity to validate susceptibility loci in independent patient groups.¹⁵

THE FUTURE OF GENETIC LINKAGE ANALYSIS

"Are Linkage Studies Boring" was the title of an editorial commenting on the fact that linkage techniques had evolved so rapidly that the scientific questions answered by these methods appeared to have become trivial.¹⁶ What the author concluded then remains valid 5 years later: genetic linkage analysis still remains the cornerstone of disease gene localization, and it will not lose its utility even after the human genome is sequenced in its entirety. However, in the area of neurogenetic testing, it can be predicted that indirect testing will be increasingly replaced by more efficient and less costly direct-testing methods.

Novel genetic markers, automated genotyping, and new mathematical models now permit application of genetic linkage analysis to nonmendelian (complex) diseases. But even for mendelian diseases, new questions can be answered, such as the identification of modifying loci that explain variability in the phenotype, reduced penetrance, or variability in age of onset. Genetic linkage analysis is here to stay and has not lost any of its luster.

Accepted for publication December 15, 1998.

The research for this article was supported by Friends of Research, Investigation, Education of Neurologic Diseases, the Drown Foundation, the Warschaw Endowment for Neurology, Los Angeles, Calif, and grants RO1 NS33123 and RO1 NS37883 from the National Institutes of Health, Bethesda, Md.

I gratefully acknowledge the members of the families who have contributed to past and ongoing genetic linkage studies.

Reprints: Stefan M. Pulst, MD, Division of Neurology and Rose Moss Laboratory for Parkinson's and Neurodegenerative Diseases, Cedars-Sinai Medical Center, the University of California, Los Angeles School of Medicine, 8631 W Third St, 1145 E, Los Angeles, CA 90048 (e-mail: Pulst@cshs.org).

REFERENCES

1. Blumberg RB. Mendel Web Archive. Available at: <http://www.netspace.org/MendelWeb/MWpaptoc.html>. Accessed November 8, 1998.
2. Bird TD, Ott J, Giblett ER. Evidence for linkage of Charcot-Marie-Tooth neuropathy to the Duffy locus on chromosome 1. *Am J Hum Genet*. 1982; 34:388-394.
3. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32:314-331.
4. Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet*. 1955;7:277-318.
5. Terwilliger J, Ott J. *Handbook for Human Genetic Linkage*. Baltimore, Md: Johns Hopkins University Press; 1994.
6. Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet*. 1989; 44:388-396.
7. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 1998;280:1077-1082.
8. Dib C, Faure S, Fizames C, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*. 1996;380:152-154.
9. The Cooperative Human Linkage Center. Available at <http://www.chlc.org/ChlcMaps.html>. Accessed November 8, 1998.
10. Nystuen A, Benke PJ, Merren J, Stone EM, Sheffield VC. 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;5:525-531.
11. Twist EC, Casaubon LK, Ruttledge MH, et al. Machado Joseph disease maps to the same region of chromosome 14 as the spinocerebellar ataxia type 3 locus. *J Med Genet*. 1995;32:25-31.
12. Tanzi RE, Blacker D, et al. The genetics of Alzheimer disease: current status and future prospects. *Arch Neurol*. 1998;55:294-296.
13. Zu L, Grewal R, Figueroa KP, et al. Mapping of a new autosomal dominant ataxia (SCA 10) to chromosome 22. *Am J Hum Genet*. 1999;64:594-599.
14. Baser ME, Mautner VF, Ragge et al. Presymptomatic diagnosis of neurofibromatosis 2 using linked genetic markers, neuroimaging, and ocular examinations. *Neurology*. 1996;47:1269-1277.
15. Bell JL, Lathrop GM. Multiple loci for multiple sclerosis. *Nat Genet*. 1996;13:377-378.
16. Bird TD. Are linkage studies boring? *Nat Genet*. 1993;4:213-214.