Published in final edited form as:

Genet Epidemiol. 2001 ; 21(Suppl 1): S800–S804.

Novel Selection Criteria for Genome Scans of Complex Traits

Anja Wille and Suzanne M. Leal

Laboratory of Statistical Genetics (A.W., S.M.L.), The Rockefeller University, New York, New York; Institute for Medical Biometry, Informatics, and Epidemiology (A.W.), University of Bonn, Bonn, Germany

Abstract

Due to their oligogenic inheritance, the identification of susceptibility loci for complex traits by classical selection criteria has not been very successful. One way to address this problem is to identify statistics that measure the effect of more than one locus simultaneously. In the approach described here, a p-value is assigned to a combination of loci under the null hypothesis that none of them is linked to the disease locus. In order to examine the power of this method to detect multiple loci, the Genetic Analysis Workshop 12 general population simulated data set was analyzed using variance component methods. Using the described novel selection criteria resulted in an increase of power, however, a rejection of the null hypothesis has to be interpreted with care.

Keywords

bootstrap resampling; quantitative traits; significance criterion; variance components analysis

INTRODUCTION

Gene mapping and identification for complex traits has been disappointing. The majority of successful gene identification for complex traits has been limited to a subset of genes with high penetrance (i.e., BRCA1, BRCA2, presenilin, alpha-synuclein). The problem with mapping and isolating genes for complex traits arises from their oligogenic inheritance and heterogeneity.

In a genome scan, a simultaneous search for more than one susceptibility locus might lead to a combination of single gene effects and facilitate the identification of interesting regions. Under the null hypothesis that there are no susceptibility loci in the whole genome, thresholds corresponding to a genome-wide p-value of 0.05 can be determined for a specified number L of loci [Lander and Kruglyak, 1995]. If L or more regions exceed this threshold in a genome scan, the null hypothesis is rejected. A simultaneous scan for several loci should perform better and detect more regions than the classical approach, especially when the single effects of genes involved in the disease etiology are weak and evenly distributed.

In order to determine the power of this approach and contrast it to conventional methods, variance components analysis was carried out with the SOLAR program version 1.5.7 [Almasy and Blangero, 1998] using the Genetic Analysis Workshop (GAW) 12 simulated data set. Since only a small number of replicates (50) was available, an approach based on bootstrap resampling with replacement was used to increase the number of replicates and to calculate the power to detect MG1 (major gene 1), MG2, and MG3 using Q1 (quantitative trait 1) and Q2.

METHODS

Significance Criteria

Under the null hypothesis of no linkage, variance component methods provide a point-wise lod score S as a test statistic that is asymptotically distributed as a $\frac{1}{2}:\frac{1}{2}$ mixture of a χ_1^2 variable and a point mass at zero, scaled by the factor $\frac{1}{2\ln 10}$,

$$S \sim \frac{1}{2\ln 10} \left(\frac{1}{2} 0 + \frac{1}{2} \chi_1^2 \right)$$

[Self and Liang, 1987]. The number of regions $N(T)$ in which S exceeds a relatively high threshold T in a whole genome scan follows a Poisson distribution with mean

$$\mu(T) = (C + 2\ln 10 \cdot 2GT)\alpha(T),$$

where C is the number of chromosomes, G is the genome length and $\alpha(T)$ is the point-wise significance level corresponding to S [Lander and Kruglyak, 1995]. Hence, the probability that the number $N(T)$ of regions exceeding T is at least 1, 2, or 3 equals

$$P(N(T) \geq 1) = 1 - e^{-\mu(T)},$$

$$P(N(T) \geq 2) = 1 - e^{-\mu(T)} - \mu(T)e^{-\mu(T)},$$

$$P(N(T) \geq 3) = 1 - e^{-\mu(T)} - \mu(T)e^{-\mu(T)} - \frac{\mu(T)^2}{2}e^{-\mu(T)}.$$

Setting all three probabilities to 0.05 and solving these equations for T leads to three critical values, CV1, CV2, and CV3, for $N(T) = 1$, $N(T) = 2$ and $N(T) = 3$, respectively. CV1 is the conventional critical lod score 3.3 [Lander and Kruglyak, 1995], CV2 = 2.4 and CV3 = 2.0.

Under the null hypothesis of no linkage, one or more regions with lod scores larger than 3.3 occur with a genome-wide p-value of 0.05. Accordingly, two or more (three or more) regions with lod scores of 2.4, (2.0) have p-values of 0.05.

Bootstrap Resampling and Analysis

Due to the small number of available replicates (50 replicates each with 23 pedigrees with a total of 1,497 individuals), bootstrap resampling with replacement was used. Here bootstrap

resampling was not used in the classical sense [Efron, 1982] but was used instead to increase the number of available simulated data sets from the 50 provided to 50²³ [Terwilliger and Ott, 1992]. A random number generator was used to select one of the 1...50 replicates for pedigree 1, this process was repeated for pedigree 2 and so on until a replicate was selected for each of the 23 pedigrees used in the analysis. This resampling scheme preserves the kinship correlation among marker alleles and among phenotypes within pedigrees.

The pedigrees in the generated data set were then analyzed using all markers within a 20-cM region of the three MGs: MG1 on chromosome 19, MG2 on chromosome 2, and MG3 on chromosome 9. For comparison purposes, a variety of analysis schemes were used for the power analysis: trait Q1 was analyzed with covariates (1) Q2, EF1, and EF2 (environmental factors); (2) EF1 and EF2; and (3) EF2. Trait Q2 was analyzed including the covariates (1) Q1, EF1, and EF2; (2) EF1 and EF2; and (3) EF2. In addition, for each analysis, age and sex were included as covariates. For the power analysis, b=350 bootstrap resamples were analyzed. The power is simply the proportion of replicates with a lod score greater or equal to a given criterion (e.g., CV1, CV2, or CV3).

It should be noted that for this example bootstrap resampling provides a total of 50²³ unique data sets, however, since bootstrap resampling was performed with replacement, each one of the data sets analyzed was not necessarily unique.

RESULTS

For trait Q1 (Table I) and Q2 (*Table II*) the power was determined using the conventional critical value CV1 = 3.3: to detect each MG separately; at least one of the MGs; and all of the MGs. When the analysis was carried out using Q2 as the dependent variable, the power to detect MG1, MG2, and MG3 was evaluated. When Q1 was the dependent variable, only the power to detect MG1 and MG2 was evaluated, since MG3 did not influence Q1. The covariates age and sex were included in each analysis. For each analysis, three different analysis schemes were undertaken with the following covariates: EF1, EF2, and the complementary quantitative trait (Q2 as a covariate when Q1 was the dependent variable and vice versa); EF1 and EF2; and only EF2.

When analyzing trait Q1, including Q2 as a covariate reduced the power to detect MG1 and MG2 (Table I). In the case where trait Q2 was analyzed and Q1 was included as a covariate the power to detect all MGs except MG3 was decreased (Table II). This finding is not unexpected, since Q1 and Q2 are correlated and both MG1 and MG2 account for the variance in Q1 and Q2.

For trait Q1, the power to detect MG1 or MG2 is 76% (covariates sex, age, and EF2 included in the model). However, the power to detect both MG1 and MG2 is much lower (27%), which is not surprising since in the GAW12 data set MG1 and MG2 account for 24% and 21% of the variance in Q1 and have therefore relatively weak effects. The addition of covariate E1 in the analysis had only a minor impact on the analysis and the power to detect MG1 or MG2. The power remained about the same whether this covariate was removed or included in the analysis (Tables I-IV).

The results for trait Q2 were similar to those of Q1 (Table II). Even if the power to identify at least one region containing a susceptibility locus was high (here above 70%), the probability of finding two or all three loci was small.

In these two examples, when multiple loci are detected, the false positive rate is 5% for each locus. The situation is different in the next two tables. If CV2 or CV3 are used as the selection criteria to detect multiple loci, the false positive rates correspond to a combination of the loci, not to each of the detected regions. For example, if CV2 is used as the selection criterion, then both loci detected may not be true susceptibility loci with a false positive rate of 5%. The rejection of the null hypothesis, however, gives ground to further pursue all of the identified regions.

The power to detect MG1 and MG2 (dependent variable Q1) is more than twice as high using CV2 as the criterion to detect linkage (Table III) as when CV1 is used as the selection criterion (Table I). The power of 18% to detect all three loci for Q2 using the CV3 criterion is still low (Table IV).

DISCUSSION

Genome scans applied to complex traits often fail to detect susceptibility loci. As seen here for the traits Q1 and Q2, the power to detect multiple loci influencing the trait is poor. Tests of linkage, which are not only based on the point-wise magnitude of the lod score but also on the number of loci with a lod score over a given threshold, can facilitate the identification of interesting regions. Our approach is based on a detection of a set of two or three loci jointly. For the general population simulated data set, it was shown that the power to detect a combination of regions containing the two or three disease-causing genes in Q1 or Q2 is higher than the power of the point-wise detection approach using a lod score of 3.3. This result was obtained for all tested analysis schemes.

It should be pointed out that the power calculation figures are not directly comparable, since the rejection of the null hypothesis of no linkage in the genome results in different interpretations for each method. A lod score of 3.3 in several regions in the genome is considered to be significant evidence that each region contains susceptibility loci. In contrast, a detection of two or more loci using CV2 leads to the conclusion that not all of them can be false positives at a significance level of 0.05. However, the emphasis in our power calculations was not only on the rejection of the null hypothesis using CV2 (or CV3), but whether MG1 and MG2 (or MG1, MG2, and MG3) were contained in the set of detected loci and should be confirmed in subsequent studies. Our approach might be considered a preselection method. It is a further application of the previously proposed two-stage procedure, in which markers are selected in stage 1 and jointly analyzed in stage 2 [Hoh et al., 2000].

The assumption of an underlying Poisson distribution for the number of regions in the genome exceeding a certain threshold holds only as long as the threshold is large and the case where a region exceeds the threshold is rare [Lander and Kruglyak, 1995]. Subsequently, this assumption might not hold for the computation of CV4, CV5, CV6, etc.

With the availability of a working draft of the human genome, it is becoming increasingly easier to use a positional candidate approach for gene identification for some phenotypes. Given this new situation, researchers may be willing to follow up regions that have a p-value greater than 0.05 genome-wide. The approach described here offers less stringent criteria for selecting loci to proceed with gene isolation. Whether or not it is acceptable to follow up two more regions when these regions may not contain any genes with a probability of 5% will depend on the resources that must be expended to further investigate these regions on a molecular level.

ACKNOWLEDGMENT

This work was supported by National Institutes of Health Grants DC03594 and MH44292.

REFERENCES

- Almasy L, Blangero J. 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198–211. [PubMed: 9545414]
- Efron B 1982 The jackknife, the bootstrap and other resampling plans. Philadelphia: SIAM.
- Hoh J, Wille A, Zee R, et al. 2000 Selecting SNPs in two-stage analysis of disease association data: A model-free approach. *Ann Hum Genet* 64:413–7. [PubMed: 11281279]
- Lander E, Kruglyak L. 1995 Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat Genet* 11:241–7. [PubMed: 7581446]
- Self SG, Liang K-Y. 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* 82:605–10.
- Terwilliger JD, Ott J. 1992 A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. *Cytogenet Cell Genet* 59:142–4. [PubMed: 1737483]

TABLE I.

Power to Detect MG1 and MG2 Using Q1 With CV1 Criterion

Detection of	Q2, EF1 & EF2 ^a	EF1 & EF2 ^a	EF2 ^a
Only MG1	0.10	0.51	0.51
Only MG2	0.05	0.52	0.51
MG1 or MG2	0.14	0.76	0.76
MG1 and MG2	0.00	0.27	0.26

^aCovariates used in the analysis besides sex and age.

TABLE II.

Power to detect MG1, MG2, and MG3 Using Q2 With CV1 Criterion

Detection of	Q1, EF1 & EF2 ^a	EF1 & EF2 ^a	EF2 ^a
Only MG1	0.00	0.05	0.04
Only MG2	0.17	0.60	0.59
Only MG3	0.70	0.40	0.40
One or more	0.74	0.76	0.76
Two or more	0.11	0.26	0.24
MG1, MG2 & MG3	0.01	0.01	0.01

^aCovariates used in the analysis besides sex and age.

TABLE III.

Power to Detect MG1 and MG2 Using Q1 with CV2 Criterion

Detection of	Q2, EF1 & EF2 ^a	EF1 & EF2 ^a	EF2 ^a
MG1 and MG2	0.05	0.60	0.59

^aCovariates used in the analysis besides sex and age.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE IV.Power to Detect MG1, MG2, and MG3 Using Q2 With CV2⁺ and CV3⁺⁺ Criterion

Detection of	Q1, EF1 & EF2 ^a	EF1 & EF2 ^a	EF2 ^a
Two or more MGs ⁺	0.28	0.58	0.55
MG1, MG2 & MG3 ⁺⁺	0.01	0.18	0.18

^aCovariates used in the analysis besides sex and age.