Genetic Epidemiology 20:409–414 (2001)

True and False Positive Peaks in Genomewide Scans: The Long and the Short of It

Peter Visscher[1]***** **and Chris Haley**[2]

[1] *Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh, United Kingdom*
[2] *Division of Genetics and Biometry, Roslin Institute (Edinburgh), Roslin, United Kingdom*

When performing a genome scan in linkage or linkage disequilibrium studies to detect loci underlying complex or quantitative traits, it is important to attempt to distinguish between true and false positives using the appropriate statistical methods. There has been some controversy in the literature regarding the use of the length of a positive peak, i.e., the length of a chromosome region displaying identity-by-descent in linkage studies among affected individuals or the length of a continuous chromosome region for which the test statistic is above a certain threshold. We show in this study, by reasoning and by simulation studies, that conditional on the strength of evidence for a locus affecting a trait of interest, i.e., conditional on the peak height of a test statistic, there is no information in the length of the peak. Our finding has implications for linkage and association studies. Genet. Epidemiol. 20:409–414, 2001. © 2001 Wiley-Liss, Inc.

Key words: genome scan; false positive; linkage; association

INTRODUCTION

A genomewide scan using highly informative marker loci can locate regions of the genome-containing genes contributing to disease susceptibility or to variation in quantitative traits. Results of such studies are often reported as a plot of an appropriate test statistic against chromosomal position. A peak produced where the test statistic exceeds a predetermined significance threshold may be a true positive effect,

*Correspondence to: Peter Visscher, University of Edinburgh, Institute of Cell, Animal and Population Biology, West Mains Road, Edinburgh EH9 3JT, UK. E-mail: peter.visscher@ed.ac.uk

Received 19 June 2000; Accepted 8 January 2001

caused by one or more loci that influence the trait, or it may be a false positive, resulting from random fluctuation. A recent study by Terwilliger et al. [1997] suggests that information to distinguish "true" peaks from "false" peaks of a similar height is contained in the length of the peak (the length of the chromosome over which the test statistic remains above the threshold). These findings were supported by an analytical analysis of Knapp [1998].

The aim of this study is to show, using reasoning and a simple simulation study, support for the contention of Lander and Kruglyak [1995] that it is not possible to distinguish between true and false peaks of similar height.

METHODS

We suggest that the flaw in the argument of Terwilliger et al. [1997] arises out of the problem of length-biased sampling that they illustrate with the "paradox of buses" described by Feller [1971]. Put simply, if the interval between buses varies, then arriving at a bus stop at an arbitrary time, a passenger is more likely to sample a long interval than a short interval. The intervals sampled in this way give a biased sample of all intervals and an overestimate of the mean interval between buses. To understand this, imagine the extreme but realistic situation where the interval is either zero, i.e., the buses all arrive together, or some long period. Then arriving at an arbitrary time, a passenger will always sample the long interval and get a very biased view of the average interval between buses. Applied to the problem of the genomewide scan, the length-biased sampling argument suggests that if we focus on any specific point, *whether or not it contains a gene affecting the trait*, we are more likely to sample long peaks than short peaks. In other words, a longer interval is more likely to overlap with any fixed genomic position, so that longer intervals are more likely to contain a disease gene just because they contain more genes. Consider an extreme case where the peak covers the entire chromosome. Then, if there is a gene on that chromosome that affects the trait of interest, the peak would always contain the gene. If this effect of "longer peaks contain more genes" is proportional to the length, i.e., on average a peak that is twice as long has a twofold increased probability of containing a gene, then there is no utility in using both the length and height of a peak in mapping a gene.

We suggest that the flaw in the arguments and simulations presented by Terwilliger et al. [1997] is that they compare the lengths of peaks at fixed positions, defined by the presence of a gene, with the average length of all peaks. We show by simulation that 1) true peaks are longer than false peaks when we focus on a fixed point even if that locus is not linked to a trait gene, and 2) the number of mapped trait genes per cM is independent of peak length, i.e., longer peaks are more likely to contain a trait gene because they span a larger proportion of the genome.

To demonstrate that the phenomenon of length biased sampling applies equally to true and false peaks, we performed the following simple simulation. We assume a basic mapping experiment and generate a backcross population of 100 individuals from a cross between two inbred lines. We focus on a single 100-cM chromosome that is genotyped for fully informative markers at 1-cM intervals. We consider a normally distributed quantitative trait with a residual standard deviation of unity, with individuals being assigned a random value from this distribution. We generate three sets of simulations, all with a quantitative trait locus, QTL, located at the chromosome midpoint (50 cM). In the first set of simulations, the QTL explains 0% of the trait variation (i.e., it has no effect). In the second and third sets of simulations, the QTL explains 5 and 10%, respectively, of the phenotypic variance in the backcross. Inheritance of all loci was determined assuming random assortment and that recombination events occurred independently, allowing use of Haldane's [1919] mapping function. The phenotype of an individual is composed of its random residual component plus any genetic effect. For each marker, we compared the means of individuals in the two possible marker classes in the backcross by analysis of variance, and used the F value as our test statistic. We chose as our significance threshold the nominal 5% significance level ($F_{1,98} = 3.94$). A significant peak is represented by F values at one or more adjacent markers greater than 3.94. Peak height is defined as the highest F value within a peak and peak length as *n* cM, where *n* is the number of adjacent significant markers within a peak.

RESULTS

We generated 100,000 replicates for each set of simulations, with the QTL having 0, 5, and 10% heritability. For each set of simulations, we recorded the height and length of all significant peaks on a chromosome. There were 77,250, 220,254, and 243,481 significant peaks for the sets of simulations with the QTL having 0, 5, and 10% heritability, respectively. These peaks were classified as "true" if they overlap the fixed 50-cM point on the chromosome at which the QTL was simulated, or "false" if they cover a random point on the chromosome not including the 50-cM point. Note, however, that the "true" peaks for the set of simulations with 0% heritability QTL are false in the sense that no QTL effect is simulated on that chromosome. The data were summarized in a similar manner to that used by Terwilliger et al. [1997], by grouping peaks into height classes based on their F value (i.e., $3.84 \leq$ Class $1 < 10$, $10 \leq$ Class $2 < 15$, $15 \leq$ Class $3 < 20$, etc.). For each class, the mean of the peak F values and the mean length of the peak were calculated.

The mean height and length of the peaks covering a fixed or random point on the chromosome are plotted in Figure 1 for each class where there were 10 or more observations. Figure 1 clearly shows that the relationship between the length and height of peaks covering a fixed point is not affected by the size of effect of a QTL. The same relationship between height and length of peaks is observed whether the QTL explains 0 or 10% of the variation in the trait. However, there is a major difference in the relationship between peak height and length according to peaks covering a fixed or random point. The difference in these relationships is thus due to whether peaks were selected because they encompass a fixed point (i.e., the 50-cM point on the chromosome), rather than the effect of the gene.

We observe the same relationship between peak height and length no matter what the heritability of the QTL on the chromosome. This tells us that once height has been taken into account, there is no information in length of the peak that would allow us to say whether a peak came from a chromosome with a 0% heritability QTL (i.e., no QTL) or one with a 5 or 10% heritability QTL. In other words, we cannot use peak length to help distinguish between chromosomes with true or false peaks. Note that the observed difference in length between the peaks covering a fixed point

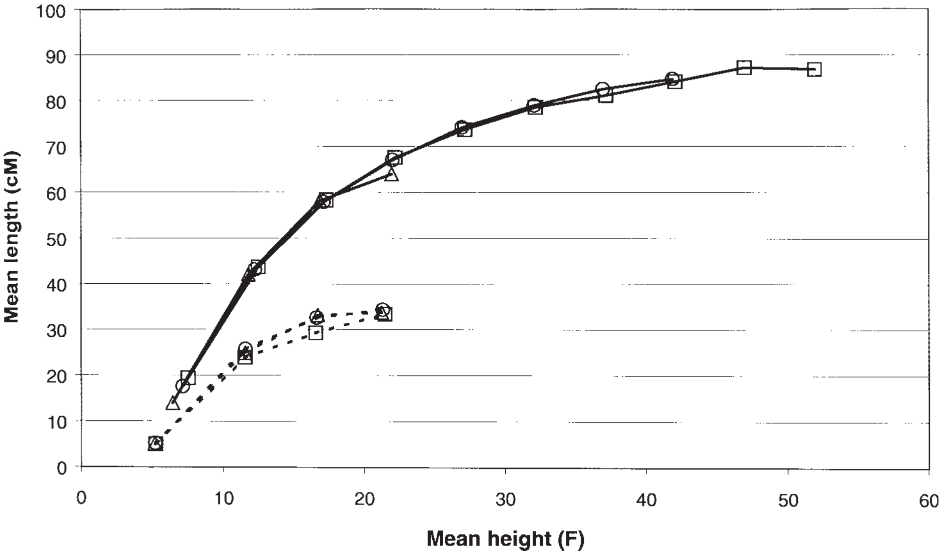

Fig. 1. Mean height and length of "true" and "false" peaks. Solid lines represent "true" peaks (i.e., they include the fixed 50-cM position on the chromosome), broken lines represent "false" peaks (not including the 50-cM position). Symbols represent the heritability of the QTL: triangle, 0%; circle, 5%; square, 10%.

and peaks covering random points is twofold, as predicted by the bus-waiting paradox with Poisson arrival times [Terwilliger et al., 1997].

In a further simulation study, we show that the number of detected QTLs per cM is independent of the length of the interval, given peak height. A long chromosome of 500 cM was simulated, with a single QTL at location 125 cM explaining 10% of the total variance in a backcross population of 100 individuals. For each of 1,000 replicates, the height and length of all significant peaks were recorded, and whether they were true (i.e., encompassed the QTL at position 125 cM) or false peaks. The variable of interest was the number of QTL (0 or 1) per peak divided by the peak length, i.e., the number of QTLs per cM. For a false positive, these values are 0 and for true positives values they are 1/length. A multiple regression of the number of QTLs/cM on the height and length of the peak revealed a significant P value for height (P < 0.0005) but no significant values for the length (P < 0.395). The same results were obtained when the regression was on height, height^2, length and length^2; both the linear and quadratic terms for height were significant (P < 0.0005 for both) whereas both terms for length were not significant at the 1% level (P < 0.683 for length and P < 0.04 for length^2). These results confirm that the number of QTLs per cM is independent of peak length, given peak height. In addition to these analyses, we applied logistic regression using data from all peaks and regressed the presence or absence of a true peak on log(height) and log(length), assuming a binomial distribution of errors and using a log link function. According to our hypothesis, for a given peak height the probability of covering a QTL is proportional to the amount of the genome which is covered by the peak. Hence, given height, we predict a coefficient of 1.0 of the regression of log[Prob(true peak)] on log(peak length). We obtained an estimate of this regression coefficient of 1.00 (SE = 0.06), and this estimate is consistent with a residual effect of peak length, which is proportional to the proportion of the genome covered by the peak.

DISCUSSION

We have argued that there is no value in the length of a peak, conditional on peak height, for distinguishing between true and false peaks. Our simulation results show no evidence that the lengths of true peaks tend to be greater than those of false peaks of similar height. The simulation models and analyses we used were greatly simplified compared to those of Terwilliger et al. [1997], but contained the same basic elements and we would not expect the conclusions to change if the models were further elaborated. Our results thus suggest that the length differences observed by Terwilliger et al. [1997] are due to length biased sampling and are a function of the comparison of peaks spanning a fixed point with all peaks, and a function of the trivial effect that longer peaks contain proportionally more genes, rather than a function of the comparison of true and false peaks. Our results provide no evidence that peak length can be used to help distinguish true from false peaks. Terwilliger and Weiss [1998] appeared to reach a similar conclusion on the utility of Terwilliger et al. [1997], at least in the context of the lengths of general shared IBD segments between distant relatives, where the same length-biased sampling principle holds, by stating that “...and it is shown that the length of a haplotype is not so useful a measure of significance as the haplotype frequency in the population, as 25% of the false positives (assuming there are false positive IBD segments shared) will be longer than the true positives.”

Knapp [1998] used a simplified study design of the number of alleles from a single parent shared IBD of an affected sib pairs, and used theoretical derivation to support the conclusion of Terwilliger et al. [1997] that true peaks are expected to be longer than false-positives. Knapp [1998] used a “genome” of two markers (M1 and M2) and considers the case when M1 is the disease locus or when the disease locus is unlinked to both M1 and M2. The case that M2 is the disease locus is not considered, hence this is another example of focusing on a single point. The conclusions reached by Knapp [1998] can also be expressed as “A peak is more likely to be a true peak if it contains M1 (the disease locus).” This is not surprising, and is another example of longer peaks being more likely to contain a disease locus because they include more genes. Note that Knapp’s conclusion that true peaks are expected to be longer than false peaks holds in case where M1 and M2 are unlinked, which demonstrates that it is the focus on the disease locus that forces this conclusion and not that there is intrinsic value in longer peaks.

Goldin and Chase [1997], Goldin et al. [1999], and Hoh and Ott [2000] appear to use the results and conclusions of Terwilliger et al. [1997] to show an improvement in the power of linkage detection by taking scans or averages of *P* values at consecutive markers. However, we would contend that these studies simply demonstrate that power can be increased by combining information from several closely linked markers that are partially informative, rather than by treating them separately, i.e., it is better to do a multi-point analysis than a single marker analysis. Thus, these studies do not support the notion that for a given height of a multi-point test statistic, longer peaks contain more disease genes.

Our results apply also to other methods in which the outcome of a study contains both a peak height, such as a maximum test statistic, and a peak length. For example, in haplotype sharing between affected individuals, one could calculate the probability of excess haplotype sharing as the test statistic, and the length of segment shared IBD. The same principle should apply here, that there is no additional information in the length of the IBD segment for a given probability of haplotype sharing between individuals that share a trait or disorder.

ACKNOWLEDGMENTS

We thank Ricardo Pong-Wong for bringing this problem to our attention and for the biscuits at the Edinburgh Animal Geneticists Discussion Group and to group members, especially Steve Bishop, Bill Hill, Ricardo Pong-Wong, Dave Waddington, and Ian White, for useful comments. We thank Mike Goddard for helpful discussions and for suggesting alternative analyses and wording, and Nik Schork and Chris Amos for encouragement.

REFERENCES

Feller W. 1971. Introduction to probability theory and its applications, volume 2, 2nd ed. New York: John Wiley & Sons.

Goldin LR, Chase GA. 1997. Improvement of the power to detect complex disease genes by regional inference procedures. Genet Epidemiol 14:785–9.

Goldin LR, Chae RA, Wilson AF. 1999. Regional inference with averaged P values increases the power to detect linkage. Genet Epidemiol 17:157–64.

Haldane JBS. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309.

Hoh J, Ott J. 2000. Scan statistics to scan markers for susceptibility genes. Proc Natl Acad Sci U S A 97:9615–7.

Knapp M. 1998. Discriminating between true and false-positive peaks in a genomewide linkage scan, by use of the peak length. Am J Hum Genet 62:1561–2.

Lander ES, Kruglyak L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–7.

Terwilliger JD, Weiss KM. 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Op Biotech 9:578–94.

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE. 1997. True and false positive peaks in genomewide scans: Applications of length-biased sampling to linkage mapping. Am J Hum Genet 61:430–8.