New Concepts for Distinguishing the Hidden Patterns of Linkage Disequilibrium Which Underlie Association Between Genotypes and Complex Phenotypes

Jacqueline Wicks,[1,2] Susan A. Treloar,[1] Nicholas G. Martin,[1] and David L. Duffy[1]

[1]*Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Brisbane, Australia*
[2]*Australian Research Council Centre for Complex Systems, School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia*

A disappointing feature of conventional methods for detecting association between DNA variation and a phenotype of interest is that they tell us little about the hidden pattern of linkage disequilibrium (LD) with the functional variant that is actually responsible for the association. This limitation applies to case-control studies and also to the transmission/disequilibrium test (TDT) and other family-based association methods. Here we present a fresh perspective on genetic association based on two novel concepts called 'LD squares' and 'equi-risk alleles'. These describe and characterize the different patterns of gametic LD which underlie genetic association. These concepts lead to a general principle – the Equi-Risk Allele Segregation Principle – which captures the way in which underlying LD patterns affect the transmission patterns of genetic variants associated with a phenotype. This provides a basis for distinguishing the hidden LD patterns and might help to locate the functional variants responsible for the association.

Genetic association studies are used in the context of analyzing candidate genes, regions under linkage peaks, and even whole genomes, to find DNA variants associated with a phenotype, and ultimately to identify causal variation. There is a substantial literature on study designs and statistical methods for detecting associations (Cardon & Bell, 2001); these include case-control studies, which have in recent years been modified to adjust for population stratification, the transmission/disequilibrium test (TDT; Spielman et al., 1993), and the plethora of family-based methods spawned by it (Zhao, 2000). A common feature of these methods is that they can detect a relationship between the variants at an observed locus and the phenotype, but they do not uncover anything of the component that connects them, namely the pattern of linkage disequilibrium (LD) with the typically unobserved causal variant.

The pattern of LD that can underlie an association can range between rather extreme cases. At one extreme the observed locus may be in *perfect* LD with a causal variant, or indeed be causal itself; at the other extreme the causal locus is some distance away but there is sufficient correlation between the genotypes at the observed locus and the unobserved causal locus to result in a detectable association between the observed locus and the phenotype. This suggests that two important priorities in association methods are to provide a careful description of the range of LD patterns that underlie genetic associations, and ultimately to develop association methods that can uncover as much as possible about these hidden patterns.

Concepts

In order to study LD patterns we introduce a new device for visualising LD, called an 'LD square'. These capture patterns of gametic LD between any two DNA variants in a completely general way. We first consider LD between two dimorphic variants, where one is an observed variant associated with a phenotype, and the other is an unobserved functional variant; polymorphic variants are discussed briefly later. We denote the alleles at the observed locus by A_1 and A_2 and the alleles at the unobserved functional locus by D_1 and D_2. An LD square has sides of length 1 and a total area of 1, and is subdivided into rectangles which capture allele and haplotype frequencies

Received 24 January, 2005; accepted 1 February, 2005.

Address for correspondence: Jacki Wicks, ARC Centre for Complex Systems, School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia. E-mail: jacki@itee.uq.edu.au

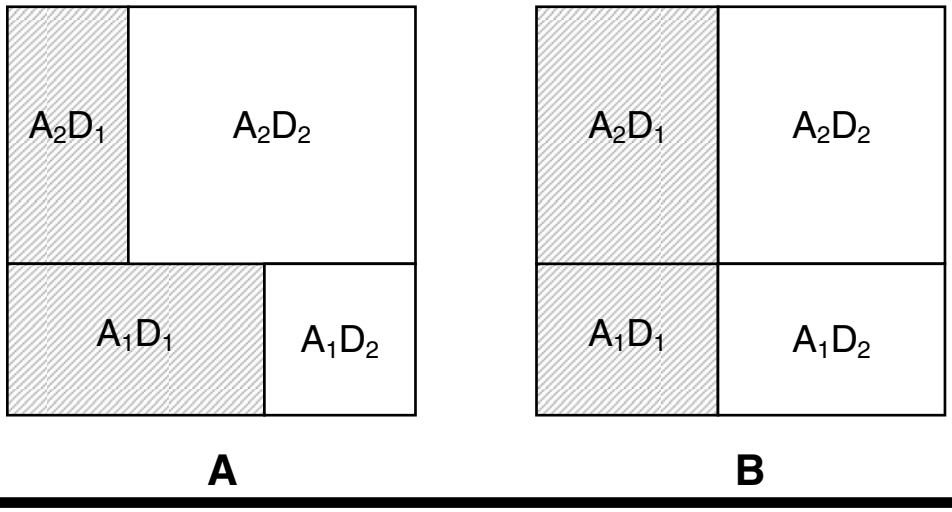

Figure 1

LD squares illustrate gametic LD between two dimorphisms. In each of the two LD squares, the frequencies of the haplotypes between the two loci are given by the areas of rectangles (as indicated). A, linkage disequilibrium. B, linkage equilibrium.

and thus the patterns of gametic LD between two loci. Two possible LD squares for dimorphic loci are given in Figure 1. In each LD square the frequencies of the four possible haplotypes, A_1D_1, A_1D_2, A_2D_1 and A_2D_2 are captured as shown and the frequencies of the alleles are obtained by summing the areas for the appropriate haplotypes. For example, the frequency of the allele A_1 is given by the sum of the rectangles for A_1D_1 and A_1D_2 and is thus the area below the horizontal line which divides each square in two. The area above the line gives the frequency of A_2. Note also that the frequency of the allele D_1 is given by the sum of the two shaded rectangles that represent the frequencies of the haplotypes A_1D_1 and A_2D_1. The remaining unshaded region gives the frequency of D_2.

To see how LD squares capture gametic LD, we compare the two squares given in Figure 1. In LD square A we see that the causal locus allele D_1 occurs in gametes proportionately more often with A_1 than with A_2. Thus there is gametic LD between the loci, and association between the observed locus and the phenotype. In LD square B, the allele D_1 occurs in gametes with alleles A_1 and A_2 in equal proportion; thus the loci are in linkage *equilibrium* (LE) and there is no association between the observed locus and the phenotype. Further, the possible patterns of gametic LD between two loci, where each has any number of alleles, can be captured using LD squares. For example, if the observed locus has three alleles and the causal locus has two, there would be a total of six rectangles representing the frequencies of the six possible haplotypes between the two loci. If two loci are in LE, the rectangles will line up as in LD square B in Figure 1.

LD squares can be used to capture the range of patterns of LD which can underlie genetic association between an observed variant and an unobserved functional variant. The case for dimorphic observed and functional variants is shown in Figure 2. Square A illustrates perfect LD between the two dimorphisms. This is the extreme case of LD in which the observed locus either *is* the functional locus or there is a one-to-one correspondence between the alleles at the two loci. Under perfect LD the only haplotypes that occur in the population are A_1D_1 and A_2D_2. This pattern can be explained by the existence of two historical haplotypes in the population and no recent recombination events. Thus, if this is the pattern of LD underlying an association between the observed locus and a phenotype, and the observed locus is not causal itself, they are likely to be very close in terms of genetic distance.

Square B illustrates the case in which A_1 occurs in gametes with both D_1 and D_2, but A_2 only occurs with D_2. Thus, this LD pattern is distinguished by an absence of A_2D_1 gametes. Further, it can be explained by the historical occurrence of two point mutations and no recombination events between them. This pattern indicates that while the observed locus is not itself causal, they are likely to be very close in genetic distance. Square C is similar except that in this case there is an absence of A_1D_2 gametes in the population. It leads to similar conclusions concerning the causal locus as for square B. Square D illustrates a mixed LD scenario in which all four possible haplotypes occur in the population. Under this LD pattern the observed locus is not causal; further, this pattern suggests that recombination events between the loci might have occurred and thus they may be more distant to one another than the LD patterns represented by squares A, B and C.

Thus we see that 'association' between a dimorphism and a phenotype can be due to a number of distinct underlying patterns of gametic LD between the dimorphism and a nearby dimorphic functional variant, each with different likely histories of mutation and recombination events. It is interesting to compare the commonly used LD measures r^2 (Hill &

Table 1

Values for the LD Measures r^2 and D' and Corresponding Equi-Risk Alleles for the LD Patterns in Figure 2

			Equi-risk*	
LD pattern	r^2	$\|D'\|$	A_1	A_2
A	1	1	Yes	Yes
B	Less than 1	1	No	Yes
C	Less than 1	1	Yes	No
D	Less than 1	Less than 1	No	No

Note: *This depends upon there being no other causal loci linked to the observed locus. There may be any number of unlinked interacting causal loci.

Robertson, 1968) and D' (Lewontin, 1964) for the LD patterns A to D. The values are given in Table 1. Thus these patterns can be seen to provide a visualization of gametic LD which corresponds to the patterns which are distinguished by these numeric measures of gametic LD.

In practice it would be useful to be able to distinguish LD patterns A to D in Figure 2 using data only from an observed locus associated with a phenotype. This would provide useful information concerning the relationship with the unobserved causal variant namely some idea of how close it is likely to be to the observed variant and also the co-occurrence of gametes that will occur between the two loci. This might, in practice, help to locate and identify the causal variant.

What we find is that the hidden LD patterns are actually revealed in the segregation patterns of the alleles at the observed locus. The basis for this is the concept of an 'equi-risk allele'. To explain this concept we consider the case in which an observed locus is closely linked to a causal locus, and there are no other causal loci linked to either of these; there may be other unlinked causal loci. Consider the special case in which an allele at the observed locus only ever occurs with one of the causal locus alleles. We call such alleles at the observed locus 'equi-risk' because they always occur in gametes *with the same causal locus allele* and therefore, via the co-occurring causal locus allele, *always impart the same risk to carriers*.

The pattern of equi-risk alleles corresponding to the LD patterns in Figure 2 are described in the final two columns of Table 1. It is interesting to note that perfect LD, for which r^2 has a value of 1, corresponds to both alleles being equi-risk. Further, D' having an absolute value of 1 corresponds to the presence of one or more equi-risk alleles. Note also that each LD pattern A to D is distinguished by its own pattern of equi-risk alleles; thus if the equi-risk alleles at an observed locus which is associated with a phenotype can be determined, then so too can the underlying pattern of gametic LD. Equi-risk alleles are distinguished by a special segregation property we have called the 'Equi-risk Allele Segregation Principle', which gives them a recognizable transmission signature:

> A parent who is homozygous for an allele which is equi-risk with respect to a particular phenotype will segregate his/her two alleles at random to any children independently of the children's values for the phenotype.

The reason the Equi-Risk Allele Segregation Principle (EAS Principle) holds is that the two alleles the parent carries at the observed locus both occur with the same allele at the causal locus and therefore their transmission to any children will bear no relationship to the children's phenotypes.

Applications

To see how the concepts can be applied in practice, consider the case of a qualitative phenotype, and an affected sib-pair (ASP) family. In such a family, it is well known that there will be excess sharing in identical-by-descent (IBD) transmissions to the ASP at a locus that is linked to a causal locus. This holds at a locus that has a true genetic association to the phenotype, as such a locus is also linked to the phenotype. Suppose that the observed locus is diallelic and is in perfect LD with a diallelic causal variant as in LD square A in Figure 2 and also that both of these loci are unlinked to any

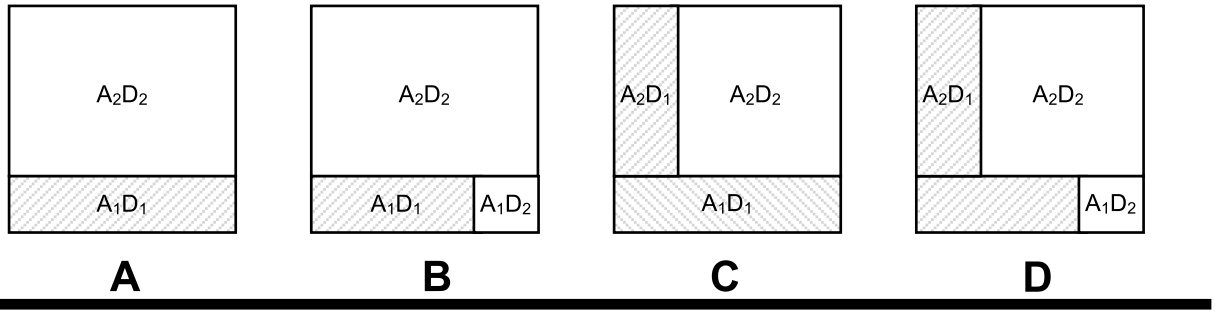

Figure 2

LD squares illustrate the full spectrum of gametic LD patterns between an observed dimorphism and an unobserved causal dimorphism, which can underlie association between the observed dimorphism and the phenotype.
A, perfect LD or co-occurance of the two dimorphisms. B, the minor allele at the observed locus occurs with both alleles at the causal locus but the major allele only occurs with one. C, the minor allele at the observed locus only occurs with one allele at the causal locus but the major allele at the observed locus occurs with both alleles at the causal locus. D, both alleles at observed locus occur with both alleles at causal locus.

other causal loci. Then both alleles at the locus will be equi-risk and the EAS Principle tells us that the excess sharing which occurs at the locus will therefore not be manifested in transmissions from the *homozygous* parents, as the segregation of their alleles will be random with respect to their children's phenotypes. Note that this result is the basis of the Homozygous Parents ASP Method (Robinson et al., 1993). This ingenious method uses the presence of statistically significant excess sharing in transmissions from homozygous parents as the basis for rejecting a locus as the only causal variant in a region. A further implication of this absence of excess IBD sharing in transmissions from homozygous parents is that all the excess IBD sharing will be concentrated in the transmissions from the *heterozygous* parents.

The other LD patterns in Figure 2 will also be distinguished by particular patterns of IBD sharing when it is divided up according to parental genotypes. At a diallelic locus which is associated with the phenotype via LD pattern B, there will be excess sharing in the expected transmissions to ASPs from parents with genotypes A_1A_1 and A_1A_2, but none in transmissions from A_2A_2 homozygous parents, as the allele A_2 is equi-risk. Similarly under LD pattern C, there will be excess sharing in expected transmissions to ASPs from parents with genotypes A_1A_2 and A_2A_2 but not from parents with genotype A_1A_1. Under LD pattern D excess sharing can be expected in transmissions from parents with any genotype. Thus, in principle at least, the EAS Principle provides a basis for distinguishing all of the underlying LD patterns A to D. It is interesting also to compare the pattern of IBD sharing across parental genotypes for loci that are linked but in LE with a causal locus such as in LD square B in Figure 1. In this case, there will be excess sharing in transmissions from parents with any genotype and the rate will be uniform across the parental genotypes. In the presence of LD, the rates vary for the different parental genotypes and the extreme case of this is where there is perfect LD so that the alleles are equi-risk and all the excess sharing occurs in transmissions from heterozygous parents.

The EAS Principle is broader than this as it applies in a similar way to sibships of any size and with any configuration of phenotypes for a qualitative or quantitative phenotype. There are well-known deviations from random segregation expected in transmissions from parents to sibships that are functions of the phenotypes of the sibship. The example of ASPs for a qualitative phenotype was given above with its deviation from random segregation of greater than 50% sharing of IBD transmissions from parents; another example is discordant sib-pairs with its deviation from random segregation of *less* than 50% sharing of IBD transmissions. The EAS Principle tells us that the deviations typical for the sibship type will not be seen in transmissions from parents who are homozygous for an equi-risk allele, and thus information to distinguish LD patterns A to D in Figure 2, as in the example of ASPs, can be gleaned from any sibship type. In particular, at a locus in perfect LD with a causal locus, and linked to no other causal loci, all of the expected deviation will be manifested in the transmissions from the heterozygous parents *only*. This provides a very specific transmission pattern for researchers studying a variant associated with a phenotype to look for.

The principle is not restricted to loci with only two alleles. Indeed a multiallelic observed locus can have any number of equi-risk alleles, each corresponding to a particular LD relationship with a multiallelic causal locus that may have a different number of alleles to the observed locus. The implication of the EAS Principle is that by looking at inheritance patterns from parents with different genotypes, evidence for equi-risk alleles can be gleaned, thereby revealing something of the likely LD pattern between a locus associated with a phenotype and an unobserved causal locus.

The final application of the EAS Principle which we discuss here is the case in which a number of tightly linked single nucleotide polymorphisms (SNPs) or other markers have been genotyped in and around a gene which shows good evidence of association to a phenotype. Suppose there is only one causal locus in the linkage region and that a collection of sibships and their parents are genotyped. A corollary of the EAS Principle is that the locus most likely to be causal or in perfect LD with the causal locus is the one for which the deviation from random segregation, which is typical for the sibship types in the data (e.g., more than 50% IBD sharing in transmissions to ASPs, less than 50% IBD sharing in transmissions to discordant sib-pairs, and so on), is maximized *when restricted to transmissions from the heterozygous parents*. To understand this result we note that at the causal locus all the tendency for deviation from random segregation will occur in transmissions from parents who are heterozygous at the causal locus. If a SNP is in LD pattern A with this causal locus, then the set of heterozygous parents at this SNP will be precisely the set that are heterozygous at the causal locus. At a SNP in LD patterns B, C or D with the causal locus, the set of heterozygous parents at this SNP will contain some of those who are heterozygous at the causal locus and also some that are homozygous. Thus the rate of deviation from random segregation seen in these sets of parents will be to some extent 'watered down' by the inclusion of the parents who are homozygous at the causal locus and who will segregate randomly with respect to the phenotype.

If we consider now the homozygous parents in this scenario we will find that the deviation from random segregation seen in transmissions from this set at an observed locus will decrease as it increases in the set of heterozygous parents, and the minimum of the former coincides with the maximum of the latter. This is because in a collection of tightly linked SNPs where there is essentially no recombination, the IBD patterns

across the region are fixed in each family. Thus if there is more deviation in random segregation in transmissions from the heterozygous parents at one SNP than at another, there must also be less deviation from the homozygous parents at the former than at the latter, as the overall amount of deviation is fixed. Thus if we were to map both the deviation observed in the heterozygous parents and the deviation observed in the homozygous parents at each of the genotyped SNPs, we would see the former reach a maximum and the latter reach a minimum at the best candidate for perfect LD with the causal variant. Further, the absence of any other linked causal loci would mean the deviation in transmissions from the homozygous parents would be expected to be zero under perfect LD with the causal locus. As with the Homozygous Parents ASP method, a locus can be rejected as the only causal locus in the region in the presence of statistically significant deviation from random segregation in transmissions from the homozygous parents. However, if the presence of another linked causal locus seems plausible, so that one of the genotyped SNPs may still be causal or in perfect LD with a nearby causal dimorphism, the best candidate for this will still be the SNP for which the deviation from random segregation in transmissions from the heterozygous parents is maximized and thus also minimized in the homozygous parents; in such a case as this the deviation from zero of the latter would be due to the additional linked causal locus.

We conclude this section by mentioning three ways in which the application of EAS Principle is limited in practice. The first and most important of these is that, in the presence of two or more linked causal loci, the alleles at the observed locus, even if it is directly causal, will only be equi-risk if there is perfect LD between it and *all the causal loci to which it is linked*. The presence of any number of additional *unlinked* causal loci is not relevant to the EAS Principle. What this means is that the usefulness of the Principle is more limited when there are linkage regions possibly containing several interacting causal loci. If, however, a causal locus has already been identified then it is possible to define 'conditionally equi-risk alleles' at another observed locus linked to it, as those which transmit the same risk when parents are homozygous at the known causal locus. This leads to a 'Conditional EAS Principle' in which such alleles are seen to segregate randomly as per the original EAS Principle, in parents who are also homozygous at the known causal locus. In this way the Conditional EAS Principle can be used to tease apart regions containing a known causal locus.

The second limitation is that while it can furnish evidence as to the underlying LD pattern between an observed and a causal variant, the EAS Principle cannot determine whether the observed locus is directly causal itself. No method using only information on the phenotype and genotypes in families can distinguish between direct causality and perfect LD with a causal locus *unless* there is complete genotyping of the entire linkage region and it just happens that there are no loci in perfect LD with the causal variant. This really defines the fundamental limit of association methods; beyond this, functional studies of one kind or another are needed to distinguish direct causality from cases of perfect or close to perfect LD. The third limitation, which is related to the second, is that while the EAS Principle provides a means of gleaning evidence for the presence of equi-risk alleles, it will in practice be difficult to distinguish alleles that are close to being equi-risk from those that are equi-risk. This means that it will be difficult to distinguish patterns very close to perfect LD from perfect LD; this limitation applies to any method and is simply due to the fact that the evidence available for distinguishing them will be limited and will decrease the closer a variant is to being in perfect LD with the casual variant.

Conclusion

In this report we have highlighted and described the range of hidden LD patterns responsible for genetic associations, and have shown that there is information in transmissions to sibships that might be helpful in revealing them. A new graphical device for visualizing gametic LD — the LD square — has been introduced, as has the concept of an equi-risk allele. In the case that an observed locus is linked to only one causal locus, the presence of equi-risk alleles distinguishes the different hidden patterns of LD between it and the unobserved causal locus. Further, equi-risk alleles are subject to a rather surprising and very specific segregation property, which we have called the 'Equi-Risk Allele Segregation Principle'. A corollary of this principle yields the well-known Homozygous Parents Affected Sib-Pair Method, and more generally the principle provides a qualitative guide as to how hidden LD patterns with a causal variant might be distinguished using data from sibships with any configuration of phenotypes for qualitative or quantitative phenotypes. It gives researchers very distinct segregation patterns to watch out for in their data when looking for sites of DNA variation that are potentially causal, and a direct application is in choosing the SNP most likely to be causal, or in perfect LD with a causal dimorphism, from a set of tightly linked SNPs in a gene associated with a phenotype. Further to this the concepts introduced in this report provide a new conceptual basis for our understanding of genetic association and for further research into dissecting the genotype–phenotype correspondence in humans.

Acknowledgments

JW would like to thank Warren Ewens, Mary Sara McPeek, Terry Speed, Glenys Thomson, Peter Visscher and Sue Wilson for discussion of the ideas. This work was supported by the Australian Cooperative

Research Centre for the Discovery of Genes for Common Human Diseases.

References

Cardon, L. R., & Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics, 2*, 91–99.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics, 38*, 226–231.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations: Heterotic models. *Genetics, 49*, 49–67.

Robinson, W. P., Barbosa, J., Rich, S. S., & Thomson, G. (1993). Homozygous parent affected sibpair method for detecting disease predisposing variants: Application to insulin dependent diabetes mellitus. *Genetic Epidemiology, 10*, 273–288.

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics, 52*, 506–516.

Zhao, H. Y. (2000). Family-based association studies. *Statistical Methods in Medical Research, 9*, 563–587