Informativeness of Polymorphic DNA Markers

Carl E. Hildebrand, David C. Torney, and Robert P. Wagner

As mentioned in "Modern Linkage Mapping," one of the five-year goals of the Human Genome Project is to find highly informative polymorphic DNA markers spaced at 2- to 5-centimorgan intervals along the genetic linkage map of each human chromosome. In this context, informative means useful for establishing through linkage analysis that the marker is near a gene or another marker of interest. Recall that linkage between two variable loci can only be determined from matings in which one parent is heterozygous (carries two different alleles) for the marker or gene at each locus (see "Classical Linkage Mapping"). Thus a marker is highly informative for linkage studies if any individual chosen at random is likely to be heterozygous for that marker. As shown below, markers with many alleles, or highly polymorphic markers, tend to be highly informative.

Informativeness can be quantitatively measured by a statistic called the polymorphism information content, or PIC. This statistic is defined relative to a particular type of pedigree: one parent is affected by a rare dominant disease and is heterozygous at the disease-gene locus (genotype *DN*, where *D* is the dominant, disease-causing allele of the gene and *N* is the normal allele of the gene). The other parent is unaffected by the disease (genotype *NN*). The polymorphic DNA marker in question has several alleles, a_i, which are codominant, that is, each one can be detected so that the genotype at the marker locus (a_ia_j) can always be determined for any individual. Moreover, the marker locus is linked to (on the same chromosome pair as) the disease-gene locus. The important property of this type of pedigree is that the genotypes of the parents and the offspring at both the marker locus and the disease-gene locus can always be inferred. In this context, an offspring is said to be *informative* if we can infer from his or her genotype which marker allele is linked to (on the same chromosome as) the disease allele and would therefore be co-inherited with the disease allele in subsequent generations.

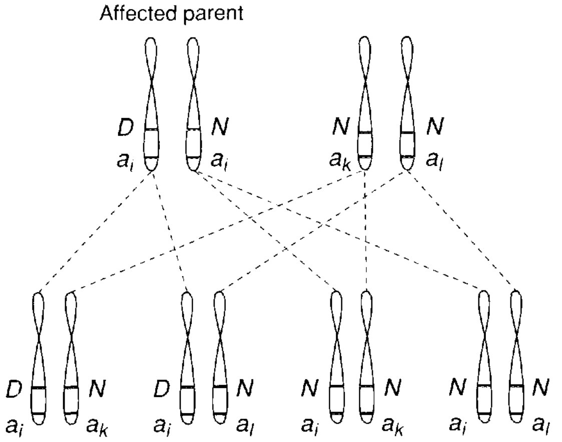
The PIC value of the marker is defined as the expected fraction of informative offspring from this type of pedigree. The figure divides the possible matings from such a pedigree into three categories depending on the genotypes of the parents at the marker locus. Each category has a different fraction of informative offspring. Note that the marker locus is assumed to be near the gene locus, so recombination between the two is a rare event and is not taken into account. In (a) the disease-affected parent is homozygous at the marker locus (genotype a_ia_i) and therefore none of the offspring are informative. In (b) both parents have the same heterozygous genotype at the marker locus (a_ia_j). Then, if each possible type of offspring is produced with equal probability, half of the offspring are informative. For all other combinations of marker alleles in the parents, all offspring are informative. The fully informative matings are summarized in (c).

PIC is the expected fraction of informative offspring from the type of pedigree shown in the figure. Under the assumption of Hardy-Weinberg equilibrium (that in the general population the frequencies of the alleles at the marker locus are independent of the frequencies of the alleles at the disease

Mating Categories for Evaluation of PIC

PIC is the expected fraction of informative offspring from a mating between an affected individual carrying a single copy of a dominant disease allele *D*, and an unaffected individual. This mating is divided into three categories depending on which alleles a_i (i = 1, 2, ...) are present at the locus of a polymorphic marker with *n* alleles. Each category produces a different fraction of informative offspring. Recall that the genotypes of each offspring are known, but the arrangement of alleles on the chromosomes is not known. Thus an offspring is informative if his or her genotype allows us to infer that *D* and a_i are linked in the affected parent and will therefore be coinherited. Informative offspring are shown in red.

(a) *k* and *l* can take on any values

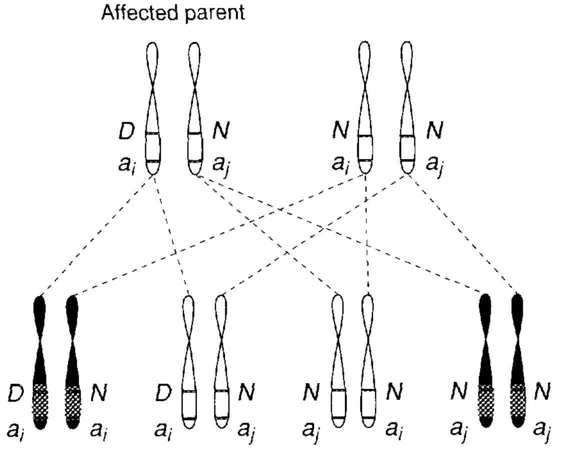


D = disease allele at disease locus
N = normal allele at disease locus
a_i = marker allele at marker locus
p_i = frequency of marker allele a_i

The affected parent is homozygous at the marker locus. Therefore, all offspring inherit a_i from the affected parent, and the inheritance of a_i cannot be used to predict the coinheritance of *D*.

Frequency of mating = p_i^2
Fraction of informative offspring = 0.

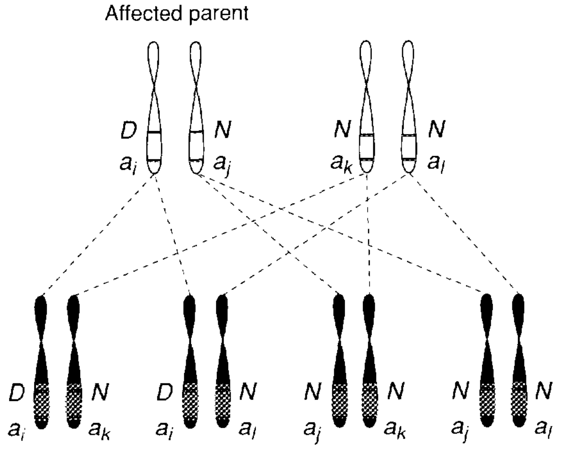
(b) $i \neq j$



Both parents are heterozygous at the marker locus (genotype a_ia_j). In the absence of crossing over two types of offspring are informative (red), that is, we can deduce from the genotypes of those offspring that *D* and a_i are linked (or on the same chromosome) in the affected parent. Specifically, the offspring genotype DNa_ia_i tells us directly that *D* and a_i were coinherited from the affected parent and therefore must be on the same chromosome. The offspring genotype DNa_ja_j, tells us that *N* and a_j were coinherited from the affected parent and by the process of elimination the *D* and a_i must be on the same chromosome in that parent.

Frequency of mating = $2p_ip_j\ (2p_ip_j)$
Fraction of informative offspring = 0.5

(c) $i \neq j$ and *k, l* can be any combination except *i, j* and *j, i*



The affected parent is heterozygous at the marker locus, and the unaffected parent carries a different combination of marker alleles than that in the affected parent. Thus the genotypes of all offspring allow one to deduce that *D* and a_i are linked in the affected parent.

Frequency of mating = $2p_ip_j\ (1-2p_ip_j)$
Fraction of informative offspring = 1.0

locus) and the further assumption that a pair of alleles occurs with a frequency equal to the product of the two frequencies, we can determine the frequency of each mating category from the frequencies p_i of each marker allele a_i. Then (following Botstein et al., 1980 or Roychoudhury and Nei, 1988), to calculate PIC we multiply the frequency of each mating type by the expected fractions of informative offspring from that mating type and add the products:

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_i^2 p_j^2 = 1 - \sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i^2\right)^2 + \sum_{i=1}^{n} p_i^4 \quad ,$$

where p_i = frequency of the marker allele, a_i and n = number of different alleles. Thus to evaluate the PIC value of a marker, we must determine the frequencies of each marker allele. We present an example (from Weber et al., 1990) in which the polymorphic marker is on human chromosome 16 and has four alleles each containing the dinucleotide repeat $(\mathrm{GT})_n$, where n takes on the values 170, 168, 166, and 154. A population of 120 chromosomes indicated that the frequencies of those four alleles are 0.01, 0.12, 0.2, and 0.67, respectively. Using the equation for PIC, we find that the PIC value for this marker equals 0.44. Thus 44 percent of the offspring should be informative in the type of pedigree illustrated in the figure. Theoretically, PIC values can range from 0 to 1. At a PIC of 0, the marker has only one allele. At a PIC of 1, the marker would have an infinite number of alleles. A PIC value of greater than 0.7 is considered to be highly informative, whereas a value of 0.44 is considered to be moderately informative. A gene or marker with only two alleles has a maximum PIC of 0.375. Clearly markers with greater numbers of alleles tend to have higher PIC values and thus are more informative.

An alternative measure of the degree of polymorphism of a marker is the heterozygosity, the probability that any randomly chosen individual is heterozygous for any two alles at a marker locus having allele frequencies p_i. Thus, heterozygosity $= 1 - \sum_{n=1}^{n} p_i^2$, where $\sum_{n=1}^{n} p_i^2$ is the homozygosity. PIC, therefore, will always be lower than the heterozygosity and can be considered to be the heterozygosity corrected for partially informative matings. Polymorphic loci containing many tandem repeats of a short sequence two to six bases long tend to have many alleles and are thus good candidates for highly informative markers. Those markers can be detected using PCR (see "The Polymerase Chain Reaction and Sequence-tagged Sites"). ■

Further Reading

David Botstein, Raymond L. White, Mark Skolnick, and Ronald W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32:314-331.

James L. Weber, Anne E. Kwitek, and Paula E. May. 1990. Dinucleotide repeat polymorphisms at the D16S260, D16S261, D16S265, D16S266, and D16S267 loci. *Nucleic Acids Research* 18:4034.

Jurg Ott. 1991. *Analysis of Human Genetic Linkage*,, revised edition. Baltimore: The Johns Hopkins University Press.

Arun K. Roychoudhury and Masatoshi Nei. 1988. *Human Polymorphic Genes*. New York/Oxford: Oxford University Press.