Ann. Hum. Genet. (2001), **65**, 313–329
Printed in Great Britain

LOD score exclusion analyses for candidate genes using random population samples

H.-W. DENG[1,2,3], J. LI[1,2], AND R. R. RECKER

[1] *Osteoporosis Research Center and* [2] *Dept. of Biomedical Sciences, Creighton University, 601 N. 30th St., Suite 6787, Omaha, NE 68131, USA*

[3] *Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, ChangSha, Hunan 410081, P.R. China*

SUMMARY

While extensive analyses have been conducted to test *for*, no formal analyses have been conducted to test *against*, the importance of candidate genes with random population samples. We develop a LOD score approach for exclusion analyses of candidate genes with random population samples. Under this approach, specific genetic effects and inheritance models at candidate genes can be analysed and if a LOD score is $\leqslant -2.0$, the locus can be excluded from having an effect larger than that specified. Computer simulations show that, with sample sizes often employed in association studies, this approach has high power to exclude a gene from having moderate genetic effects. In contrast to regular association analyses, population admixture will not affect the robustness of our analyses; in fact, it renders our analyses more conservative and thus any significant exclusion result is robust. Our exclusion analysis complements association analysis for candidate genes in random population samples and is parallel to the exclusion mapping analyses that may be conducted in linkage analyses with pedigrees or relative pairs. The usefulness of the approach is demonstrated by an application to test the importance of vitamin D receptor and estrogen receptor genes underlying the differential risk to osteoporotic fractures.

INTRODUCTION

Association studies that depend on linkage disequilibrium between markers and genes underlying complex traits have helped to decipher some genetic bases of differential susceptibility to complex diseases (e.g. Chagnon *et al.* 1998). In association studies, usually, case-control analyses have been employed by comparing genotype or allele frequencies of candidate genes in unrelated cases and controls (e.g. Blum *et al.* 1990, 1991; Holden, 1994). However, despite extensive efforts, the results of independent association studies often fail to reach consensus and result in controversy. Such examples are the association between the dopamine D2 receptor gene and alcoholism (Blum *et al.* 1990, 1991; Holden, 1994; Gelernter *et al.* 1993; Pato *et al.* 1993) and the association between vitamin D receptor (*VDR*) genotypes and osteoporosis (Eisman, 1995; Peacock, 1995; Gong *et al.* 1999).

One of the most important causes, that may result in the inconsistent results from association studies, is population admixture (Chakraborty & Smouse, 1988; Lander & Schork, 1994; Weir, 1996; Deng & Chen, 2000; Deng *et al.* 2001). Population admixture causes deviation from Hardy–Weinberg (HW) equilibrium and may also yield linkage disequilibrium (and thus association) between a marker

Correspondence: Hong-Wen Deng, Ph.D., Osteoporosis Research Center, Creighton University, 601 N. 30th St., Suite 6787, Omaha, NE 68131, USA. Tel: 402–280–5911; Fax: 402–280–5034.
E-mail: deng@creighton.edu

locus and a disease susceptibility locus (DSL) (Weir, 1996). Family based analyses such as the transmission disequilibrium test (TDT, Spielman *et al.* 1993) have been developed specifically to control for population admixture in association studies. However, compared with the case-control studies that employ random samples of cases and controls, the samples for the family based studies such as the TDT are generally more difficult to obtain. Therefore, case-control studies that employ random (unrelated) population samples are still commonly used (e.g. Deng *et al.* 1999) and advocated (e.g. Risch & Teng, 1998; Morton & Collins, 1999) for populations in which admixture is not much of a concern. A robust approach has been developed to test for the importance of candidate genes with random population samples, even in the presence of population admixture (Pritchard *et al.* 2000). The focus of all of the early analyses of candidate genes in random population samples is to test *for* the importance of candidate genes and little effort has been made to test explicitly *against* their importance. Statistically speaking, a simple qualitative measure of lack of significance in previous association analyses may not be formally employed as the evidence against the importance of candidate genes.

It is known that in linkage analyses with pedigrees or relative pairs, the information collected for linkage analyses may also be employed for exclusion analyses (Edwards, 1980; Ott, 1999; Kruglyak & Lander, 1995). The traditional criteria are that a Logarithm-of-Odds (LOD) score $\geqslant 3.0$ is taken as evidence for significant linkage, a LOD score $\leqslant -2.0$ is taken as evidence against linkage and a LOD score between -2.0 and 3.0 is not conclusive concerning linkage and exclusion for the genomic region under test. Exclusion mapping analyses have been conducted in practice in conjunction with linkage analyses (e.g. Hanis, 1996).

In this study, we develop a LOD score method for exclusion analyses of candidate genes in random population samples. The approach may be employed to test explicitly *against* the importance of candidate genes and may be applied as a complement to regular association analyses to test *for* the importance of candidate genes. The LOD score exclusion analysis to be developed here is parallel to the exclusion mapping analyses in linkage studies. By computer simulations, the performance (the power) of our exclusion analyses is investigated in relation to population parameters and the genetic effects and models assumed in exclusion analyses. Examples of application of our exclusion analyses are provided for three published data sets on *VDR* and estrogen receptor (*ER*) genotypes, for their significance to osteoporotic fractures (a complex genetic disease, Deng *et al.* 2000*a*).

METHODS

For a random population sample with each subject ascertained for the disease status and genotype at the candidate gene under test, we will first construct maximum likelihood functions under the null hypothesis (H_0) and under the alternative hypothesis (H_1). The H_0 is that the locus under test is not a DSL and the H_1 is that the locus under test is a DSL. At a DSL, the penetrances of the genotypes are not all equal. A LOD score for exclusion analysis will subsequently be constructed by the H_0 and H_1. Then, we will derive the maximum likelihood estimates (MLEs) of the penetrances of the candidate gene genotypes under the following four typical inheritance models: dominant, recessive, additive and multiplicative.

LOD Score Construction

n random individuals are sampled from a population and are diagnosed for their disease status and genotyped at a candidate gene under test. Let n_D and n_C be the numbers of affected and nonaffected individuals in the sample, respectively. Assume that the candidate gene is biallelic with alleles A and a. If the candidate gene is a DSL, let A denote the disease susceptible allele and a denote the normal

allele. Let n_{AA}, n_{Aa}, and n_{aa} be the numbers of individuals with genotypes AA, Aa and aa, respectively in the sample. For each genotype in the sample, we can also count the numbers of affected and nonaffected persons. Hence we have the numbers $n_{AA,D}$, $n_{Aa,D}$, $n_{aa,D}$, $n_{AA,C}$, $n_{Aa,C}$, and $n_{aa,C}$ that denote the sample sizes of individuals of different genotypes and disease status. Apparently,

$$n_{AA,D}+n_{AA,C}=n_{AA},\ n_{Aa,D}+n_{Aa,C}=n_{Aa},\ n_{aa,D}+n_{aa,C}=n_{aa},\ n_{AA}+n_{Aa}+n_{aa}=n$$

and

$$n_{AA,D}+n_{Aa,D}+n_{aa,D}=n_D,\ n_{AA,C}+n_{Aa,C}+n_{aa,C}=n_C,\ n_D+n_C=n.$$

Let ψ denote the population prevalence of the disease. Let ψ_g denote the penetrance of the genotype g, where g can be AA, Aa or aa, respectively. Let y be a variable denoting a person's disease status, and let $y=1$ when the person is affected and $y=0$ when he/she is not affected. The probability function for diseases status of an individual with genotype g can be written as:

$$l(y\,|\,g)=(\psi_g)^y(1-\psi_g)^{1-y}.$$

The likelihood L of the n random individuals' disease status given their genotypes at the candidate gene is

$$L=\prod_{i=1}^{n} l_i(y\,|\,g).$$

Under the H_0 that the candidate gene is not a DSL, the penetrances of the three genotypes are the same and equal to ψ, the likelihood L is:

$$L_0=\psi^{n_D}(1-\psi)^{n_c}.$$

Under the H_1 that the candidate gene is a DSL, the likelihood L is:

$$L_1=\psi_{AA}^{n_{AA,D}}(1-\psi_{AA})^{n_{AA,C}}\psi_{Aa}^{n_{Aa,D}}(1-\psi_{Aa})^{n_{Aa,C}}\psi_{aa}^{n_{aa,D}}(1-\psi_{aa})^{n_{aa,C}}$$

Therefore, we can construct a LOD score as follows:

$$Lod=\log_{10}\left[\frac{\hat{L}_1}{\hat{L}_0}\right],$$

where

$$\hat{L}_1=\hat{\psi}_{AA}^{n_{AA,D}}(1-\hat{\psi}_{AA})^{n_{AA,C}}\hat{\psi}^{n_{Aa,D}}\hat{\psi}_{aa}^{n_{aa,D}}(1-\hat{\psi}_{aa})^{n_{aa,C}}$$

and

$$\hat{L}_0=\hat{\psi}^{n_D}(1-\hat{\psi})^{n_c}.$$

The $\hat{\psi}$, $\hat{\psi}_{AA}$, $\hat{\psi}_{Aa}$, and $\hat{\psi}_{aa}$ are the MLEs of ψ, ψ_{AA}, ψ_{Aa} and ψ_{aa}, , respectively. Let $(\partial \log_{10}L_0)/(\partial\psi)=0$, it is straightforward to obtain the MLE of ψ as $\hat{\psi}=(n_D)/n$. It is noted that in the construction of the above likelihood functions, no assumption is involved about the experimental designs of population association studies to be analysed, except that individuals are assumed to be randomly ascertained so that they are not related. Study subjects may be ascertained randomly with regard to their genotypes at the candidate genes to be tested but with their phenotypes considered, a situation in case-control study designs. Random samples that can be analysed in association studies and for this exclusion analyses refer to samples of unrelated individuals. In the following text, we will derive $\hat{\psi}_{AA}$, $\hat{\psi}_{Aa}$ and $\hat{\psi}_{aa}$ under the four typical inheritance models with the maximum likelihood function L_1.

MLE of Penetrances under Four Inheritance Models

Dominant model

Under the dominant model for the disease susceptible allele, we have $\psi_{AA} = \psi_{Aa}$. Let $t = \dfrac{\psi_{AA}}{\psi_{Aa}}$. We term t the genetic effect at the DSL. t is equivalent to the measure of relative risk (Khoury *et al.*, 1993) if allele A is regarded as a risk factor. If $t > 1$, the test locus is a DSL. If $t = 1$, the test locus is not a DSL. t quantifies the magnitude of genetic effects under specified genetic models and it is an important parameter to be tested in our exclusion analyses. L_1 can be rewritten as:

$$L_1 = (t\psi_{aa})^{n_{AA,D}+n_{Aa,D}}(1-t\psi_{aa})^{n_{AA,C}+n_{Aa,C}}\psi_{aa}^{n_{aa,D}}(1-\psi_{aa})^{n_{aa,C}}.$$

Let $(\partial \operatorname{Log}_{10} L_1)/(\partial\psi_{aa}) = 0$ and after some algebraic simplification, we have:

$$nt\psi_{aa}^2 - [n_D(t+1) + (n_{AA,C} + n_{Aa,C})t + n_{aa,C}]\psi_{aa} + n_D = 0.$$

Let $a = nt, b = -[n_D(t+1) + (n_{AAC} + n_{AaC})t + n_{aaC}], c = n_D$, we can obtain the MLEs as follows:

$$\hat{\psi}_{aa} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

$$\hat{\psi}_{AA} = \hat{\psi}_{aa} = t\hat{\psi}_{aa}$$

By definition that A is a potential disease susceptibility allele under H_1, we have $t \geqslant 1$; therefore,

$$\begin{aligned} b^2 - 4ac &= [n_D(t+1) + (n_{AAC} + n_{AaC})t + n_{aaC}]^2 - 4ntn_D \\ &> (n_D t + n_D + n_{AAC} + n_{AaC} + n_{aaC})^2 - 4ntn_D \\ &= (n_D t + n)^2 - 4ntn_D = (n_D t - n)^2 \geqslant 0 \end{aligned}$$

Hence we will have two real roots of $\hat{\psi}_{aa}$. If both of these roots satisfy the conditions that $0 < \psi_{AA}$, ψ_{Aa}, $\psi_{aa} < 1$, we will compare the likelihoods corresponding to these two roots. The root corresponding to a larger likelihood is the MLE of $\hat{\psi}_{aa}$. Therefore, under the dominant inheritance model, the LOD score can be computed as:

$$LOD = \log_{10}\left[\frac{\hat{L_1}}{\hat{L_0}}\right] = (n_{AA,D} + n_{Aa,D})\log_{10}(t\hat{\psi}_{aa}) + (n_{AA,C} + n_{Aa,C})\log_{10}(1 - t\hat{\psi}_{aa})$$

$$+ n_{aa,D}\log_{10}(\hat{\psi}_{\alpha\alpha}) + n_{aa,C}\log_{10}(1-\hat{\psi}_{aa}) - n_D\log_{10}\left(\frac{n_D}{n}\right) - n_C\log_{10}\left(\frac{n_C}{n}\right).$$

Recessive model

Under the recessive model, we have $\psi_{Aa} = \psi_{aa}$. Let $t = (\psi_{AA})/(\psi_{aa})$, we have:

$$L_1 = t^{n_{AA,D}}\psi_{aa}^{n_D}(1-t\psi_{aa})^{n_{AA,C}}(1-\psi_{aa})^{n_{aa,C}+n_{aa,C}}.$$

Let $(\partial \operatorname{Log}_{10} L_1)/(\partial\psi_{aa}) = 0$ and after some algebraic simplification, we have:

$$nt\psi_{aa}^2 - [n_D(t+1) + (n_{aa,C} + n_{Aa,C}) + n_{AA,C}\,t]\psi_{aa} + n_D = 0.$$

Let $a = nt, b = -[n_D(t+1) + (n_{aaC} + n_{AaC}) + n_{AAC}\,t], c = n_D$, the MLEs are as follows:

$$\hat{\psi}_{Aa} = \hat{\psi}_{aa} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$\hat{\psi}_{AA} = t\hat{\psi}_{aa}.$$

As earlier, by noting that $t \geqslant 1$, we can show that $b^2-4ac \geqslant 0$. One of the two real roots to be obtained will be chosen as the $\hat{\psi}_{aa}$ as earlier. The LOD score is:

$$LOD = \log_{10}\left[\frac{\hat{L}_1}{\hat{L}_0}\right] = n_{AA,D}\log_{10}(t\hat{\psi}_{aa}) + n_{AA,C}\log_{10}(1-t\hat{\psi}_{aa}) + (n_{aa,D}+n_{aa,D})\log_{10}(\hat{\psi}_{aa})$$

$$+(n_{aa,C}+n_{Aa,C})\log_{10}(1-\hat{\psi}_{aa}) - n_D\log_{10}\left(\frac{n_D}{n}\right) - n_C\log_{10}\left(\frac{n_C}{n}\right)$$

Additive model

Under the additive model, we have $\psi_{AA}+\psi_{aa} = 2\psi_{Aa}$. Let $t = (\psi_{Aa})/(\psi_{aa})$. We have $\psi_{AA} = 2\psi_{Aa}-\psi_{aa} = (2t-1)\psi_{aa}$; therefore,

$$L_1 = (2t-1)^{n_{AA,D}}t^{n_{Aa,D}}\psi_{aa}^{n_D}(1-(2t-1)\psi_{aa})^{n_{AA,C}}(1-t\psi_{aa})^{n_{Aa,C}}(1-\psi_{aa})^{n_{aa,C}}.$$

Let $(\partial \operatorname{Log}_{10}L_1)/(\partial\psi_{aa}) = 0$, we have:

$$\frac{n_D}{\psi_{aa}} - (2t-1)\frac{n_{AA,C}}{1-(2t-1)\psi_{aa}} - t\frac{n_{Aa,C}}{1-t\psi_{aa}} - \frac{n_{aa,C}}{1-\psi_{aa}} = 0$$

This is a cubic equation of ψ_{aa}. With the aid of Mathematica (Wolfram, 1996), we can calculate the roots of this equation. The root corresponding to the largest likelihood is the MLE of ψ_{aa}. The LOD score can be computed easily as earlier.

Multiplicative model

Under the multiplicative model, we have the equation $\psi_{AA}*\psi_{aa} = \psi_{Aa}^2$. Let $t = (\psi_{Aa})/(\psi_{aa})$, then $\psi_{AA} = t^2\psi_{aa}$. L_1 can be rewritten as:

$$L_1 = t^{2n_{AA,D}+n_{Aa,D}}\psi_{aa}^{n_D}(1-t^2\psi_{aa})^{n_{AA,C}}(1-t\psi_{aa})^{n_{Aa,C}}(1-\psi_{aa})^{n_{aa,C}}$$

Let $(\partial \operatorname{Log}_{10}L_1)/(\partial\psi_{aa}) = 0$, we have,

$$\frac{n_D}{\psi_{aa}} - t^2\frac{n_{AA,C}}{1-t^2\psi_{aa}} - t\frac{n_{Aa,C}}{1-t\psi_{aa}} - \frac{n_{aa,C}}{1-\psi_{aa}} = 0.$$

Similar to the additive model, the above cubic equation can be solved for the MLE of ψ_{aa} to obtain the LOD score.

Although we only examined the four typical inheritance models at the candidate gene for the demonstration of our exclusion analyses, other inheritance models (such as partial dominant, partial recessive) may also be investigated in a similar manner. For example, let $t = (\psi_{AA})/(\psi_{aa})$ $(t > 1.0)$, under a partial dominant or partial recessive genetic model, we have $\psi_{Aa} = \Omega\psi_{aa}$, where Ω is a parameter for specific genetic models. The value of Ω is fixed once a genetic model is specified (or inferred from the data) for testing. The values of $\Omega = \mathrm{t}$ and $\Omega = 1$ represent the dominant and recessive genetic models investigated earlier. A specified value of Ω with $t > \Omega > (t+1)/2$ represents partial dominant models and a specified value of Ω with $1 < \Omega < (t+1)/2$ represents models for partial recessive.

The MLEs of the prevalence $\hat{\psi}$ and the penetrances, $\hat{\psi}_{AA}$, $\hat{\psi}_{Aa}$ and $\hat{\psi}_{aa}$, and thus the LOD scores are functions of $n_{AA,D}$, $n_{Aa,D}$, $n_{aa,D}$, $n_{AA,C}$, $n_{Aa,C}$, and $n_{aa,C}$, and the genetic effect t that needs to be assumed under a specific genetic model for exclusion analyses. The above n's can be ascertained from the population sample. The hypothesis to be tested in exclusion analysis is that the candidate gene is a DSL with a genetic effect t under a specific genetic model. With the traditional criteria (Ott, 1999) for exclusion in linkage analyses, a LOD score $\leqslant -2.0$ is taken as evidence to exclude a candidate gene as being a DSL with a genetic effect t under a specified inheritance model. The choice of a LOD

score of -2.0 is a reasonable, though somewhat arbitrary, criterion. A LOD score of -2.0 corresponds to a likelihood ratio of 1:00, i.e., 100:1 odds against linkage, which is a very stringent criteria for exclusion analyses even in whole genome-wide analyses (Ott, 1999).

Computer simulations

To investigate the performance of our exclusion analyses of candidate gene(s) with random population samples, we perform computer simulations. The simulated population may not be randomly mating and thus may not be in Hardy-Weinberg equilibrium. Let f_A be the frequency of allele A, P_{AA} be the frequency of the genotype AA, δ be the Hardy–Weinberg disequilibrium (HWD) coefficient. By definition (Weir, 1996), we have:

$$P_{AA} = f_A^2 + \delta.$$

For a biallelic marker, we have (Weir, 1996):

$$P_{Aa} = 2f_A(1-f_A) - 2\delta$$
$$P_{aa} = (1-f_A)^2 + \delta,$$

where P_{Aa} and P_{aa} are the frequencies of the genotypes Aa and aa, respectively. Considering all the possible values of genotype frequencies, the range of δ is (Weir, 1996):

$$\delta_{\min} = \max[-f_A^2, -(1-f_A)^2] \leqslant \delta \leqslant f_A(1-f_A) = \delta_{\max},$$

since $\delta_{\min} < 0$ and $\delta_{\max} > 0$, δ could be either negative or positive. With admixture of randomly mating sub-populations, a scenario that is found in genetic analyses for candidate genes, δ is negative due to the well-known Walnut effect (Hartl & Clark, 1989). However, it is not unusual for δ estimated from humans and animal populations to be positive for genetic loci examined (Nei, 1987; Lynch & Spitze, 1994). Therefore, while we will mainly test the situations when δ is negative, the situations when δ is positive and 0 (for unstructured and randomly mating populations) will also be investigated.

In simulations when the candidate gene is not a DSL, for a given allele frequency f_A, δ may be $0.5\,\delta_{\min}$, 0 or $0.5\,\delta_{\max}$, respectively. The expected genotype frequencies in the whole population can be obtained based on f_A and δ. For the simulations in which the candidate gene is not a DSL, under a specified population prevalence ψ, n random individuals can be easily simulated for their genotypes and disease status and $n_{AA,D}$, $n_{Aa,D}$, $n_{aa,D}$, $n_{AA,C}$, $n_{Aa,C}$, and $n_{aa,C}$ in the sample can be obtained. For a particular inheritance model and a specified genetic effect t under test, we can calculate the MLEs, $\hat{\psi}$, $\hat{\psi}_{AA}$, $\hat{\psi}_{Aa}$ and $\hat{\psi}_{aa}$, and thus the LOD score as detailed earlier for the random population sample simulated. For each parameter set (such as n, f_A and δ), we perform 10000 repeated simulations. In the 10000 repeated simulations, the proportion of the times that the LOD score is $\leqslant -2.0$ is a measure of our approach for its ability to exclude the candidate gene as a DSL under the specified inheritance model and genetic effect t. We term this proportion as the exclusion power. We investigate how the parameters δ, f_A, ψ, n, t and the inheritance models assumed in analyses will influence the exclusion power.

In simulations, we also examine the performance of our exclusion analyses when the candidate gene is indeed a DSL with the genotype penetrances not all being equal and the true genetic effect being t_1. In this situation, the proportion of the times that the LOD score is $\leqslant -2.0$ with assumed t ($> t_1$) is also a measure of our exclusion analyses on how powerfully it may exclude the candidate gene as a DSL with an assumed genetic effect $t(> t_1)$. In simulations, with the genotype frequencies and y specified, we can calculate the genotype penetrances by the following equation:

$$P_{AA} * x_1(t_1) * \psi_{aa} + P_{Aa} * x_2(t_1) * \psi_{aa} + P_{aa} * \psi_{aa} = \psi,$$

Table 1. *Distribution of candidate gene genotypes according to history of osteoporotic fracture*

ER gene PvuII genotypes (Vandevyver *et al.* 1999)			VDR gene *Fok*I genotypes (Gennari *et al.* (1999)			VDR gene *Bsm*I genotypes (Houston *et al.* 1996 and Gomez *et al.* 1999)		
Genotype	No fracture	Fracture	Genotype	No fracture	Fracture	Genotype	No fracture	Fracture
PP	23	31	FF	138	21	BB	47	21
Pp	84	60	Ff	156	30	Bb	128	57
pp	59	51	ff	38	17	bb	105	42

where

$$\psi_{AA} = x_1(t_1)*\psi_{aa}, \psi_{Aa} = x_2(t_1)*\psi_{aa},$$

x_1 and x_2 are functions of t_1, which are decided by the inheritance models we assume in simulations. For example,

$$x_1(t_1) = \begin{cases} t_1 & \text{Dominant} \\ t_1 & \text{Recessive} \\ 2t_1 - 1 & \text{Additive} \\ t_1^2 & \text{Multipicative} \end{cases} \quad \text{and} \quad x_2(t_1) = \begin{cases} t_1 & \text{Dominant} \\ 1 & \text{Recessive} \\ t_1 & \text{Additive} \\ t_1 & \text{Multipicative} \end{cases}$$

Then, with the genotype frequencies, genetic effect t_1 and ψ, we can simulate a population sample of n individuals and to ascertain $n_{AA,D}$, $n_{Aa,D}$, $n_{aa,D}$, $n_{AA,C}$, $n_{Aa,C}$, $n_{aa,C}$ in the sample. Then we can apply our LOD score method for exclusion analyses with an assumed genetic effect t. In simulation analyses for this case, the inheritance models assumed in exclusion analyses are the same as those simulated. This may be justified, since in practice some clues can be obtained about the inheritance modes of a candidate locus (if it is a DSL) by data of the genotypes and diseases status of a random population sample. In fact, based on the data, a most likely model may also be chosen by comparing the LOD scores computed under different models for the same t. The model with the largest LOD score best fits the data and thus may be chosen in analyses.

The exclusion power is equivalent to the statistical power when the candidate gene is not a DSL and t assumed in analyses is larger than 1.0, or when the candidate gene is a DSL with a true effect t_1 but an effect t ($t > t_1$) is assumed in analyses. The exclusion power is equivalent to type one error rate when t assumed in analyses is the true effect t_1 at the candidate gene ($t_1 = 1$ when the locus is not a DSL and $t_1 > 1$ when it is a DSL).

Exclusion analyses of VDR and ER genes for osteoporotic fractures

We collected genotype data for the polymorphisms at the *Fok*I and *Bsm*I restriction sites in the *VDR* gene and at the *Pvu*II restriction site in the *ER* gene, together with the phenotype data for osteoporotic fractures (Table 1). We perform exclusion analyses against the importance of these two genes for osteoporotic fractures with our LOD score method described above. For the *Fok*I polymorphism in the *VDR* gene, the data employed are from Gennari *et al.* (1999). For the *Bsm*I polymorphism in the *VDR* gene, the data are from Houston *et al.* (1996) and Gomez *et al.* (1999). For the *Pvu*II polymorphisms in the *ER* gene, the data are from Vandevyver *et al.* (1999). The analyses are performed under the four typical inheritance models.

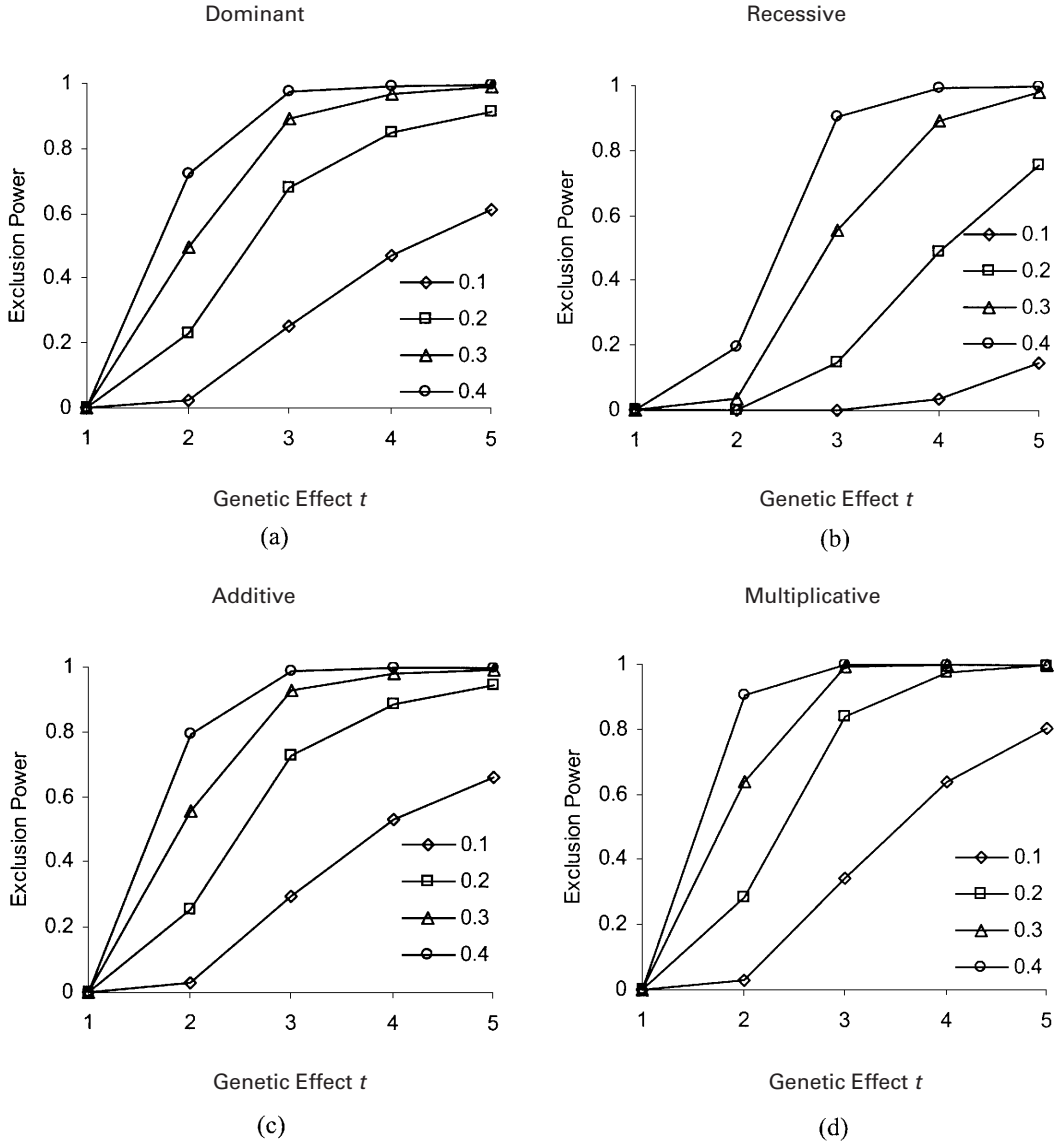

Fig. 1. The effect of population penetrance on exclusion analysis power. Populations with HWD coefficient $\delta = 0.5\ \delta_{\min}$ are simulated. 200 individuals are sampled in each simulation. The allele frequency f_A is 0.3. The population prevalence ψ of 0.1, 0.2, 0.3 and 0.4 are simulated respectively as indicated on each plot. X-axis is the genetic effect t assumed in exclusion analyses. In Figs 1–4, the candidate gene is not a DSL.

RESULTS

In Figs 1–5, we present some representative data of our extensive simulation results for a range of parameter values with different inheritance models. It can be seen that, the larger the genetic effect t assumed in analyses, the higher the power in our exclusion analyses when the candidate gene is not a DSL (Figs 1, 2, 4). This result is straightforward since the candidate gene we simulate is not a DSL, the larger the genetic effect t that we assume this locus has, the less likely it is that the hypothesis is compatible with the simulated data and the more likely we can exclude the locus as a DSL. When t is assumed to be 1, the exclusion power (equivalent to the type one error rate) is very small (< 0.005). This result is consistent with our expectation. This is because when t is assumed to be 1 in analyses, we assume that the locus is not a DSL, which is consistent with the simulated fact.

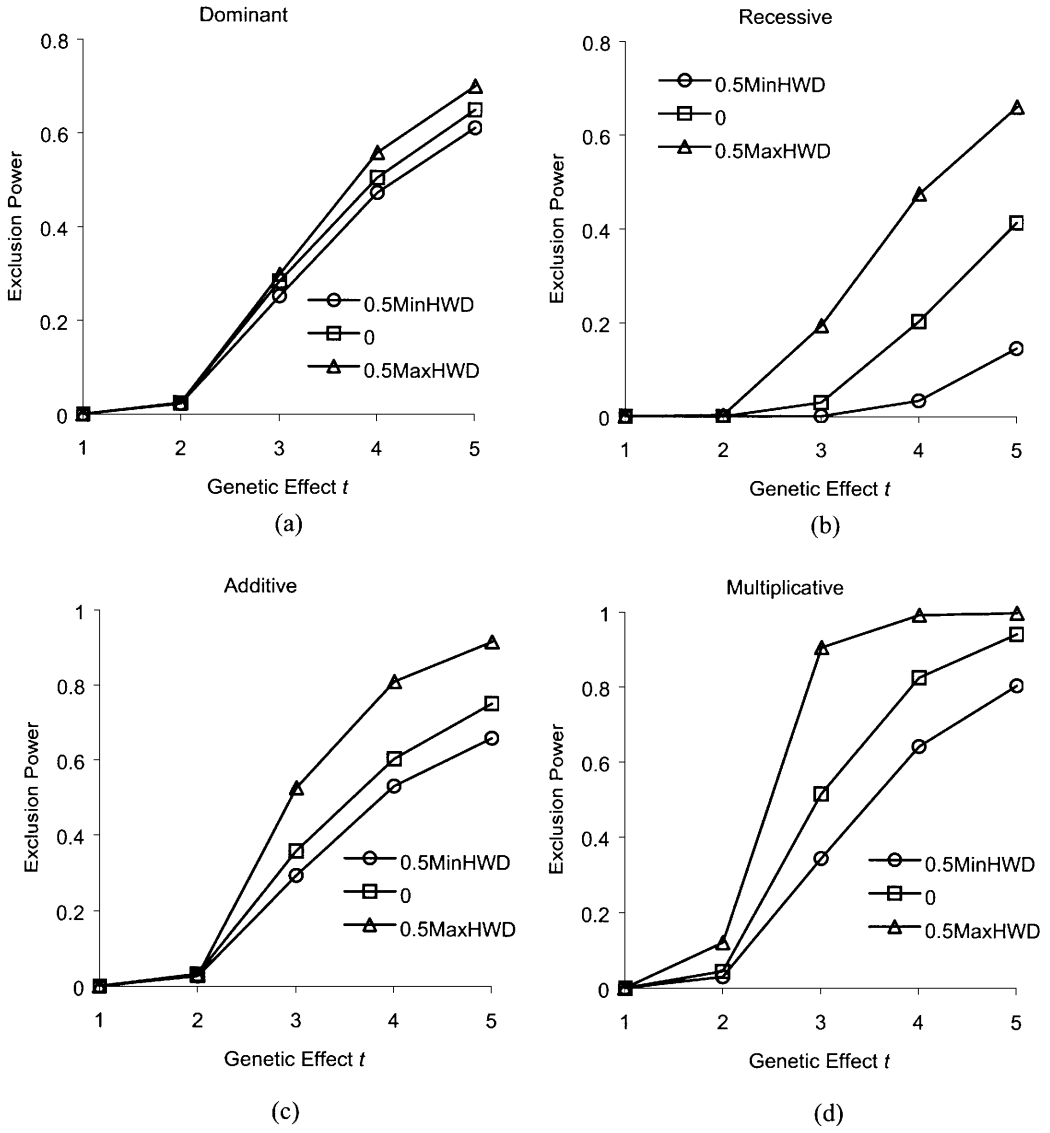

Fig. 2. The effect of HWD coefficient on exclusion analysis power. Populations with $\delta = 0.5\ \delta_{\min}$, 0, $0.5\delta_{\max}$ are simulated, respectively as indicated on each plot. $n = 200, f_A = 0.3, \psi = 0.1$. X-axis is the t assumed in exclusion analyses.

Therefore, the exclusion power for it not to be a DSL (i.e., $t = 1$) should be nearly 0. When the simulated locus is indeed a DSL (Fig. 5) with a true effect t_1, if we assume $t = t_1$ in analyses, the exclusion power is also very small (< 0.005), again consistent with expectation, since the locus is indeed a DSL with an effect t_1. However, if we assume $t > t_1$ in analyses, we can exclude the locus as having a larger effect t than its true effect t_1, the larger the effect t we assume in analyses, the higher is our exclusion power. This result demonstrates that even if the candidate gene is a DSL with an effect t_1, our exclusion approach is still useful for excluding this locus as having a genetic effect exceeding a specified value t ($t > t_1$). This feature is useful, since generally those DSLs with certain effects may be useful in practical and clinical settings and thus we can exclude a locus that is not a DSL or is a DSL but with only a minor effect. There may be cases in which a DSL is of small effects in terms of relative risk but of significance in terms of population attributable risk due to a high frequency (> 0.5) of the disease allele (Altshuler *et al.* 2000). However, such cases are rare since

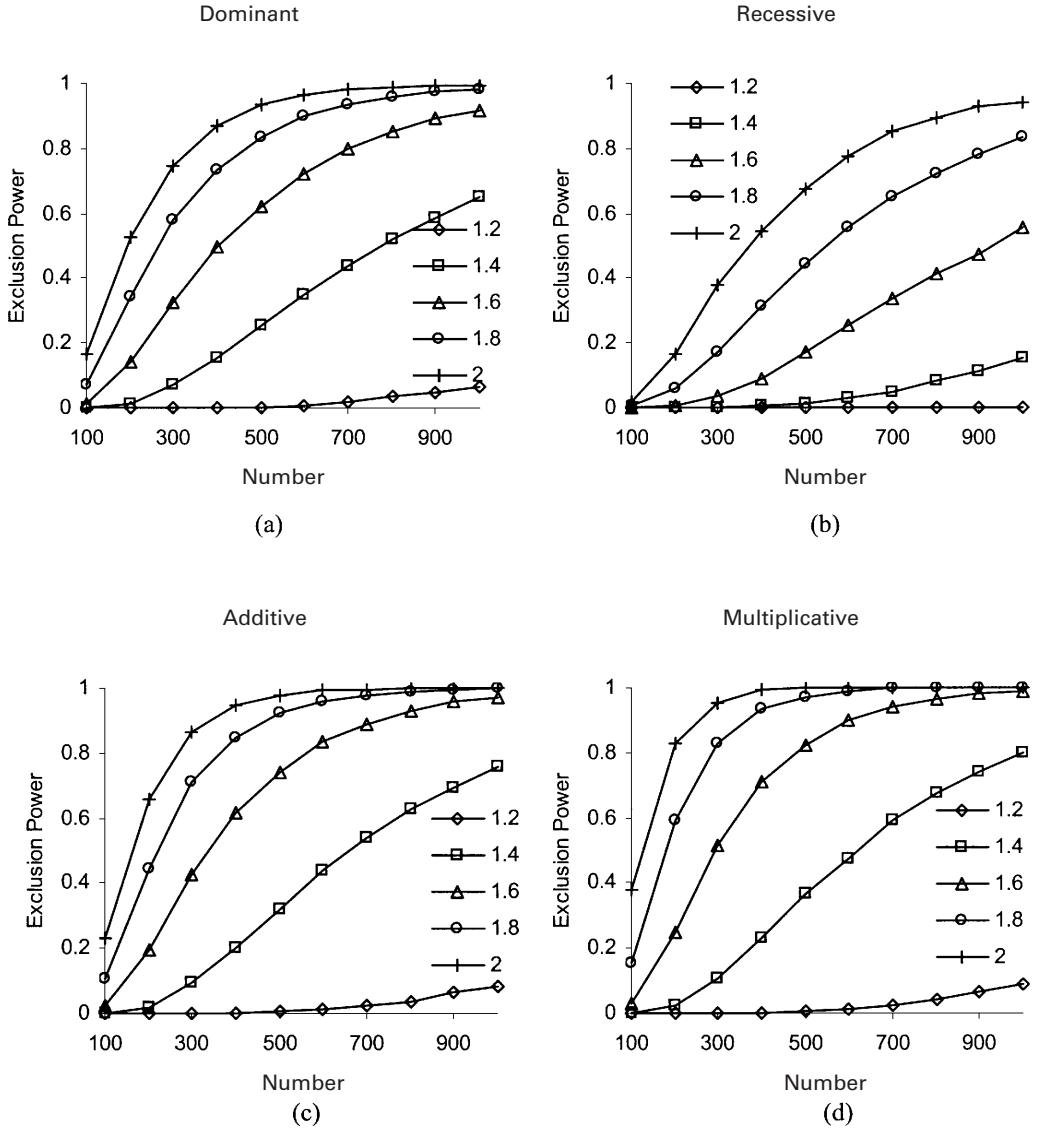

Fig. 3. The number (sample size) needed to achieve certain powers. Populations with $f_A = 0.3$, $\delta = 0$ and $\psi = 0.30$ are simulated. The t assumed in the exclusion analyses is indicated on each plot. X-axis is the number of individuals needed to achieve the power for the exclusion of the assumed genetic effect t.

disease alleles do not generally achieve a high frequency, especially when the disease is important for an individual's fitness (Hartl & Clark, 1989). In these rare cases, the DSL may be of epidemiological significance; however, its clinical importance may be relatively minor, since it does not contribute much to individuals' differences in their susceptibilities to the disease.

In addition to the t assumed in analyses, the exclusion power is also affected by population disease prevalence ψ (Fig. 1), population HWD coefficient δ (Fig. 2), sample size n (Fig. 3), frequency f_A (< 0.5) of the disease allele (Fig. 4), and the inheritance model assumed in analyses (Figs 1–5). Generally speaking, other things being equal, the exclusion power increases with ψ, and the increase is more significant when t assumed in analyses is small (Fig. 1). The exclusion power also increases with δ, and the increase is more significant when t assumed in analyses is large (Fig. 2). Therefore, under population admixture when δ is smaller than 0, the exclusion power is smaller than unstructured and randomly mating populations. This will lead to conservative conclusions from exclusion analyses, as

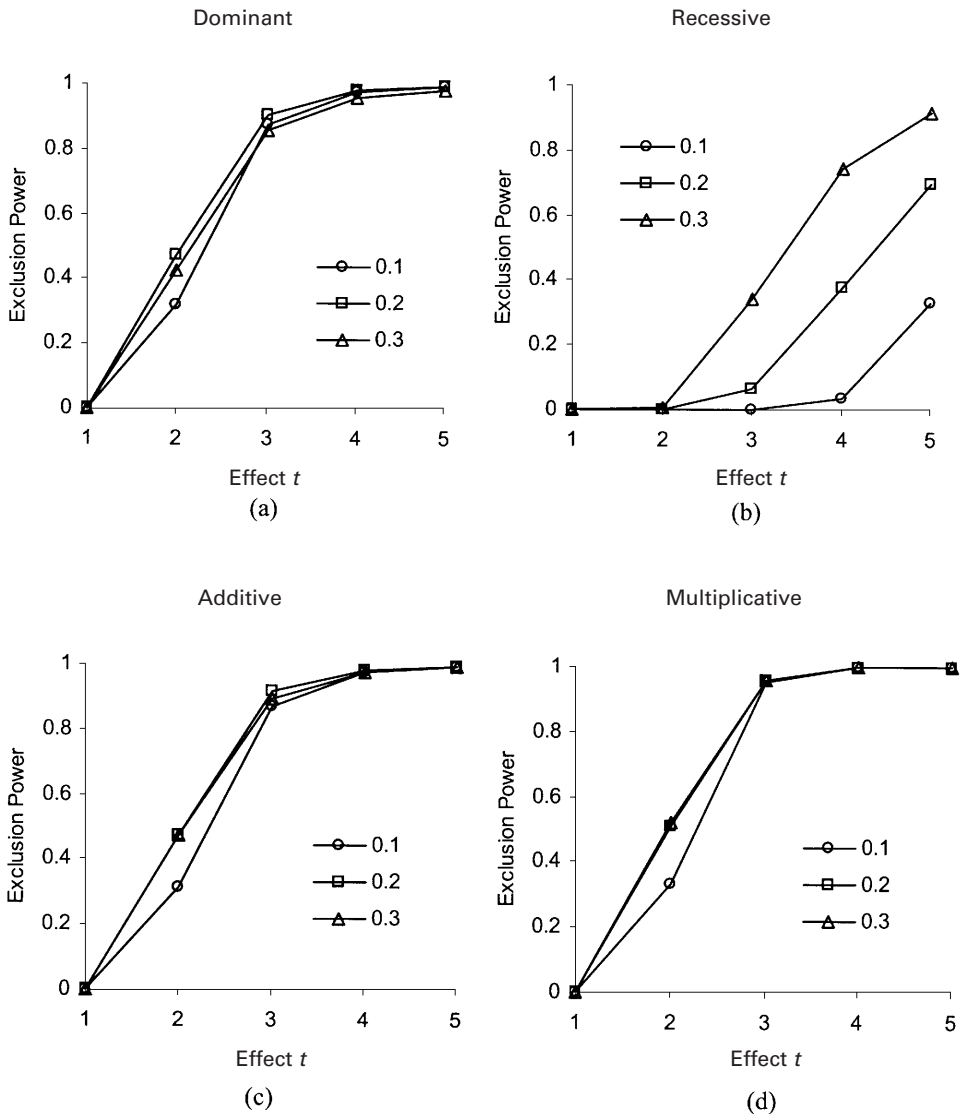

Fig. 4. The effect of the frequency of allele A on exclusion analyses. Populations are simulated with $\delta = 0.5\,\delta_{\min}$, $n = 300$, and $\psi = 0.1$. $f_A = 0.1$, 0.2 or 0.3, respectively as indicated on each plot. X-axis is the t assumed in exclusion analyses and is the genetic effect we wish to exclude.

opposed to the potentially liberal conclusions from usual association analyses (due to the inflated power under population admixture). The exclusion power increases with larger n and the increase is more dramatic for larger t's assumed in analyses (Fig. 3). The larger the f_A (when it is < 0.5), the higher the exclusion power and difference is more significant with relatively small t's assumed in analyses (such as $t = \sim 2$) (Fig. 4). The dependency of our exclusion power on the genetic model assumed in analyses is apparent (Figs 1–5); however, no clear pattern seems to exist although the exclusion power under the multiplicative model is usually higher than under the other three inheritance models.

To exclude a candidate gene with a moderate effect t (> 1.4), the sample size required by our approach is reasonable and moderate; it is within the range of the sample sizes employed in many association studies (Fig. 3). Generally, depending on the inheritance model of the locus, for a t assumed between 1.4 to 2.0, a sample size of 200–1000 may achieve an exclusion power of more than 80% (Fig. 3) when the disease is relatively prevalent ($\psi = 0.3$).

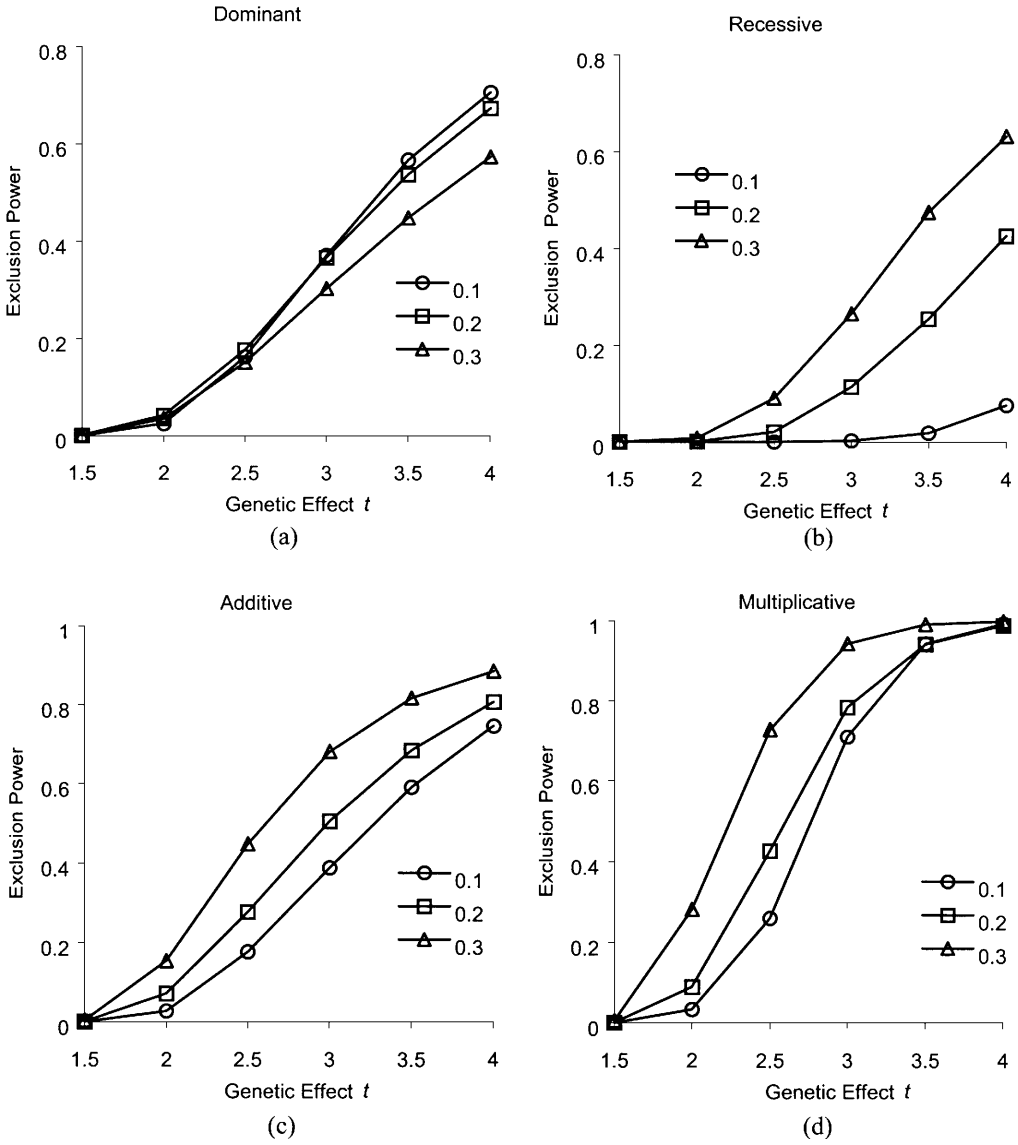

Fig. 5. The exclusion analysis when the candidate gene is a DSL. Populations are simulated with $\delta = 0$, $n = 300$, and $\psi = 0.2$. $f_A = 0.1$, 0.2 or 0.3, respectively as indicated in each plot. The genetic effect t_1 simulated for the DSL is 1.5. The X-axis is the genetic effect we would like to exclude and we assume in exclusion amylases.

The results of the exclusion analyses for the *ER* and *VDR* genes are summarized in Fig. 6. For all three analyses, it can be seen that the LOD score for the recessive inheritance model is higher than the corresponding LOD scores for the other three typical inheritance models. Therefore, recessive inheritance is the most likely inheritance model among the four inheritance models typically analysed in genetics. The genetic effect t's under various inheritance models that can be excluded are those with LOD scores below -2.0, which is clearly indicated on each exclusion plot. For example, under the recessive inheritance model, a genetic effect t greater than 2.04 can be excluded at the *ER* locus. Equivalently, under the recessive model, the genotype penetrance of PP individuals is no more than 2.04 times larger than that of Pp and pp individuals at the *ER* locus. Under the recessive inheritance model, a genetic effect t greater than 1.81 can be excluded at the *VDR Bsm*I locus. Under the recessive inheritance model, a genetic effect t greater than 5.21 can be excluded at the *VDR Fok*I

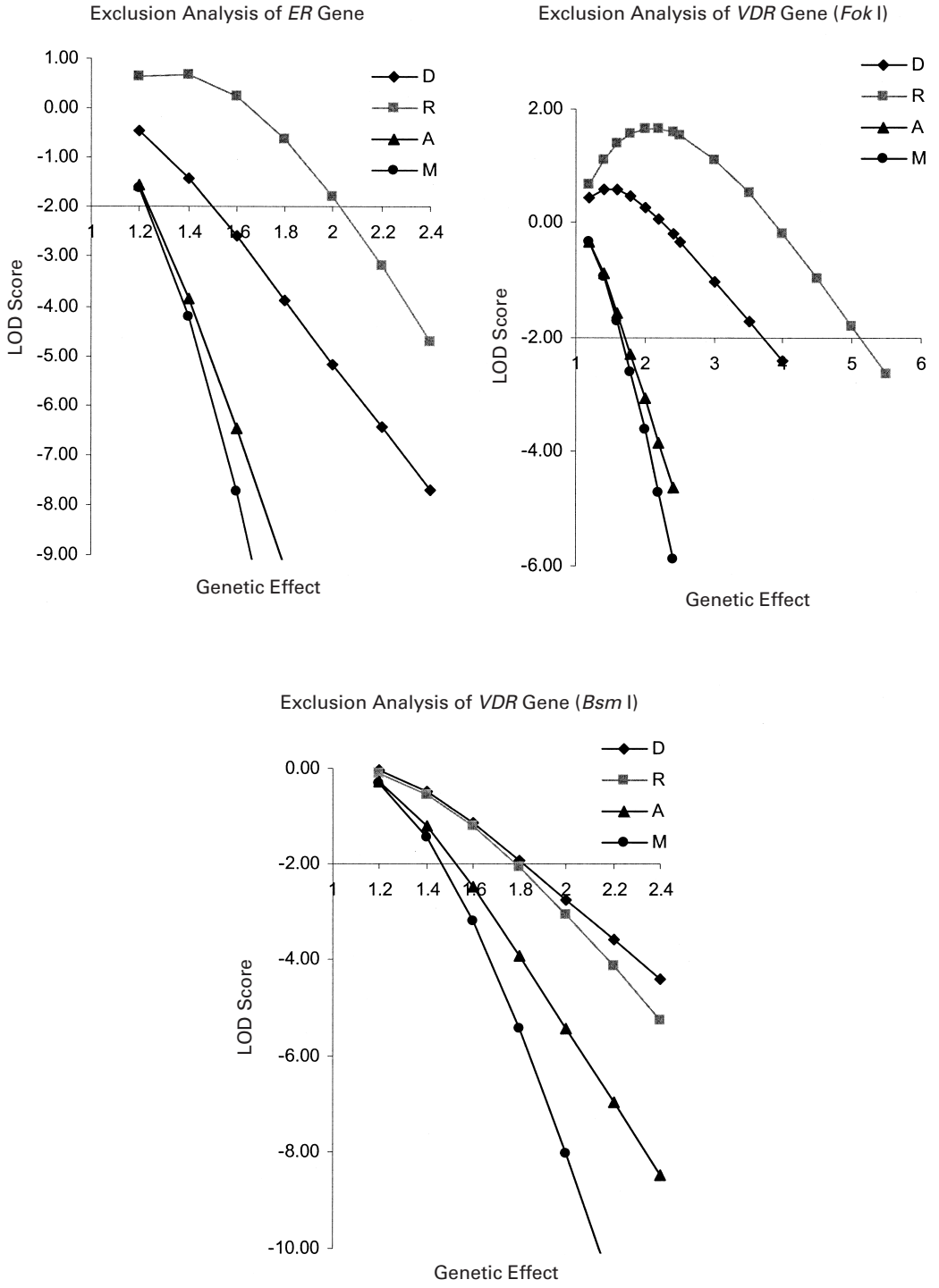

Fig. 6. Exclusion analyses of *VDR* and *ER* genes as DSLs osteoporotic fractures. X-axis is the genetic effect *t* assumed in exclusion analyses. In each plot, 'D', 'R', 'A', 'M' denotes respectively for dominant, recessive, additive and multiplicative inheritance models at the candidates under test. The allele for which the homozygote has higher disease incidence than the other allele (reflected in Table 1) is designated as the disease allele A. For the *ER* gene, it is the P allele; for the *VDR Fok*I site, it is the f allele; for the *VDR Bsm*I site, it is the B allele.

locus. Under the other inheritance models, much smaller genetic effects may be excluded. Interestingly, under the recessive inheritance model, a maximum LOD score of 1.65 is achieved for the genetic effect t of about 2.2 at the *VDR Fok*I locus. A LOD score of 1.65 corresponds roughly to a p-value of 0.0058 (Lander & Kruglyak, 1995) and suggests the potential association of the *VDR Fok*I genotypes with the risk to osteoporotic fracture.

discussion

Random population samples have been widely employed to test association of candidate genes and complex diseases. It is well know that such studies are prone to false positive findings due to population admixture and the results are usually inconsistent and controversial. While extensive analyses have been conducted to test *for*, no formal analyses have been conducted to test *against*, the importance of candidate genes within random population samples. Statistically speaking, nonsignificance of the usual association analyses for candidate genes cannot be taken as formal evidence to unambiguously exclude their importance.

In this study, we develop a LOD score approach for exclusion analyses of candidate genes with random population samples. Under this approach, specific genetic effects and inheritance models at candidate genes can be analysed. If a LOD score is $\leqslant -2.0$, the locus can be excluded from having an effect larger than that specified under the specified inheritance model. The most likely inheritance model at the candidate gene locus may be inferred by our LOD score approach as demonstrated by our analyses for the three published data sets for the two candidate genes for osteoporotic fractures. Computer simulations show that, if a candidate gene is not a DSL, with sample sizes commonly employed, this approach has high power to exclude a gene from having moderate genetic effects. Even if the candidate gene is a DSL, our analyses may still be applied to exclude it from having an effect different from its true effect. In contrast to regular association analyses for the importance of candidate genes, population admixture will not affect the robustness of our analyses; in fact, it usually renders our exclusion analyses more conservative. Under population admixture, the LOD score is higher than when population admixture is absent, which yields lower exclusion power (Fig. 2) and thus more conservative conclusions. The conservative and robust conclusions from our exclusion analyses *against* the importance of candidate genes is also consistent with the well know fact that population admixture increases false positive associations *for* the importance of candidate genes which may lead to liberal conclusions. When samples are recruited with regard to the source sub-populations in an admixed population, situations can be deliberately constructed so that even when a factor is associated with a disease in sub-populations, no association is shown in an admixed population (Armitage & Berry, 1987). This will lead to exclusion of a true DSL when applying our analyses. However, it is extremely rarely (if ever) the case that one can actually identify the source sub-populations for individuals in admixed populations. Therefore, as long as unrelated individuals are recruited randomly without regard to the sources of the sub-populations, our exclusion mapping analyses should be robust. The robustness of our exclusion analyses is also reflected by the very small type-one-error rate (< 0.005) as revealed in simulations. This is largely because of the stringent criterion of LOD score of -2.0 adopted for exclusion.

Our exclusion analysis is powerful in that the sample size required by our approach is moderate and is within the range of the sample sizes employed in many association studies, particularly for common diseases. Our exclusion approach should also be useful for testing the importance of candidate genes under specific genetic effects and models. Usually, for a candidate gene to be of significant interest from the perspective of clinical and basic research, it should have a certain

magnitude of effect on the differential susceptibilities to a disease. Thus, exclusion analyses against candidate genes are significant not only when a candidate gene is not a DSL but also when it is a DSL but with a minor effect. Our exclusion analysis complements association analysis for candidate genes in random population samples and is parallel to the exclusion mapping analyses that may be conducted in linkage analyses with pedigrees or relative pairs. Thus, the exclusion analyses developed here should be of significant practical value and the significance is demonstrated by an application to test the importance of the *VDR* and *ER* genes for osteoporotic fractures. The *VDR* and *ER* genes have been extensively studied in osteoporosis research (Eisman, 1995; Peacock, 1995; Deng *et al*. 1999; Gong *et al*. 1999) and their importance is under debate. Our LOD score analyses exclude that, under the recessive model that is most compatible with the data, the genetic effect t cannot be greater than, respectively, 2.04 at the *ER* gene (*Pvu*II polymorphisms), 1.81 at the *VDR* gene (*Bsm*I polymorphisms), and 5.21 at the *VDR* gene (*Fok*I polymorphisms).

It is known that the LOD score approach may be used for both linkage and exclusion analyses (Edwards, 1980; Ott, 1999; Kruglyak & Lander, 1995) for human pedigrees or relative pairs. Our LOD score analyses may also be employed for association as well as exclusion analyses for random population samples. However, association analysis is not a focus here, since there have already been many other association analyses methods. The utility of our LOD score analyses for association studies is demonstrated for the analyses of the *VDR* gene (*Fok*I polymorphisms). Under the recessive inheritance model that is identified by our LOD score analyses to be most compatible with the data, a maximum LOD score of 1.65 is achieved for the genetic effect t of about 2.2 at the *VDR Fok*I locus. This magnitude of genetic effect is in agreement with the result for a significant relative risk of 2.5 revealed for the genotypes at this locus (Gennari *et al*., 1999). The more important effect identified for the *VDR Fok*I polymorphisms relative to the *Bsm*I polymorphisms is compatible with our knowledge of the locations of these two polymorphisms inside the *VDR* gene. The *VDR* gene is a gene of a relatively long DNA sequence and has 10 exons (Audi *et al*. 1999). The *Fok*I polymorphism site is at the transcription initiation site of the *VDR* gene and thus may be of important biological significance (Audi *et al*. 1999). On the other had, the *Bsm*I polymorphism site is at the intron VIII with no apparent biological significance.

It should be noted that, in general and from the above practical example of the application to the *VDR* gene, what are tested in our exclusion analyses are the specific markers genotyped inside candidate genes. This is also true for association analyses. A marker inside a candidate gene may be excluded (likewise, a marker may also not be significant in association analyses) even if there may be a functional mutation elsewhere inside the candidate gene. This may occur if the candidate gene is large and if the marker is relatively far away from the functional mutation and is not in (or is in weak) linkage disequilibrium with the functional mutation. Therefore, in both association and exclusion analyses, only after thorough examination of polymorphisms densely located throughout the candidate gene (or even outside of the gene, in its transcription regulation regions) may we conclude on the importance or lack of importance of the candidate gene *per se*.

In summary, the exclusion analyses developed here should be of practical value. Our exclusion analysis complements association analysis for the importance of candidate genes in random population samples and is parallel to the exclusion mapping analyses that may be conducted in linkage analyses with pedigrees or relative pairs. Its usefulness, high power and robustness may warrant it being implemented with random population samples to rule out candidate genes with no, or minor, effects. Although the current exclusion analysis approach is developed for testing candidate genes as DSL for dichotomous disease traits, like many other analyses for DSL, the approach developed here may also be directly applied for exclusion analyses of candidate genes as QTL for quantities traits (Deng *et al*. 2000*b*).

This study was supported by grants from NIH, Health Future Foundation, State of Nebraska (Department of Health and Human Services Cancer and Smoking Related Disease Research Program, LB595) and HuNan Normal University, and by a graduate student tuition waiver to J. L. from Creighton University. We are grateful to the three anonymous reviewers for helpful comments that improved the manuscript.

REFERENCES

Altshuler, D., Hirschhorn J. N., Klannemark M. & Lander E. S. (2000). The common PPARγPro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nature Genet. **26**, 76–80.

Armitage, P. & Berry G. (1987). *Statistical methods in medical research*. Oxford, Blackwell. UK.

Audi, L., Garcia-Ramirez, M. & Carrascosa, A. (1999). Genetic determinants of Bone Mass. *Hormone Research* **51,** 105–123.

Blum, K., Nobel, E. P. & Sheridan, P. J., *et al*. (1990). Allelic association of human dopamine D2 receptor gene in alcoholism. *JAMA* **263**, 2055–2060.

Blum, K., Noble, E. P. & Sheridan, P. J., *et al*. (1991). Association of the A1 allele of the D2 Dopamine receptor gene with severe alcoholism. *Alcohol* **8**, 409–416.

Chagnon, Y. C., Perusse, L. & Bouchard, C. (1998). The human obesity gene map: the 1997 update. *Obes. Res.* **5**, 76–92.

Chakraborty, R. & Smouse, P. (1988). Recombination in haplotypes leads to biased estimates of admixture proportions in human populations. *Proc. Natl. Acad. Sci. USA* **85**, 3071–3074.

Deng, H. W., Li, J., Li, J. L., Johnson, M., Gong, G. & Recker, R. R. (1999). Association of VDR and Estrogen Receptor Genotypes with Bone Mass in Postmenopausal Caucasian Women: Different Conclusions with Different Analyses and the Implications. *Osteoporosis Intl*. **9**, 499–507.

Deng, H. W. & Chen, W. M. (2000). Re: 'Biased tests of association: comparison of allele frequencies when departing from Hardy-Weinberg proportions'. *Am. J. Epidemiology* **151**, 335–357.

Deng, H. W., Chen, W. M., Recker, S., Stegman, M. R., Li, J. L., Davies, K. M., Zhou, Y., Deng, H. Y., Heaney, R. R. & Recker, R. R. (2000*a*). Genetic determination of Colles' fractures and differential bone mass in women with and without Colles' fractures. *J. Bone. and Mineral Research* **15**, 1243–1252.

Deng, H. W., Chen, W. M. & Recker, R. R. (2000*b*). QTL fine mapping by measuring and testing for Hardy–Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of populations. *Am. J. Hum. Genet.* **66**, 1027–1045.

Deng, H. W., Chen, W. M. & Recker, R. R. (2001). Population admixture: detection by Hardy–Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. *Genetics*, **157**, 885–897.

Edwards, J. H. (1980). Exclusion mapping. *J. Med. Genet.* **24**, 539–543.

Eisman, J. A. (1995). Vitamin D receptor gene alleles and osteoporosis: An affirmative view. *J. Bone. Miner. Res.* **10**, 1289–1293.

Gelernter, J., Goldman, D. & Risch, N. (1993). The A1 allele at the D2 dopamine receptor gene and alcoholism. *JAMA* 269, 1673–1677.

Gennari, L., Becherini, L., Mansani, R., Masi, L., Falchetti, A., Morelli, A., Colli, E., Gonnelli, S., Cepollaro, C. & Brandi, M. L. (1999). FokI polymorphism at translation initiation site of the vitamin D receptor gene predicts bone mineral density and vertebral fractures in postmenopausal Italian women. *J. Bone. Miner. Res.* **14**, 1379–86.

Gomez, C., Naves, M. L., Barrios, Y., Diaz, J. B., Fernandez, J. L., Salido, E., Torres, A. & Cannata, J. B. (1999). Vitamin D receptor gene polymorphisms, bone mass, bone loss and prevalence of vertebral fracture: differences in postmenopausal women and men. *Osteoporosis Intl*. **10**, 175–82.

Gong, G., Stern, S., Cheng, S. C., Fong, F., Mordeson, N., Deng, H. W. & Recker R. R. (1999). On the association of bone mass density and Vitamin-D genotype polymorphisms. *Osteoporosis Intl*. **9**, 55–64.

Hanis, C. L., Boerwinkle, E., Chakraborty, R., Ellsworth, D. L. & Bell, G. I. (1996). A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genetics* **13**, 161–166.

Hartl, D. L. & Clark, A. G. (1989). *Principles of population genetics*. Sinauer Associates, Sunderland, MA.

Holden, C. (1994). A cautionary genetic tale: the sobering story of D2. *Science* **264**, 1696–1697.

Houston, L. A., Grant, S. F., Reid, D. M. & Ralston, S. H. (1996). Vitamin D receptor polymorphism, bone mineral density, and osteoporotic vertebral fracture: studies in a UK population. *Bone*. **18**, 249–52.

Khoury, M. J., Beaty, T. H. & Cohen, B. H. (1993). Fundamentals of Genetic Epidemiology Oxford University Press, NY.

Kruglyak, L. & Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitaitve traits. *Am. J. Hum. Genet.* **57**, 439–454.

Lander, E. S. & Schork, N. J. (1994). Genetic dissection of complex traits. *Science* 265, 2037–2048.

Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting the reporting results. *Nature Genetics* **11**, 241–247.

Lynch, M. & Spitze, K. (1994). Evolutionary genetics of Daphnia. Pp. 109–128 in *Ecological Genetics*, edited by Real L, Princeton University Press, Princeton, NJ.

Morton, N. E., Collins, A. (1999). Tests and estimates of allelic association in complex inheritance *Proc. Natl. Acad. Sci. USA* **95**, 11389–11393.

Nei, M. (1987). Molecular population genetics and evolution. North-Holland/American Elsevier, Amsterdam.
Ott, J. (1999). Analysis of human genetic linkage. 3rd Ed. Johns Hopkins University Press, Baltimore.
Pato, C. N., Macciardi, F., Pato, M. T., *et al* (1993). Review of the putative association of dopamine D2 receptor and alcoholism: a meta-analysis. *Am. J. Med. Genet.* **48**, 78–82.
Peacock, M. (1995). Vitamin D receptor gene alleles and osteoporosis: A contrasting view. *J. Bone. Miner. Res.* **10**, 1294–1297.
Pritchard, J. K., Stephens M., Rosenberg N. A. & Donnelly P. (2000). Association mapping in structured populations. *Am J Hum Genet.* **67**, 170–181.
Risch, N. & Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA polling. *Genome Research* 8, 1273–1288.
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). *Am. J. Hum. Genet.* **52**, 506–516.
Vandevyver, C., Vanhoof, J., Declerck, K., Stinissen, P., Vandervorst, C., Michiels, L., Cassiman, J. J., Boonen, S., Raus, J., Geusens, P. (1999). Lack of association between estrogen receptor genotypes and bone mineral density, fracture history, or muscle strength in elderly women. *J. Bone. Miner. Res.* **14**, 1576–82.
Weir, B. S. (1996). *Genetic data analysis II*. Sinauer, Sunderland, MA.
Wolfram, S. (1996). *The Mathematica*. Wolfram Research, Inc., Champaign, IL.