False Disease Region Identification From Identity-By-Descent Haplotype Sharing in the Presence of Phenocopies

Stuart Macgregor,[1,2] Sara A. Knott,[1] and Peter M. Visscher[1,2]

[1] *Institute of Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh, United Kingdom*
[2] *Genetic Epidemiology, Queensland Institute of Medical Research, Brisbane, Australia*

Linkage analysis (either parametric or nonparametric) is commonly applied to identify chromosomal regions using related individuals affected by disease. In complex disease the incomplete relationship between phenotype and genotype can be modeled using a phenocopy parameter, the probability that an individual is affected given they do not carry the disease mutation of interest, and a nonpenetrance parameter, the probability that an individual is not affected given they do carry the disease mutation of interest. If the linkage phase between multiple markers and a putative disease locus is known, then haplotypes carrying the mutation can, in principle, be identified by comparing the chromosome segments that are shared identical-by-descent (IBD) across affected individuals. We consider here the effect of a nonzero phenocopy rate on the linkage peak and hence upon the identification of disease haplotypes that are shared IBD between affected individuals. We show, by theory and computer simulation, that in diseases for which there is a nonzero phenocopy rate, the chromosomal regions identified may not include the true disease locus. We utilize a LOD-1 confidence interval for a widely used nonparametric linkage statistic. We find that in small/moderate samples this confidence interval may be inappropriate. We give specific examples where the phenocopy rates are nonnegligible in some complex diseases. The success of further work to identify the causal mutations underlying the linkage peaks in these diseases will depend on researchers allowing for the presence of phenocopies by examining appropriately wide regions around the initial positive linkage finding.

In an attempt to locate disease genes many researchers have applied linkage analysis to identify chromosomal regions which segregate with the disease of interest in a pedigree. Under linkage peaks, regions that are shared identical-by-descent (IBD) in affected relatives, and therefore unbroken by recombination ('disease haplotypes') are sought out. For Mendelian disorders there is usually a single disease region which is completely associated with the disease phenotype. In complex disorders, there are typically multiple disease regions, some of which may be the result of mutations at distinct loci in the genome. Such regions may only be partially associated with the disease phenotype (region is neither necessary nor sufficient for disease). Complex diseases can be modeled as if they were Mendelian, with individuals carrying a disease mutation but not exhibiting the disease phenotype labeled as nonpenetrant and affected nondisease mutation carriers labeled as phenocopies. The focus of this report is these phenocopies. Here we define the *phenocopy parameter* as P (individual in sample is affected/individual does not carry disease mutation of interest) whilst the nonpenetrance rate is P (individual in sample is not affected/individual does carry the disease mutation of interest); these are the penetrance parameters required in a parametric linkage analysis (Ott, 1991). A related measure of phenocopy frequency is the *phenocopy rate*, the proportion of affected individuals that are phenocopies (Sham, 1998; Ott, 1991). For convenience we henceforth use the phenocopy parameter definition to describe the frequency of phenocopies in the sample of interest.

Since many complex diseases are caused by multiple unlinked loci (which are not all required for affection), all affected individuals, even those affected as a result of other unlinked disease loci, must be regarded as phenocopies if they do not carry the mutation at the locus of interest. If there are many unlinked disease loci, as is the case in diseases such as Alzheimer's disease (OMIM 104300) and Breast Cancer (OMIM 114480), the number of phenocopies (with respect to the locus of interest) may be large even when the disease carries a substantial genetic component. Individuals may also be phenocopies if they do not carry the disease haplotype of interest and are affected as a result of environmental factors. Since phenocopies either have the disease as a result of nongenetic factors or because of other mutations at unlinked chromosomal regions, the haplotype they have at the putative disease region under consideration

Received: ??

Address for correspondence: Stuart Macgregor, Genetic Epidemiology, Queensland Institute of Medical Research, 300 Herston Road, Herston, Queensland 4006, Australia. E-mail: stuart.macgregor@qimr.edu.au

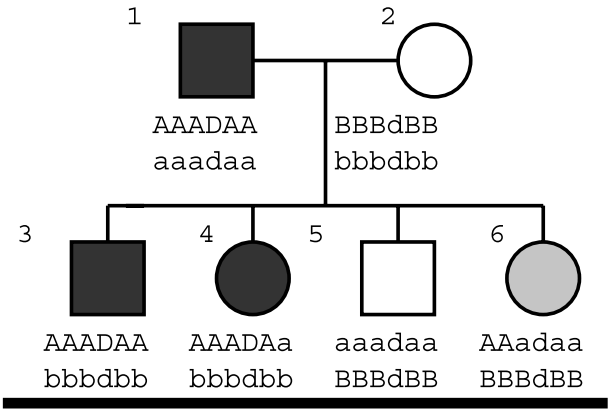

Figure 1

Nuclear family with a phenocopy.

Affected individuals are shaded in black, phenocopies are shaded in grey and unaffected individuals are unshaded. Assume the disease mutation in this region is dominant in its effect on phenotype, with alleles D and d.

will be different to that of the other affected individuals. In this report the effect of phenocopies upon the identification of an IBD disease region is considered and it is shown that the regions inferred in the presence of phenocopies may not include the true disease locus. Such errors will impact significantly on subsequent attempts to identify disease causing mutations.

Methods

Theory

Phenocopies and Disease Regions

To identify a disease region using affected individuals in families, one uses the marker information to assess where recombination events have occurred. The length of chromosome shared by all affected individuals in the region of interest is called the minimal disease region or MDR. When one or more phenocopies are present within a sample, the chromosomal recombination pattern of these phenocopies is erroneously used to narrow the disease region. Consider the nuclear family in Figure 1. In this family a recombination event in affected individual 4 is used to narrow the disease region on the right of the true disease locus. Call the disease region inferred from the affected individuals carrying the D allele the minimal disease region for individuals carrying the mutation or MDRM. Suppose one of the individuals, numbered 6, is a phenocopy. This individual does not carry the disease allele, D, but has inherited part of the disease mutation carrying chromosome from its affected parent, but not the disease mutation itself because of a recombination event. This means that the genomic region shared by all of the affecteds (D allele carriers and phenocopies), the MDR, spans only the leftmost two markers and does not include the actual disease locus of primary interest. If the phenocopy rate is high, the probability of an incorrectly identified disease haplotype will be nonnegligible.

Note that although some phenocopies will occur in families otherwise unaffected by the disease (sporadic cases), linkage analysis samples are typically ascertained to have a large number of affected relatives. The phenocopies ascertained and analyzed are therefore likely to occur in circumstances similar to that of individual 6 in Figure 1.

We consider a range of rates (penetrance parameters between 0.01 and 0.08) at which phenocopies occur within a sample of nonmutation carriers. To assess the probability of phenocopies causing the disease locus not to be contained within the MDR, we investigate the distribution of MDR lengths and the likely number of phenocopies.

Distribution of Minimal Disease Region Lengths

The length of the MDR can be calculated by considering the distribution of recombination events. If one ignores linkage interference, the number of recombination events follows a Poisson distribution with parameter equal to 1 per Morgan of genome per meiosis (Haldane, 1919; Sham, 1998). Consider a putative disease locus on a chromosome. Assume for the moment that the disease locus is dominant in its effect on the phenotype, that is, there is one chromosome of interest per person (see discussion for recessive case). The map distance to the first recombination event to the right of the disease locus is distributed exponentially with parameter 1 Morgan. Given a number of such inherited chromosomes n, the distance from the putative disease locus to the nearest recombination on the right (over all available chromosomes) is then distributed as exponential with parameter $1/n$. The distribution of the distance between the first recombination to the left and the first to the right is thus the sum of 2 exponential distributions. This has a gamma distribution with alpha equal to 1 and beta equal to $2/n$ (Hanson, 1959).

Quantifying the Effect of Phenocopies

Affected individuals who do not share any of the MDR (which, by definition, must be phenocopies) are assumed to have been removed from the sample. This will usually happen in practice since otherwise it will not be possible to identify a disease region at all. In a nuclear family the probability of a phenocopy causing the disease locus to be outside the MDR depends on the average length of the MDRM and the probability distribution of the number of phenocopies.

If the likely number of phenocopies is small, then one can calculate the probability of at least one phenocopy having a recombination in the MDRM (and hence carrying part of it, but not the disease mutation of interest) by

$$1 - (1 - L)^w$$

where w is the number of phenocopies and L is the length of the MDRM measured in Morgans (we assume for simplicity the Morgan map function in which recombination fraction equals map distance). In the presence of phenocopies the MDR may be smaller than the unobserved MDRM; the MDRM is of length L and is defined by the mutation carriers.

To evaluate the likely number of phenocopies, consider a sample of pedigreed individuals, with m individuals not carrying the mutation of interest (these will commonly be unaffected individuals; if there is only one mutation causing the disease and no environmental factors generating phenocopies then these individuals will definitely be unaffected). If each of these m individuals has probability p of being a phenocopy, the number of phenocopies in the sample will have a binomial distribution with parameters m and p; r is the number of phenocopies. The probability of at least one phenocopy causing the disease locus to be outside the MDRM is therefore

$$\sum_{r=1}^{m} \binom{m}{r} p^r (1-p)^{m-r} (1-(1-L)^r) \qquad [1]$$

Equation 1 will not hold exactly when there exist two or more phenocopies in a sample. An exact equation is given in the appendix. To quantify the effects of the phenocopies we evaluate the exact equation for a variety of phenocopy rates and sample sizes.

Computer Simulation

We used simulation to assess the effects of phenocopies on the LOD score profile in linkage analysis. In the simulation five nuclear families, each with four affecteds, were generated. Chromosomes with 24 highly polymorphic, 2 cM spaced markers were passed (with recombination) from parents to offspring. A disease locus with a fully dominant disease allele was placed midway between markers 11 and 12 (21 cM). LOD score profiles from multipoint parametric linkage analyses and from multipoint nonparametric linkage (exponential model with 'all' scoring function) were calculated using the program Allegro (Gudbjartsson et al., 2000). This set-up allowed us to check the theoretical results presented and examine the effect of phenocopies on the LOD score profile.

To assess the impact of phenocopies, a phenocopy was added to one family and the LOD profile recalculated. Assuming phenocopies to be binomially distributed, 1 phenocopy would arise in this way 37% of the time if 100 nonmutation carrying individuals were ascertained with a phenocopy rate of 0.01. Analyses were first conducted with the phenocopy rate parameter in the parametric linkage set to 0. Subsequent parametic analyses considered changing the phenocopy rate parameter to .1 or .2. For the nonparametric analysis we calculated the peak LOD position and the LOD –1 (or 1 LOD drop) confidence interval the peak for 300 replicates. The LOD –1 cutoff yields a 97% confidence interval asymptotically (Mangin et al., 1994).

Results

Theory

Distribution of Minimal Disease Region Lengths

The distribution of MDRM lengths for 20, 40, 60 and 80 affected individuals all carrying the disease locus of

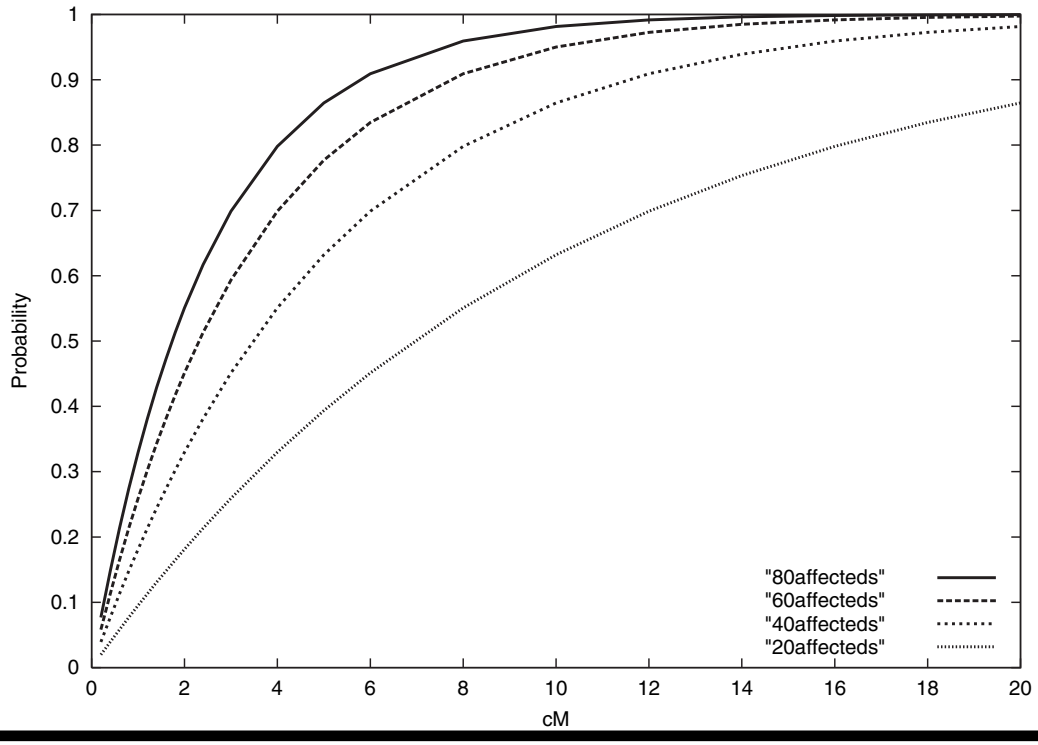

Figure 2

Distribution of MDRM lengths.

The distribution of MDRM lengths given varying numbers of affected individuals all carrying the disease locus of interest.

Table 1

Effect of Varying Phenocopy Rate on the Probability of the MDR Containing the Disease Locus

Phenocopy rate	Probability MDR does include actual disease locus	Probability MDR does not include actual disease locus
.01	.91	.09
.02	.83	.17
.03	.76	.24
.05	.66	.35
.08	.56	.44

Note: The probabilities in the table are based upon 20 affected mutation carriers, assuming that the mutation is fully penetrant in its effect on the phenotype. A total of 100 individuals not carrying the mutation (who may or not be affected depending on phenocopy rate) have been considered alongside the affected individuals.

interest is given in Figure 2. The mean MDRM lengths in the four cases are 10 cM, 5 cM, 3.3 cM and 2.5 cM, respectively.

The Effect of Varying Phenocopy Rate and Sample Size

In Table 1 the effects of changing the phenocopy rate are shown. The probability of the MDR falsely ruling out the genomic region where the locus actually resides reaches high levels (greater than 40%) if the phenocopy rate exceeds a few per cent.

In Table 2 the effects of altering the number of mutation carrying individuals are shown. The probability of obtaining an MDR that includes the actual disease locus is high provided that the sample of affected individuals is large ($n > 50$). This is because increasing the number of affecteds carrying the mutation decreases the MDRM and, hence decreases the probability of a phenocopy sharing some of it.

Computer Simulation

No Phenocopies

Twenty affected individuals were generated. As predicted by the theory, the MDR in each case was visible as a plateau (region in which no recombinations occurred in the genotyped individuals, on average 10 cM long) in the LOD profile; 1 replicate is displayed in Figure 3 (solid line).

For the nonparametric analysis of the simulated data the LOD –1 confidence interval contained the simulated disease location in 100% of cases. Although the LOD –1 method should asymptotically give a 97% confidence interval, for the data simulation model utilized here there is limited scope for variability in the location parameter. It is therefore unsurprising that none of the 300 replicates yield intervals that fail to include the simulated disease locus.

Table 2

Effect of Varying Number of Affected (Mutation Carrying) Individuals on the Probability of the MDR Containing the Disease Locus

No. of affected mutation carriers	Probability MDR does include actual disease locus	Probability MDR does not include actual disease locus
10	.70	.30
20	.83	.17
30	.88	.12
50	.92	.08
100	.96	.04

Note: The probabilities in the table are based on a phenocopy rate of .02 and the ascertainment of 100 individuals not carrying the mutation.

One Phenocopy Added

Figure 3 shows the LOD profiles of three replicates (broken lines) where the MDR was falsely narrowed by recombination(s) in the added phenocopy. When the phenocopy has recombination(s) in the MDRM there is a region shared by 21 affected individuals, generating a LOD around 3.6. Conversely, when there are no recombinations around the disease locus in the phenocopy, there are 20 individuals with a common set of alleles and one without this set of alleles. This typically generated a LOD of around 2.8. The addition of a phenocopy increases the maximum LOD score achieved but, crucially, indicates a genomic region which does not include the true location of the disease locus (since the phenocopy cannot actually share the genomic region with the disease gene on it, only a nearby region via a recombination in the affected parent). The discrepancy in location was up to 20 cM. As predicted by the above theory, ~10% of these phenocopy individuals shared some of the MDRM (based on 300 replicates).

Allowing for phenocopies in the analysis does not improve the situation since the LOD peak is still at the point where most individuals share the same set of alleles. The only effect of setting the phenocopy rate parameter to .1 or .2 is to reduce the overall LOD scores achieved. With 300 replicates the percentage of replicates with the peak LOD distinct from simulated disease locus (i.e., there was a region over which all individuals, including the phenocopies, shared a haplotype) was 10% for an analysis with phenocopy rate parameter 0 and 11% for both the .1 and .2 analysis.

For the nonparametric analysis of the simulated data with one phenocopy added the LOD –1 confidence interval did not contain the simulated disease location in 8% of replicates. Examination of these replicates revealed that these replicates were a subset of the 10% of replicates that gave a peak LOD distinct from the simulated disease location in the parametric analysis. These results indicate that the problem of phenocopies cannot always be avoided by simply applying standard LOD –1 confidence intervals; the proportion of confidence intervals that include the true disease locus will not necessarily be the expected 97%. The coverage probability of the confidence interval will vary depending on the distribution of phenocopies in the

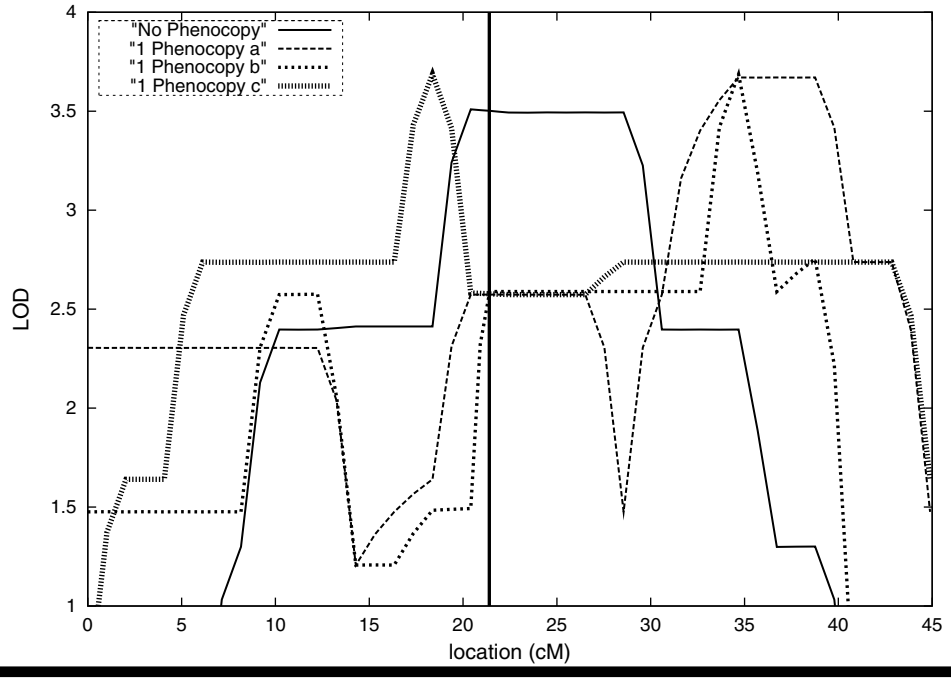

Figure 3

Four simulation replicates.

Three simulation replicates (broken lines) displaying the effect of a phenocopy on the LOD score profile. One replicate (solid line) where there are no phenocopies simulated is also shown.

sample. In larger samples, where the linkage signal comes from a larger number of affected individuals (where single phenocopies are less likely to have a large effect on the location of the LOD peak), we would expect the LOD –1 to become accurate asymptotically. What we have demonstrated here is that there is a real danger if researchers overinterpret the LOD score profiles calculated from small samples.

Discussion

We have shown that the 'identification' of a disease region from a comparison of haplotypes shared IBD between relatives can be severely biased if some of the relatives have the disease phenotype but do not carry the disease causing mutation that is prevalent in the pedigree. In the presence of such phenocopies the length of the MDR may be substantially less than that of the MDRM. This potential problem essentially occurs because complex diseases are treated as if they are Mendelian.

This work was motivated by attempts to identify MDRs from linkage peaks for various complex disorders. For example, Angius et al. (2002) looked at essential hypertension, considering 35 affected individuals. Hypertension is almost certainly caused by multiple loci (Wright et al., 1999) and all affected individuals not carrying a mutation at the main locus (2p24) they reported will be phenocopies with respect to this locus. Angius et al. (2002) were unable to identify a single set of alleles in this region carried by all the affecteds, indicating the existence of at least one phenocopy. The MDR that Anguis et al. (2002) reported may, therefore, not include the actual disease locus as a result of these phenocopies. In a study of bipolar disorder (BP), a single large family affected by BP and recurrent major depression (RMD) generated a LOD of 4.8 on chromosome 4p (Blackwood et al., 1996). Although this gives strong evidence for the relevance of this locus to disease susceptibility it is far from clear whether RMD and BP have the same genetic cause(s). This means that the phenocopy rate of relevance to this region is unlikely to be zero. In an attempt to narrow the disease region indicated by this initial linkage, another three families that also showed evidence of linkage of the disease phenotype to this region of 4p were collected. There was some overlap between the regions identified in the families (Figure 4) but there was no single region implicated by all four families (Porteous et al., 2004). This may indicate that some of the affected individuals considered in these families segregated mutations at loci other than the one of interest on chromosome 4p. That is, some of the affected individuals were phenocopies (with respect to the 4p locus). A false positive linkage finding in one or more of the pedigrees would have the same effect.

Other researchers have encountered similar problems in identifying a single MDR in all affected individuals. Camp et al. (2001) performed linkage analyses on a relatively small schizophrenia data set (less than 50 affecteds) and concluded that, on the basis of traits known to have nonnegligible phenocopy rates, the two regions of interest for further work were 4.3 and 19.75 cM in length. Whilst these are the

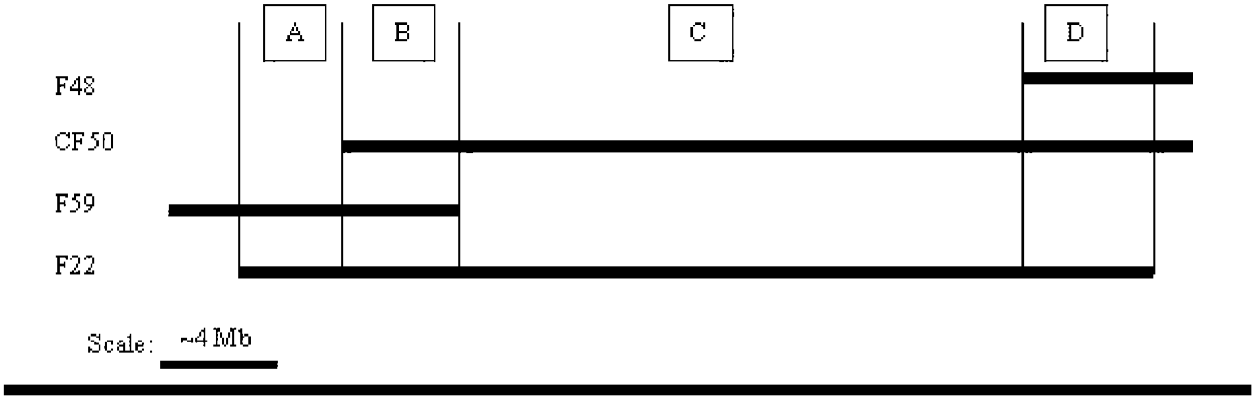

Figure 4

4p regions.

Overlap of disease haplotypes in four families (F48, CF50, F59, F22) affected by BP/RMD.

regions indicated by recombination events in the affected individuals, in the presence of phenocopies, these regions will not necessarily include the disease locus sought out in each case. Further, investigators performing multiple statistical tests (e.g., fitting a dominant model, a recessive model, a model with broad/narrow disease definition) will normally report the smallest possible 'region of interest' without due regard to the number of tests done.

Parametric linkage techniques are fairly robust to miss-specification of parameters such as penetrance (Clerget-Darpoux et al., 1986). This only applies to the detection of linkage, however. The simulations described here show that correctly specifying the phenocopy rate in a parametric analysis will not prevent phenocopies from interfering with disease region identification.

Extensions From Dominant Nuclear Families

The results in Tables 1 and 2 were obtained by assuming that the phenocopies appeared in nuclear families in which there was dominant disease inheritance. However, similar problems will often arise when larger families and recessive types of inheritance patterns are considered. The extension of the above argument to cases other than nuclear families is possible because of two factors:

(i) Singleton affecteds are not ascertained for linkage studies. Hence, a phenocopy will be included in a disease mapping study alongside a number of other closely related affected individuals (whose affection status is at least in part due to them possessing a particular gene). This will mean that the families used will be relatively densely affected and a number of affecteds will likely have some chromosomal regions in common.

(ii) It is common for investigators to remove individuals whose haplotypes are completely distinct from that of the other affecteds (i.e., phenocopies who share none of the MDR).

It is argued that because of the ascertainment procedure and the discarding of incongruous phenocopies (i and ii, above), nuclear families with phenocopies often provide a good approximation to the situation where more general extended pedigrees are analyzed.

Extension to Larger Families (Dominant Inheritance)

Consider extending a nuclear family through the offspring. There are three ways in which the grandchildren of the original founders can be phenocopies. First, these grandchildren may be the offspring of an affected parent and be phenocopies (Figure 5, case 1). In this case they may still inherit a section of disease haplotype via recombination (this is the same situation as in the original nuclear family: unaffecteds and phenocopies can inherit regions of the genome near the disease locus by recombination). Second, there may be phenocopies who are the children of an unaffected individual (Figure 5, case 2). This unaffected individual may possess regions of the genome near the disease locus (via recombination) and will pass this chromosome on to its offspring 50% of the time (any further recombination in the meioses forming the phenocopy will still result in the phenocopy getting at least some of the disease haplotype). The other 50% of the time individual 4 in the pedigree will pass on a chromosomal region unrelated to any of the other affecteds. In this case the phenocopy will often be removed from the group of affecteds since it does not share the MDR with them (issue 2). Third, the grandchildren may be the children of a phenocopy (Figure 5, case 3). This case is the same as the case where the grandchildren's parent is unaffected but, unless the phenocopy rate is rather high, it is unlikely that two such phenocopies will occur.

The family may be further extended to consider great-grandchildren. However, if a branch of the pedigree stems from a second generation individual who is unaffected and who has no affected offspring then the fourth generation is less likely to have been considered for inclusion in the study. In the unlikely event of one being included it will often be excluded because of

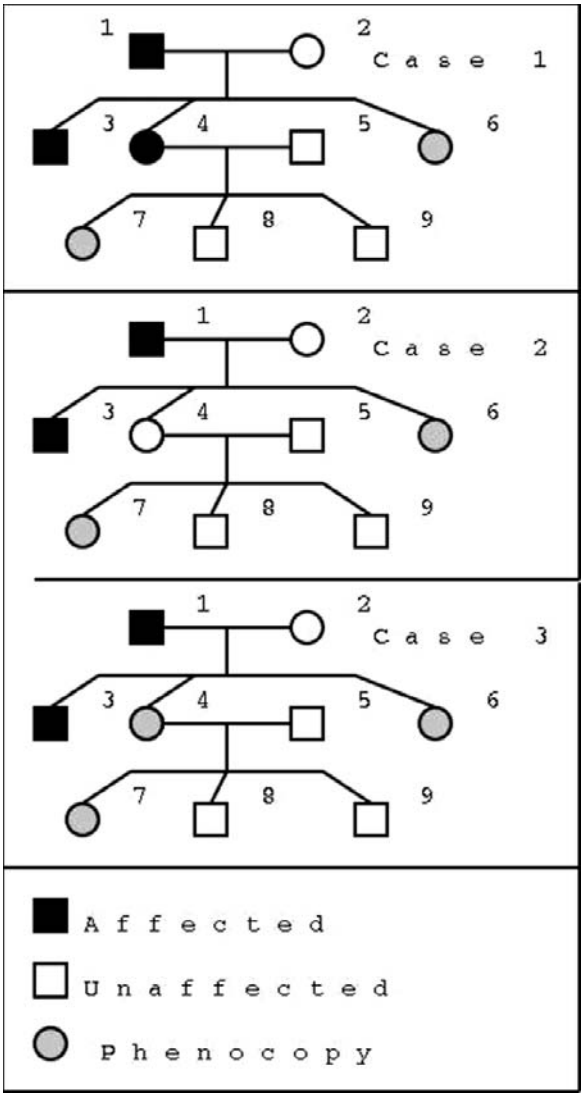

Figure 5
Extensions of nuclear families: dominant case.

issue 2, above. Branches with many affected individuals are much more likely to be included (issue 1). Any phenocopies arising in such a branch will hence share much of their genome with the true affecteds. In summary, in many cases the problems caused by phenocopies in nuclear families will also occur in extended families.

Extension to Recessive Cases

If a disease gene that is recessive in its effect upon the disease is considered, a MDR can be identified where affecteds share two copies of a particular haplotype. In the case of a nuclear family, phenocopy offspring will cause problems similar to those in the dominant case (any recombinations in the transmission of alleles from unaffected carrier parents may falsely narrow the MDR if there are phenocopy offspring). Phenocopy parents will rule out part of the MDR (including the disease locus) obtained from the other affecteds (since they have at most one copy of the disease gene) but, in practice they will usually be removed from the group of affecteds.

Unlike the dominant case, recessive type families are less likely to extend beyond the offspring generation. Clearly, in the dominant case, the disease will often be transmitted over multiple generations. In the recessive case, a new disease allele must be introduced for the disease to be transferred over more than one generation.

In the analyses of quantitative traits uncertainty in the position of the trait locus is dealt with by constructing appropriate confidence intervals (Atwood & Heard-Costa, 2003; Hsueh et al., 2001; Visscher & Goddard, 2004). However, in the analyses of discrete complex traits some investigators are wont to forget that the affected individuals do not all necessarily carry the mutation of interest, resulting in the reporting of untenably small chromosomal regions as the MDR. Confidence intervals in discrete trait linkage analyses have been considered (Roberts et al., 1999) but, in the presence of phenocopies these may exclude the true disease locus. In the simulations presented here, utilizing a LOD –1 confidence interval from nonparametric linkage analysis was often insufficient to guard against the problem of phenocopies. Although such confidence intervals will become more appropriate in larger samples, in small/moderate samples there may be danger when researchers overinterpret the LOD score profile (or equivalently the calculated MDR). Researchers should not simply report the smallest region of allele sharing they find in their samples.

Acknowledgments

We acknowledge the financial support of Organon, the Biotechnology and Biological Sciences Research Council and the Royal Society. We thank Andrew Carothers and Peter Holmans for useful comments.

References

Angius, A., Petretto, E., Maestrale, G. B., Forabosco, P., Casu, G., Piras, D., Fanciulli, M., Falchi, M., Melis, P. M., Palermo, M., & Pirastu, M. (2002). A new essential hypertension susceptibility locus on chromosome 2p24-p25, detected by genomewide search. *American Journal of Human Genetics, 71,* 893–905.

Atwood, L. D., & Heard-Costa, N. L. (2003). Limits of fine-mapping a quantitative trait. *Genetic Epidemiology, 24,* 99–106.

Blackwood, D. H. R., He, L., Morris, S. W., McLean, A., Whitton, C., Thomson, M., Walker, M. T., Woodburn, K., Sharp, C. M., Wright, A. F., Shibasaki, Y., StClair, D. M., Porteous, D. J., & Muir, W. J. (1996). A locus for bipolar affective disorder on chromosome 4p. *Nature Genetics, 12,* 427–430.

Camp, N. J., Neuhausen, S. L., Tiobech, J., Polloi, A., Coon, H., & Myles-Worsley, M. (2001). Genomewide multipoint linkage analysis of seven extended Palauan

pedigrees with schizophrenia, by a Markov-chain Monte Carlo method. *American Journal of Human Genetics, 69,* 1278–1289.

Clerget-Darpoux, F., Bonaiti-Pellie, C., & Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics, 42,* 393–399.

Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., & Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics, 25,* 12–13.

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distance between the loci of linked factors. *Genetics, 8,* 299–309.

Hanson, W. D. (1959). Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. *Genetics, 44,* 833–837.

Hsueh, W. C., Goring, H. H. H., Blangero, J., & Mitchell, B. D. (2001). Replication of linkage to quantitative trait loci: Variation in location and magnitude of the lod score. *Genetic Epidemiology, 21,* S473–S478.

Mangin, B., Goffinet, B., & Rebai, A. (1994). Constructing confidence-intervals for qtl location. *Genetics, 138,* 1301–1308.

Ott, J. (1991). *Analysis of human genetic linkage*. Baltimore, MD: Johns Hopkins University Press.

Porteous, D., Evans, K., Millar, J., Pickard, B., Thomson, P., James, R., Macgregor, S., Wray, N. R., Visscher, P. M, Muir, W. J., & Blackwood, D. H. (2004). Genetics of schizophrenia and bipolar affective disorder: Strategies to identify candidate genes. In *Symposium 68: The Genome of Homo sapiens* (pp. 383–394). Woodbury, NY: Cold Spring Harbor Press.

Roberts, S. B., MacLean, C. J., Neale, M. C., Eaves, L. J., & Kendler, K. S. (1999). Replication of linkage studies of complex traits: An examination of variation in location estimates. *American Journal of Human Genetics, 65,* 876–884.

Sham, P. (1998). *Statistics in human genetics*. London: Arnold.

Visscher, P. M., & Goddard, M. E. (2004). Prediction of the confidence interval of quantitative trait loci location. *Behavior Genetics, 34,* 477–482.

Wright, A. F., Carothers, A. D., & Pirastu, M. (1999). Population choice in mapping genes for complex diseases. *Nature Genetics, 23,* 397–404.

Appendix

Equation 1 is not strictly correct. When there are two or more phenocopies in a sample it is possible for more than one to have a recombination in the MDRM. These recombinations may indicate different regions of the MDRM and hence together rule out the whole region. The probability of this happening can be incorporated into equation 1 above. Equation 1 needs to altered to include $.5^{k-1}$ (where k is the number of phenocopy haplotypes that have recombination events in the MDRM). This takes into account when there are two or more phenocopies with recombination events on the same side. The full equation for the probability of a phenocopy making the MDR too small is therefore

$$\begin{aligned} &Pr(\text{phenoc. has rec. in MDRM}) + Pr(\text{2 phenocs. have rec. in MDRM}) \times 0.5 + \\ &\qquad Pr(\text{3 phenocs. have rec. in MDRM}) \times 0.5^2 + \ldots \\ &= \sum_{r=1}^{m} \binom{r}{1} L^1(1-L)^{r-1} Pr\,(\text{r phenocs. in sample}) + \\ &\sum_{r=2}^{m} 0.5 \binom{r}{2} L^2(1-L)^{r-2} Pr\,(\text{r phenocs. in sample}) + \\ &\sum_{r=3}^{m} 0.5^2 \binom{r}{3} L^3(1-L)^{r-3} \mathrm{Pr}\,(\text{r phenocs. in sample}) + \ldots \\ &= \sum_{k=1}^{m}\sum_{r=1}^{m} 0.5^{k-1} \binom{\mathrm{r}}{k} L^k(1-L)^{r-k} \binom{m}{r} \mathrm{p}^r(1-p)^{r-m} \; for\; r \geq k \end{aligned} \qquad [2]$$

where m is the number of individuals not carrying the mutation of primary interest, r is the number of phenocopy haplotypes, k is as above (for both r and k a dominant disease model is assumed so the number of haplotypes equals the number of individuals), L is the MDRM length and p is the phenocopy rate. Phenoc. is short for phenocopy, rec. is short for recombination. With 20 affecteds carrying the mutation of interest, 100 individuals not carrying the mutation and a phenocopy rate of .01, equations 1 and 2 give .095 and .093 respectively. With a phenocopy rate of .05, the difference is more substantial (.346 cf. .393).