Letters to the Editor

Am. J. Hum. Genet. 62:994–996, 1998

Linkage Thresholds for Two-stage Genome Scans

To the Editor:

Two-stage genome scans are a common design in linkage studies. In the first stage, genotypes are generated for a standard set of ~300 microsatellite markers, and linkage analysis is carried out. In the second stage, any interesting "regions" (or "hits") that were found during the first stage are followed with a much denser map of markers to extract all available linkage information.

How should the statistical significance of the final results be determined? We have argued elsewhere that thresholds computed for a dense map of markers covering the entire genome should be applied (Lander and Kruglyak 1995). The logic behind this recommendation is simple. False-positive peaks that would have been detected with a dense whole-genome map almost always come up as hits in the initial stage of a two-stage scan. The reason for this is that the marker density for the initial stage is specifically chosen to detect peaks of moderate size (whether real or false positive). Since these hits are then followed up with a dense map, the final results are virtually the same as if a dense whole-genome map had been used at the outset.

Recently, Sawcer et al. (1997) challenged this recommendation. They carried out simulations to determine the significance of results obtained in a genome scan for multiple sclerosis (MS), and they concluded that "the practice of adding markers around provisional linkage hits in a genome screen has relatively little effect on the false-positive rate" (Sawcer et al. 1997, p. 227). The goal of this letter is to point out the fallacy in the study of Sawcer et al. (1997) and to demonstrate unequivocally that increasing map density around linkage hits has a strong effect on the false-positive rate, making it essentially the same as that expected with a dense whole-genome map.

In the MS genome scan, Sawcer et al. (1996) identified provisional hits in the initial screen and then followed them up by increasing the marker density in the surrounding regions "to increase the information extraction in those areas showing possible linkage" (Sawcer et al. 1997, p. 224). Sawcer et al. (1996, 1997) carried out simulations (with marker genotypes generated under the hypothesis of no linkage) to evaluate the significance of the genome scan findings. However, the simulations had a serious flaw. Instead of using increased marker density around the hits occurring in each simulated genome scan (which would accurately model the actual follow-up strategy), they used increased marker density around the hits obtained in the actual MS genome scan, despite the fact that these regions did not correspond to hits in most simulated genome scans. Thus, the simulations modeled a strategy of increasing map density in arbitrary locations rather than around provisional hits. Not surprisingly, Sawcer et al. (1997) observed that the locations of hits in simulated genome scans showed little correlation with the regions of higher map density. The simulations did not accurately model the experiment and, as a consequence, considerably underestimated the false-positive rate.

To illustrate this effect, we carried out simulations modeling the different approaches. We generated 1,000 replicates of a genome scan of 100 sib pairs, under the null hypothesis of no linkage, with each replicate containing genotype data on 23 chromosomes of length 150 cM each. Markers were assumed to have four equally frequent alleles (heterozygosity 0.75). For each replicate, we examined four scenarios of marker density: sparse, dense, follow-up of hits (FH), and follow-up of arbitrary regions (FA). In the sparse scenario, markers were spaced every 10 cM. In the dense scenario, markers were spaced every 1 cM. In the FH scenario, markers were spaced every 10 cM, as in the sparse scenario, except that regions in which the score from the sparse map exceeded a given threshold were saturated with markers, at 1-cM density, in a 10-cM window around each peak. This scenario was intended to model an actual two-stage study. Finally, in the FA scenario, markers were spaced every 10 cM, as in the sparse scenario, except that the marker density was increased to 1 cM in arbitrarily selected 10-cM regions, with the number of such regions constrained to equal the number of regions followed up in the FH scenario. This scenario was intended to model the incorrect simulation approach of Sawcer et al. (1997). Linkage analysis was carried out by using MAPMAKER/SIBS (Kruglyak and Lander 1995) to compute

the maximum LOD score (MLS) statistic of Holmans (1993). The threshold for follow-up was an MLS of 1.0. On average, 5.7 such regions were followed up per genome scan; higher or lower thresholds for follow-up did not change the results substantially.

The results are shown in figure 1, which plots the expected number of false positives in a genome scan for the four scenarios. As expected, the false-positive rate for follow-up of arbitrary regions closely follows that for the sparse map, while the rate for follow-up of hits closely follows that for the dense map. The respective thresholds for suggestive linkage (one expected false positive per genome scan) are ~1.9 for the sparse and FA scenarios and ~2.3 for the dense and FH scenarios. The thresholds for significant linkage (one expected false positive per 20 genome scans) are ~3.4 for the sparse and FA scenarios and ~3.8 for the dense and FH scenarios. The use of simulations with follow-up of arbitrary regions underestimates the true false-positive rate by a factor of ~2. The thresholds for follow-up of hits are not very different from the theoretical dense-map values of 2.6 for suggestive and 4.0 for significant linkage (Lander and Kruglyak 1995).

We therefore conclude that the original recommendation to use dense-map thresholds to assess the significance of results from two-stage genome scans is appropriate. Of course, it may still be desirable to carry out simulations, to take into account specific features of a particular study and to avoid relying on asymptotic assumptions. In this case, such simulations must accurately model the methodology of the study, including follow-up of interesting regions. Otherwise, the false-positive rate can be underestimated considerably.

LEONID KRUGLYAK AND MARK J. DALY
Whitehead Institute for Biomedical Research
Cambridge, Massachusetts

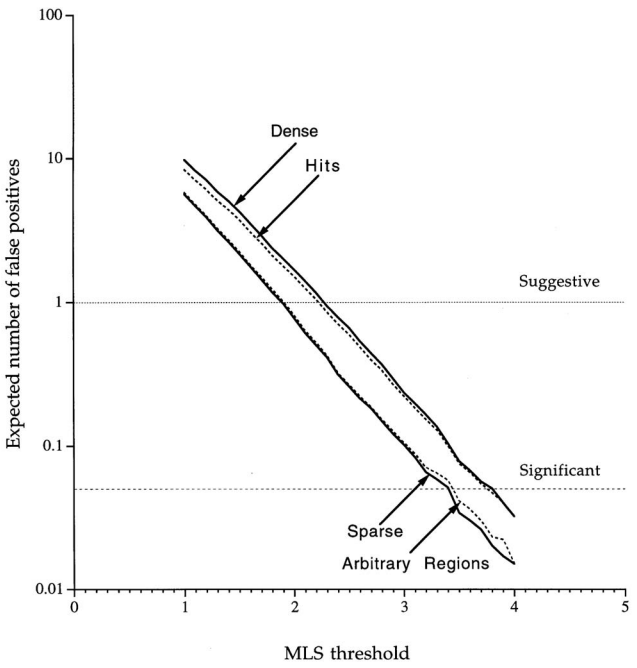

Figure 1 The expected number of false positives at or above a given MLS threshold for the four scenarios described in the text, plotted on a log scale.

References

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Kruglyak L, Lander ES (1995) Complete multipoint sib pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Sawcer S, Jones HB, Feakes R, Gray J, Smaldon N, Chataway J, Robertson N et al. (1996) A genome screen in multiple sclerosis reveals susceptibility loci on chromosomes 6p21 and 17q22. Nat Genet 13:464–468

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN, Clayton D (1997) Empirical genomewide significance levels established by whole genome simulations. Genet Epidemiol 14:223–229

Address for correspondence and reprints: Dr. Leonid Kruglyak, Whitehead Institute for Biomedical Research, 1 Kendall Square, Building 300, Cambridge, MA 02139. E-mail: leonid@genome.wi.mit.edu

0002-9297/98/6204-0038$02.00