Journal of Human Genetics (2017) 62, 191–198

www.nature.com/jhg

ORIGINAL ARTICLE

Estimation of the risk of a qualitative phenotype: dependence on population risk

Naoyuki Kamatani[1], Shigeo Kamitsuji[1], Yasuaki Akazawa[2], Takashi Kido[3] and Masanori Akita[4]

Individual disease risk estimated based on the data from single or multiple genetic loci is generally calculated using the genotypes of a subject, frequencies of alleles of interest, odds ratios and the average population risk. However, it is often difficult to estimate accurately the average population risk, and therefore it is often expressed as an interval. To better estimate the risk of a subject with given genotypes, we built R scripts using the R environment and constructed graphs to examine the change in the estimated risk as well as the relative risk according to the change of the average population risk. In most cases, the graph of the relative risk did not cross the line of $y=1$, thereby indicating that the order of the relative risk for given genotypes and the population average risk does not change when the average risk increases or decreases. In rare cases, however, the graph of the relative risk crossed the line of $y=1$, thereby indicating that the order of the relative risk for given genotypes and the population average risk does change owing to the change in the population risk. We propose that the relative risk should be estimated for not only the point average population risk but also for an interval of the average population risk. Moreover, when the graph crosses the line of $y=1$ within the interval, this information should be reported to the consumer.
Journal of Human Genetics (2017) **62,** 191–198; doi:10.1038/jhg.2016.106; published online 25 August 2016

INTRODUCTION

Personal genome tests have been offered directly to individual consumers since 2007.[1] These data can be used to estimate the individual risk of diseases using the genotypes at single-nucleotide polymorphisms reported to be associated with diseases. However, genetic risk models based on known single-nucleotide polymorphisms typically have only low to moderate predictive ability for most diseases, because of the relatively low effect sizes of previously reported single-nucleotide polymorphisms.

Kalf *et al.*[2] compared various algorithms used by three private genetic risk service companies (23 and Me, deCODEme and Navigenics), and reported that the area under the curve (AUC) values of receiver operating characteristic curves differed between these companies even for the same given genotypes. In addition, previous reports showed that the predicted risks differed among companies and were divergent for some traits in some individuals.[3–6] Although previous reports suggested that the discriminative accuracy reflected by the area under the curve of the receiver operating characteristic curve using currently available single-nucleotide polymorphisms is not sufficiently high,[7–9] additional variations that have not yet been discovered along with more sophisticated algorithms may improve the accuracy of this method.

In the present study, we examined the characteristics of estimated risks based on individual genotypes from single and multiple loci to evaluate the validity of estimating such risks.

To estimate the risk of an individual to express a qualitative phenotype such as a disease based on single or multiple associated genetic loci, it is first necessary to determine the average risk in the population in addition to the population allele frequency and odds ratio of the association. However, it is often difficult to obtain an accurate average risk of a population. The risk is usually estimated either from the results of an epidemiological study or from a meta-analysis of multiple studies, and is expressed as an interval such as the 95% confidence interval rather than as a point estimation.

This type of interval estimation means that the calculated risk of a subject is likely to be influenced by any change in the average risk within the interval. According to such analyses, a graph of the estimated relative risk (y axis) against the average risk of the population (x axis) can be constructed. In the present context, a relative risk is defined as the individual risk divided by the average risk of the population. In general, it is more important to know whether an individual risk is higher or lower compared with the average risk of the population rather than estimating the absolute individual risk. Therefore, it is essential to determine whether the relative risk vs average risk graph crosses the line of $y=1$, and if so, to determine the point at which the average population risk (x) is equal to the estimated risk of the subject.

Here, we examine the conditions in which the graph of the estimated relative risk of a subject crosses the line of $y=1$, and propose methods to cope with that situation.

[1]Department of Data Analysis, StaGen, Tokyo, Japan; [2]Department of Electrical Engineering and Bioscience School of Advanced Science and Engineering, Waseda University, Tokyo, Japan; [3]Riken Genesis, Tokyo, Japan and [4]MTI, Tokyo, Japan
Correspondence: N Kamatani, Department of Data Analysis, StaGen, KUGA Building 8F, 4-11-6 Kuramae, Taito-ku, Tokyo 111-0051, Japan.
E-mail: kamatani@msb.biglobe.ne.jp
Received 30 June 2016; revised 10 July 2016; accepted 13 July 2016; published online 25 August 2016

MATERIALS AND METHODS

First, we describe the algorithm used to calculate individual risk based on genotypes examined in this study.

Estimating the individual risk based on a single-locus genotype

Among two alleles, *A* and *a*, at a given locus, we designate *a* as the allele of interest. Accordingly, the number of the alleles of interest in the genotype of a subject is 0, 1 or 2 for the genotypes *AA*, *Aa* and *aa*, respectively.

Let d_1, d_2 and d_3 be the absolute risks (e.g., the probability of developing a disease) of the subjects with the genotypes *AA*, *Aa* and *aa*, respectively. Let r_1 be the odds ratio of the risk for the comparison of genotypes *Aa* and *AA*, and let r_2 be the odds ratio of the risk for the comparison of genotypes *aa* and *Aa*. Then, because of the definition of the odds ratio, the following equations hold:

$$r_1 = \frac{d_2}{1-d_2} \Big/ \frac{d_1}{1-d_1}, \tag{1}$$

$$r_2 = \frac{d_3}{1-d_3} \Big/ \frac{d_2}{1-d_2}. \tag{2}$$

Let p denote the frequency of allele *a* and let m denote the average risk in the population, which is usually calculated from either the incidence or prevalence in the population. If Hardy–Weinberg's equilibrium is assumed, then the following equation holds:

$$m = (1-p)^2 d_1 + 2p(1-p)d_2 + p^2 d_3. \tag{3}$$

By removing the variables d_2, d_3 using Equations (1)–(3), the following equation is obtained, in which d_1 remains as a variable:

$$\begin{aligned} 0 = {} & -m + (1-p)^2(1-r_1)(1-r_1r_2)d_1^3 + [(3r_1 - r_1^2r_2 - 2)p^2 + 2(2 \\ & - 2r_1 - r_1r_2 + r_1^2r_2)p - 2 - m + r_1 + mr_1 + r_1r_2 + mr_1r_2 \\ & - mr_1^2r_2]d_1^2 + [(1 - 2r_1 + r_1r_2)p^2 - 2(1-r_1)p - mr_1r_2 - mr_1 \\ & + 2m + 1]d_1. \end{aligned} \tag{4}$$

Although this equation can be solved mathematically with respect to the variable d_1, the solution is too complex to present here.

When assigning values to p, m, r_1 and r_2, we can obtain the solution for d_1 using Cardano's method.[10]

Estimation of the risk based on genotypes for a single locus

After deriving an appropriate solution of d_1 using Cardano's method based on Equation (4) as described above, d_2 and d_3 can be derived using the following

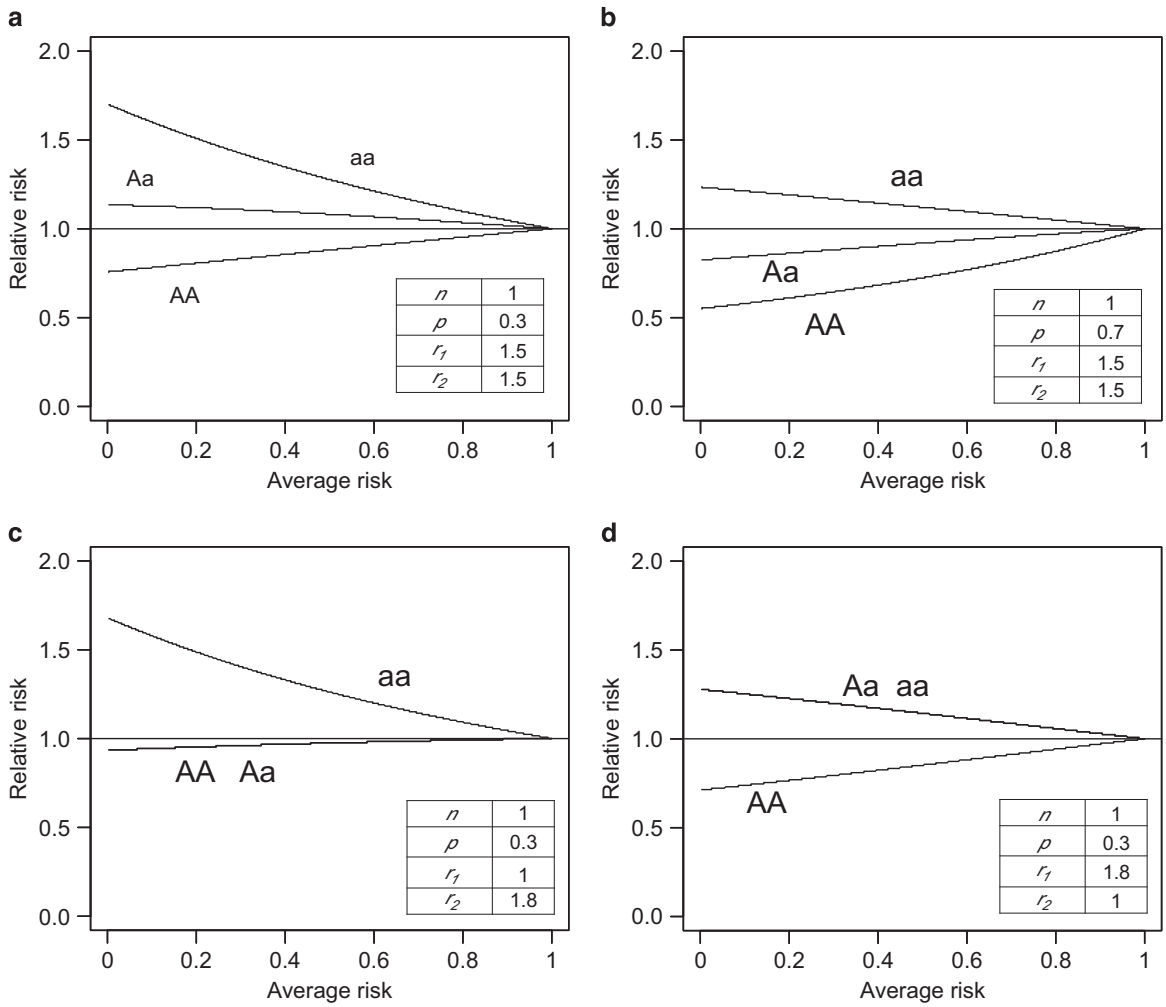


Figure 1 Relative risk of subjects with three different genotypes at a single locus calculated using the R script Singlelocus. R at varying population average risk values, *m*. The frequency *p* of the allele of interest *a* and two odds ratios, r_1 *and* r_2, were given as shown. *n* indicates the number of loci and $n=1$ was assumed to derive these graphs.

equations obtained from Equations (1) and (2):

$$d_2 = \frac{d_1 r_1}{1 - d_1 + d_1 r_1},$$

$$d_3 = \frac{d_2 r_2}{1 - d_2 + d_2 r_2}.$$

Thus, d_1, d_2 and d_3 can be obtained if the values of p, m, r_1 and r_2 are known. The relative risk of an individual with the genotype *AA*, *Aa* or *aa* as compared with the average risk of the population; that is, d_1/m, d_2/m or d_3/m can also be derived. From Equation (3), we obtain

$$1 = (1-p)^2 d_1/m + 2p(1-p)d_2/m + p^2 d_3/\mathrm{m}. \quad (5)$$

When $0 <= d_1 < d_2 < d_3 <= 1$, d_1/m and d_3/m must be $<$ or >1, respectively, for $0 < p < 1$.

The effects of changing the values of m, p, r_1 or r_2 on the relative risk of an individual were analyzed for each genotype within an appropriate interval, and graphs were constructed to examine these effects visually.

Estimation of the risk based on genotypes for multiple loci

The risk of a subject based on the data from multiple loci is calculated using the following multivariate logistic model:

$$\log\frac{P}{1-P} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \beta_0 + \sum_{i=1}^{n} \beta_i X_i, \quad (6)$$

where β_i, $(i = 1,2,\ldots,n)$ denote coefficients, and P denotes the variable for the risk. For the multilocus model, we assume that two odds ratios, r_1 and r_2, are equal (i.e., the effect of an allele is additive), and X_i denotes the number of the allele of interest a in the genotype at the ith locus; X_i is 0, 1 and 2 for genotypes *AA*, *Aa* and *aa*, respectively.

Since $\frac{P}{1-P}$ is the odds, the odds ratio of the risk for the comparison of genotypes *aa* (i.e., $x_i = 2$) and *Aa* (i.e., $x_i = 1$) is e^{β_i}, which is equal to r_i, denoting the odds ratio at the ith locus. Note that r_i is the same as r_1 and r_2 for the ith locus in Equations (1) and (2). Therefore,

$$\beta_i = \log\ r_{i.} \quad (7)$$

By solving Equation (6) with respect to P, we get the following logistic function:

$$P = 1/(1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}). \quad (8)$$

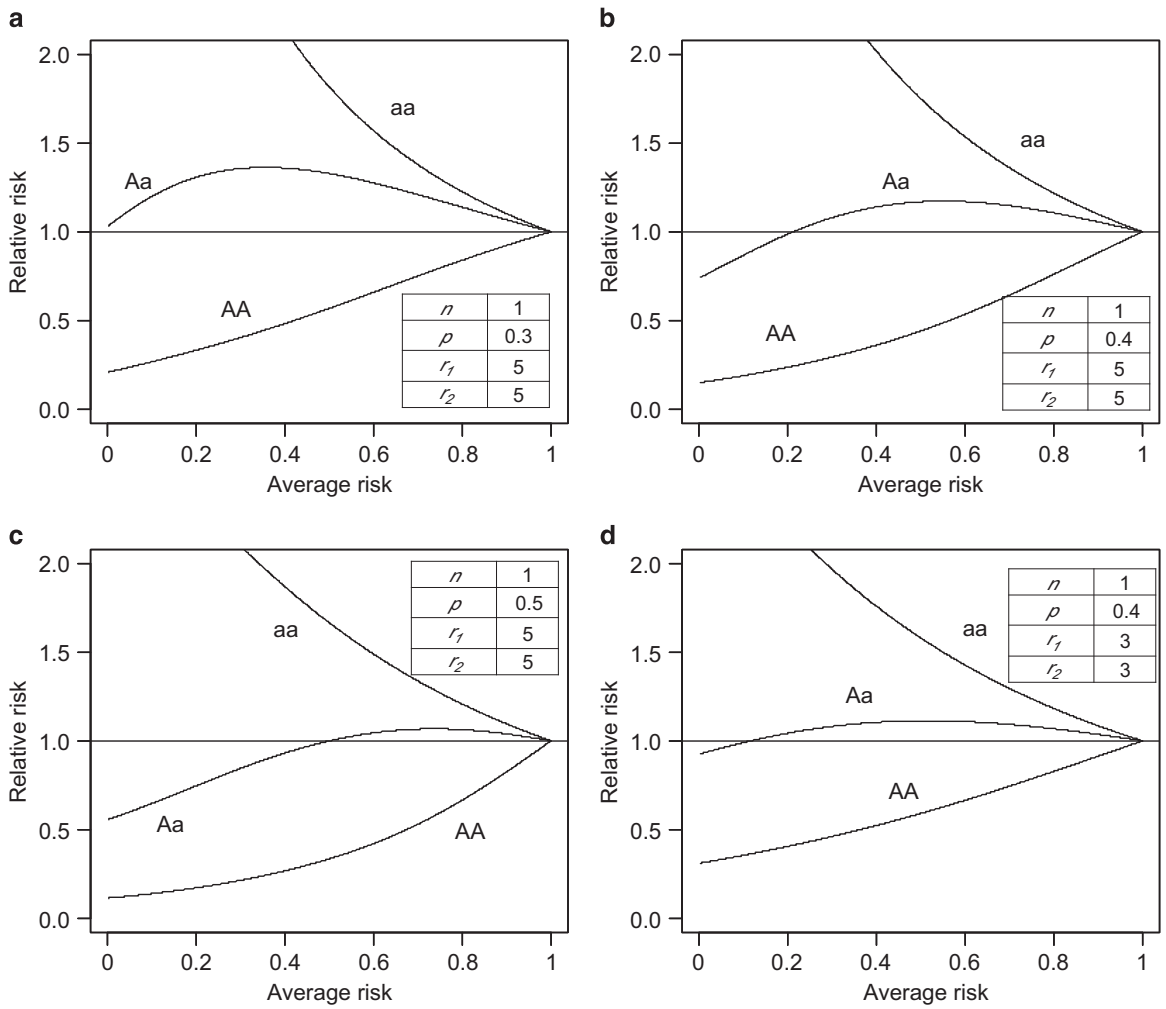


Figure 2 Relative risk of subjects with three different genotypes at a single locus calculated using the R script Singlelocus. R at varying population average risk values, m, but with higher values of the odds ratios r_1 *and* r_2. Other conditions and parameters are the same as described for Figure 1.

The average of P in the population is

$$E(P) = \sum_{s \in S} 1/(1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}) \prod_{i=1}^{n} h_i, \quad (9)$$

where $s = (x_1, x_2, \ldots, x_n)$, and S denotes the set of all s. In Equation (9),

$$h_i = \begin{cases} (1-p_i)^2 & (\text{for genotype } AA) \\ 2p_i(1-p_i) & (\text{for genotype } Aa) \\ p_i^2 & (\text{for genotype } aa) \end{cases}, \quad (10)$$

where p_i denotes the frequency of the allele of interest at the ith locus.

When β_i ($i = 1, 2, \ldots, n$) and p_i ($i = 1, 2, \ldots, n$) as well as $s = (x_1, x_2, \ldots, x_n)$ are given, the right-hand side of Equation (9) is a monotone increasing function with respect to β_0.

Therefore, if $E(P)$ is given, β_0 can be numerically determined by solving Equation (9).

When β_0 is determined, P for each given s, which is defined as P_s is determined using Equation (8), and the relative risk $P_s/E(P)$ for each subject based on the observed genotypes can be obtained using Equations (8) and (9).

Thus, the relative risk $P_s/E(P)$ based on a combination of multiple genotypes can be determined for different values of $E(P)$. Accordingly, the graph was drawn for $P_s/E(P)$ for the given values of $m = E(P)$ between 0 and 1.

RESULTS

Effect of the average population risk on the relative risks for different genotypes at a single locus

We developed an R script named Singlelocus.R (Supplementary Material 1) to solve Equations (1)–(3) for determining d_1, d_2 and d_3 (penetrance parameters for genotypes *AA*, *Aa* and *aa*, respectively) from the odds ratios r_1 and r_2, the population frequency p of the allele of interest a and the average population risk m. First, we determined d_1 by solving Equation (4), and then determined d_2 and d_3. This R script also draws the curves of d_1, d_2 and d_3 as a function of the average population risk m.

Using this R script, we examined the effect of changes in the average population risk (m) on the relative risks for the different genotypes (*AA*, *Aa*, *aa*) based on the data from a single locus. Figures 1 and 2 show the results with various values of p, r_1 and r_2. All of the graphs reached a line of $y = 1$ when m was 1, where y denotes the relative risk (Figures 1a–d and 2a–d). In all cases, the order of the relative risks for the genotypes *AA*, *Aa*, *aa* was $AA \leq Aa \leq aa$, and the relative risks of

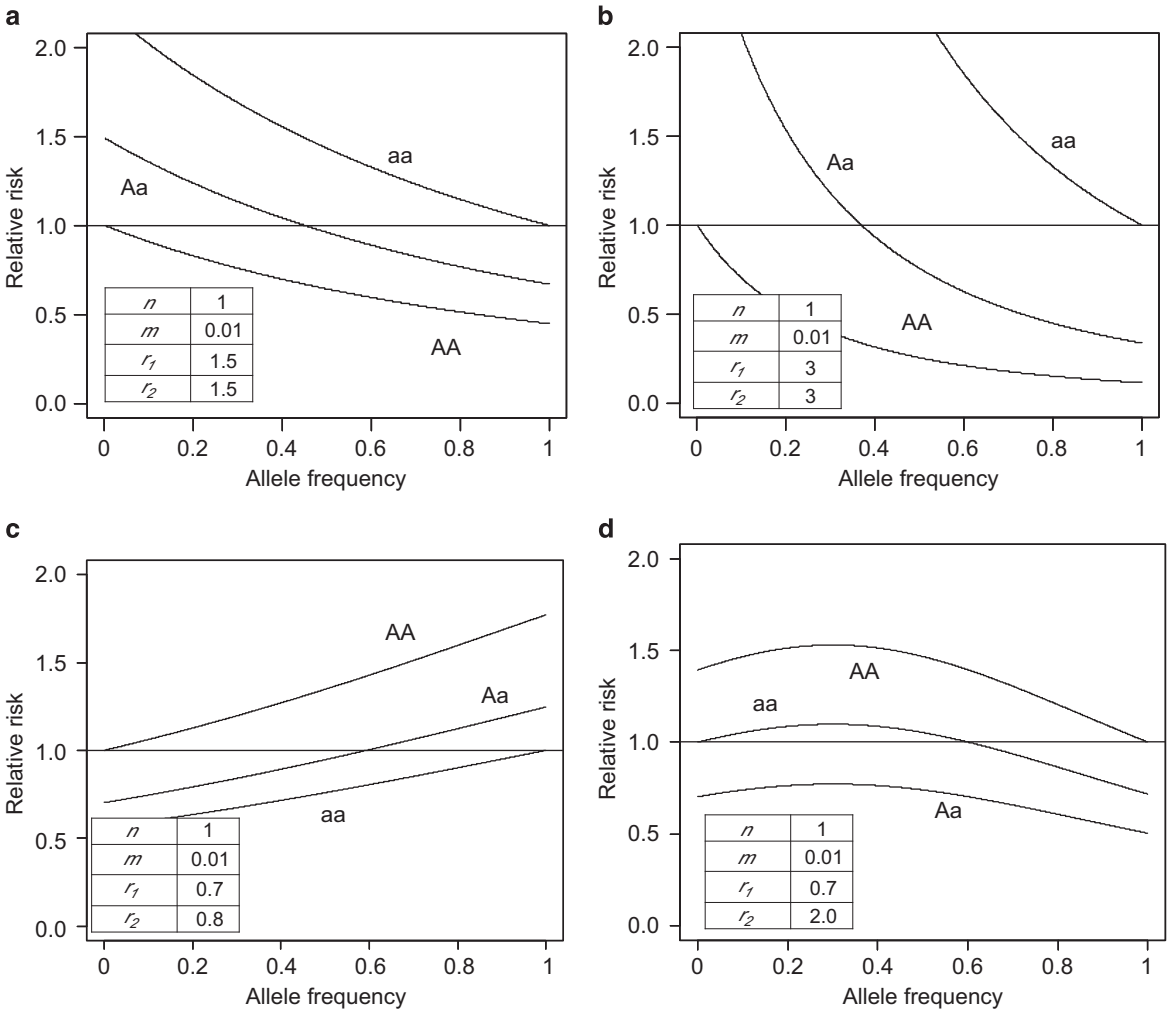


Figure 3 Relative risk of subjects with three different genotypes at a single locus calculated by the R script Singlelocus. R with varying frequencies of the allele of interest *a*. The average population risk *m* and two odds ratios, r_1*and* r_2, were given as shown. *n* indicates the number of loci and $n = 1$ was assumed to derive these graphs.

AA and *Aa* were equal when $r_1 = 1$ (Figure 1c), whereas those of *Aa* and *aa* were equal when $r_2 = 1$ (Figure 1d).

Figures 1a–d indicates that under the applied conditions (r_1 and r_2 are close to 1), the relative risks of the individuals change in an almost linear manner because of the change of the average risk, and the differences between different genotypes tend to decrease when the average risk increases. Thus, when $r_1 > 1$ and $r_2 > 1$, the relative risk for *AA* tends to increase, whereas that for *aa* tends to decrease when the average risk increases (Figures 1a and b). The relative risk for *Aa* tends to increase when it is lower than 1, whereas it tends to decrease when it is higher than 1 (Figures 1a and b). None of the lines for the relative risks of different genotypes crossed the horizontal line of 1.0, indicating that the order of relative risk of an individual and the population average risk does not change when the average risk changes (Figures 1a–d).

When r_1 and r_2 are rather high, the relationship between the average risk and the relative risks of the subjects with different genotypes are no longer nearly linear (Figures 2a–d). Furthermore, the relative risk for genotype *Aa* changes from lower to higher compared with the average when the average risk increases (Figures 2b–d), as reflected by the fact that the graph crosses the line of $y = 1$.

Effect of the frequency of the allele of interest on the relative risks for different genotypes based on single-locus data

We next examined the effect of the frequency of the allele of interest on the relative risks for different genotypes, and the results are shown in Figures 3a–d. With odds ratios of $r_1 > 1$ and $r_2 > 1$, the relative risks for all genotypes tend to decrease when the frequency of the allele of interest increases (Figures 3a and b). By contrast, with odds ratios of $r_1 < 1$ and $r_2 < 1$, the relative risks for all genotypes tend to increase when the frequency of the allele of interest increases (Figure 3c). However, when $r_1 = 0.7$ and $r_2 = 2.0$ (i.e., overdominance), the lines neither increase nor decrease monotonously, but instead show peaks between 0 and 1 (Figure 3d).

Effect of the odds ratio on relative risks for different genotypes based on single-locus data

We also examined the effect of the odds ratio on the relative risks for different genotypes, and the results are shown in

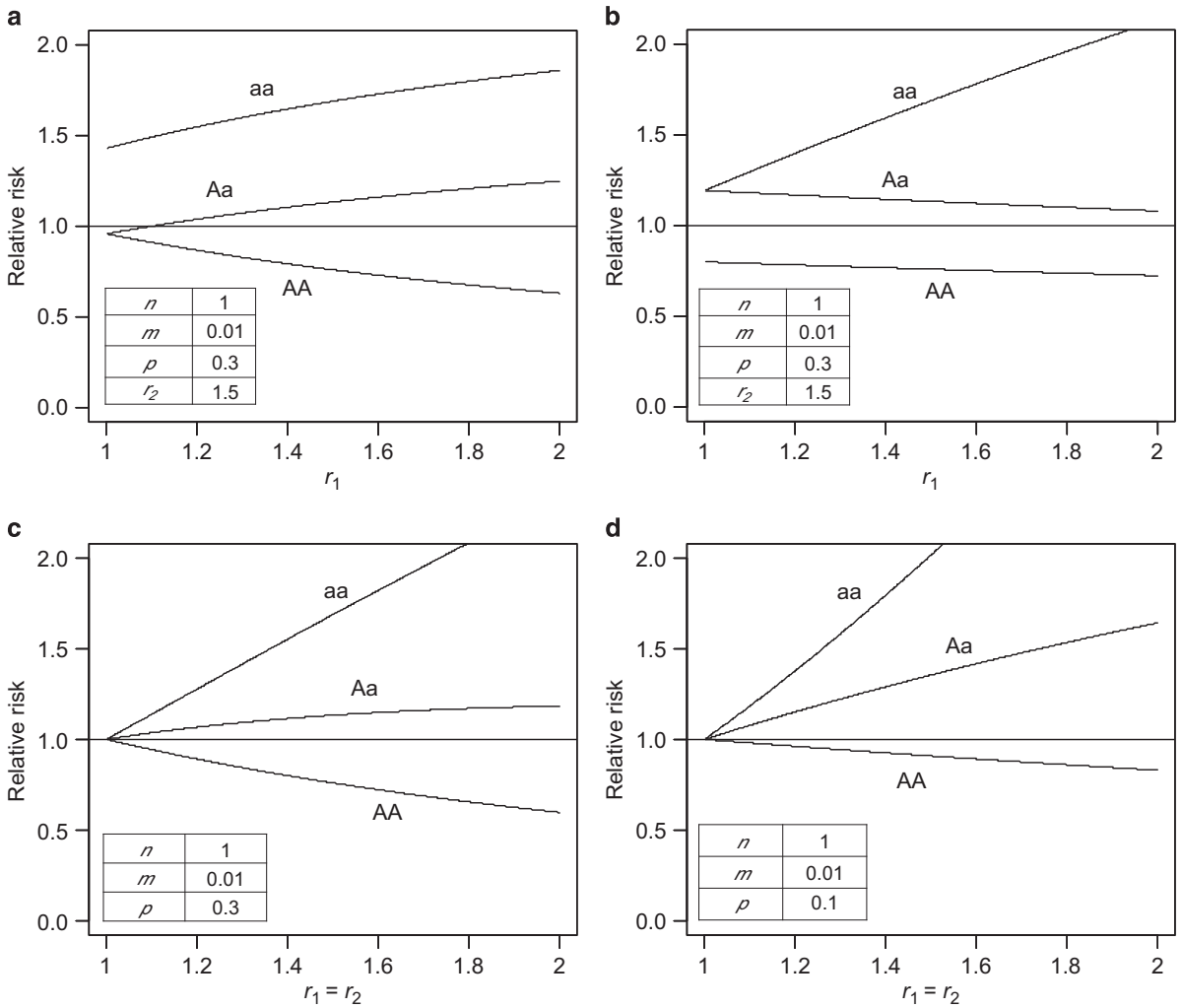


Figure 4 Relative risk of subjects with three different genotypes at a single locus calculated by the R script Singlelocus. R at varying odds ratios, r_1 and r_2. The average population risk *m* and the frequency *p* of the allele of interest *a* were given as shown. *n* indicates the number of loci and $n = 1$ was assumed to derive these graphs.

Figures 4a–d. These graphs indicate that the relative risks for genotype *AA* decrease, whereas those for genotype *aa* increase when the odds ratio increases (Figures 4a–d). The change in the relative risk for genotype *Aa* in response to changes in the odds ratio depended on the specific conditions (Figures 4a–d).

Estimation of the risk based on genotypes at multiple loci using R scripts

For calculation of the risk based on multiple loci, an R script, MultilocusSubject.R (Supplementary Material 2), was developed according to given frequencies of the allele of interest, odds ratios based on the additive model and genotypes of the subject at multiple loci, as well as the average population risk m. This script also calculates the relative risk in comparison with the average risk.

We also developed another R script, MultilocusCurve.R (Supplementary Material 3), to draw a graph showing the change in the individual relative risk because of the change of the average risk m. We performed an extensive simulation by inputting a variety of data to MultilocusCurve.R. All of the graphs reach $y=1$ when m reaches 1 (Figures 5–7). In general, the relative risk either increases or decreases monotonously when m increases from 0 to 1, and it finally reaches 1 when m is equal to 1 (Figures 5a,b,6a,c and 7b). Therefore, in these cases, the relative risk reaches 1 only when $m=1$. When $r_i>1$ and all loci have the genotype *AA*; that is, $x_i=0$, the relative risk is always below 1, and increases monotonously to reach 1 when $m=1$ (Figure 5b). However, when $r_i>1$ and all loci have the genotype *aa*; that is, $x_i=2$, the relative risk is always above 1, and decreases monotonously to reach 1 when $m=1$ (Figure 5a). The graph does not cross the line of $y=1$ in any of these cases.

In rare cases, however, the graph does cross the line of $y=1$ in the interval $0<m<1$ and reaches 1 at $m=1$, similar to the case of using data from a single locus (Figures 5c,d and 6b,d). This phenomenon occurred irrespective of the number of loci

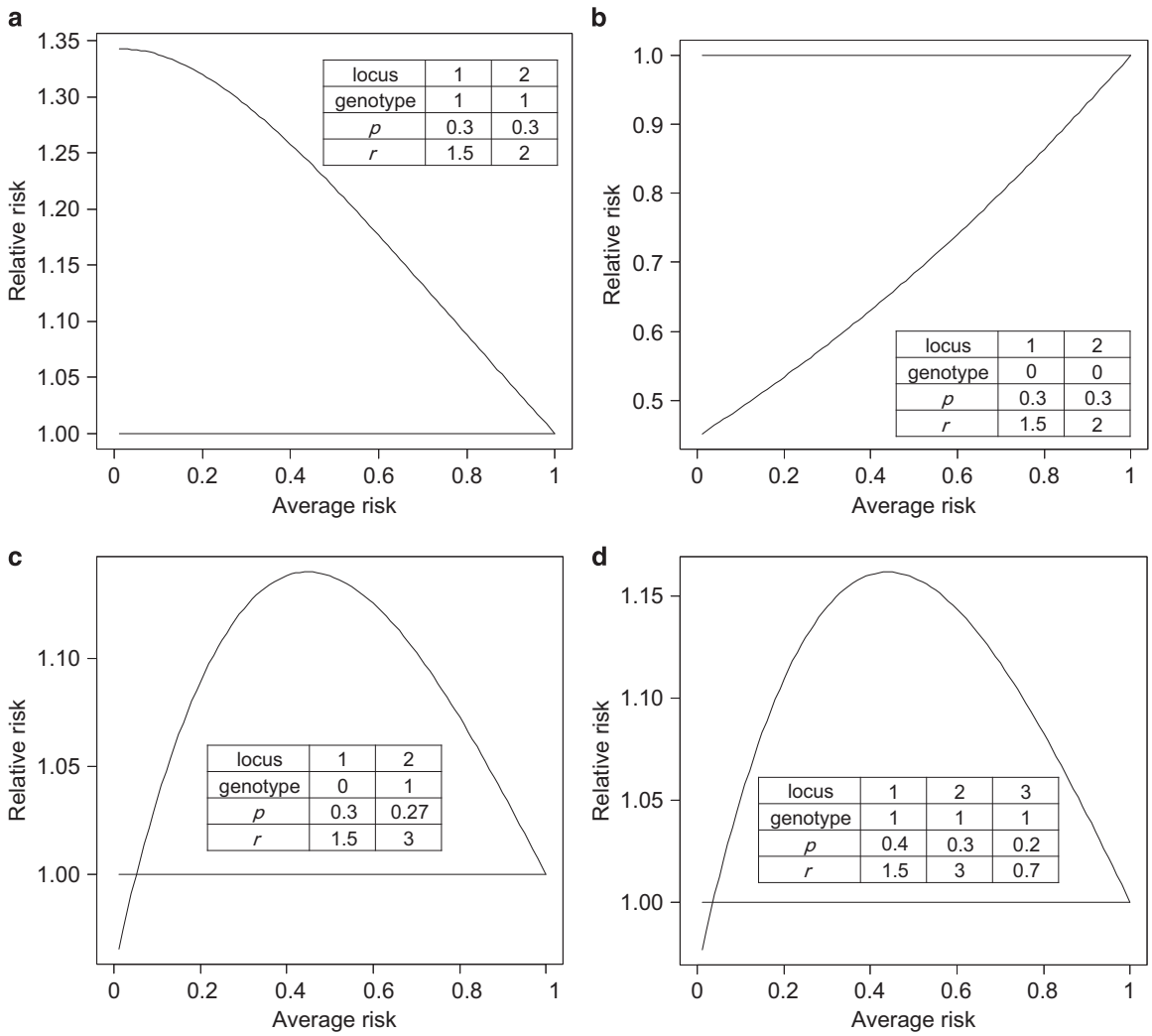


Figure 5 Relative risk of subjects with different genotype frequencies at multiple loci calculated using the R script Multilocus Curve. R with varying average population risk values, m. The genotype is expressed according to the number of the allele of interest a; that is, 0 for *AA*, 1 for *Aa* and 2 for *aa*. The frequency p of the allele of interest a, and a single odds ratio r for each locus were given as shown. The loci were numbered, and two or three loci were assumed for the construction of these graphs.

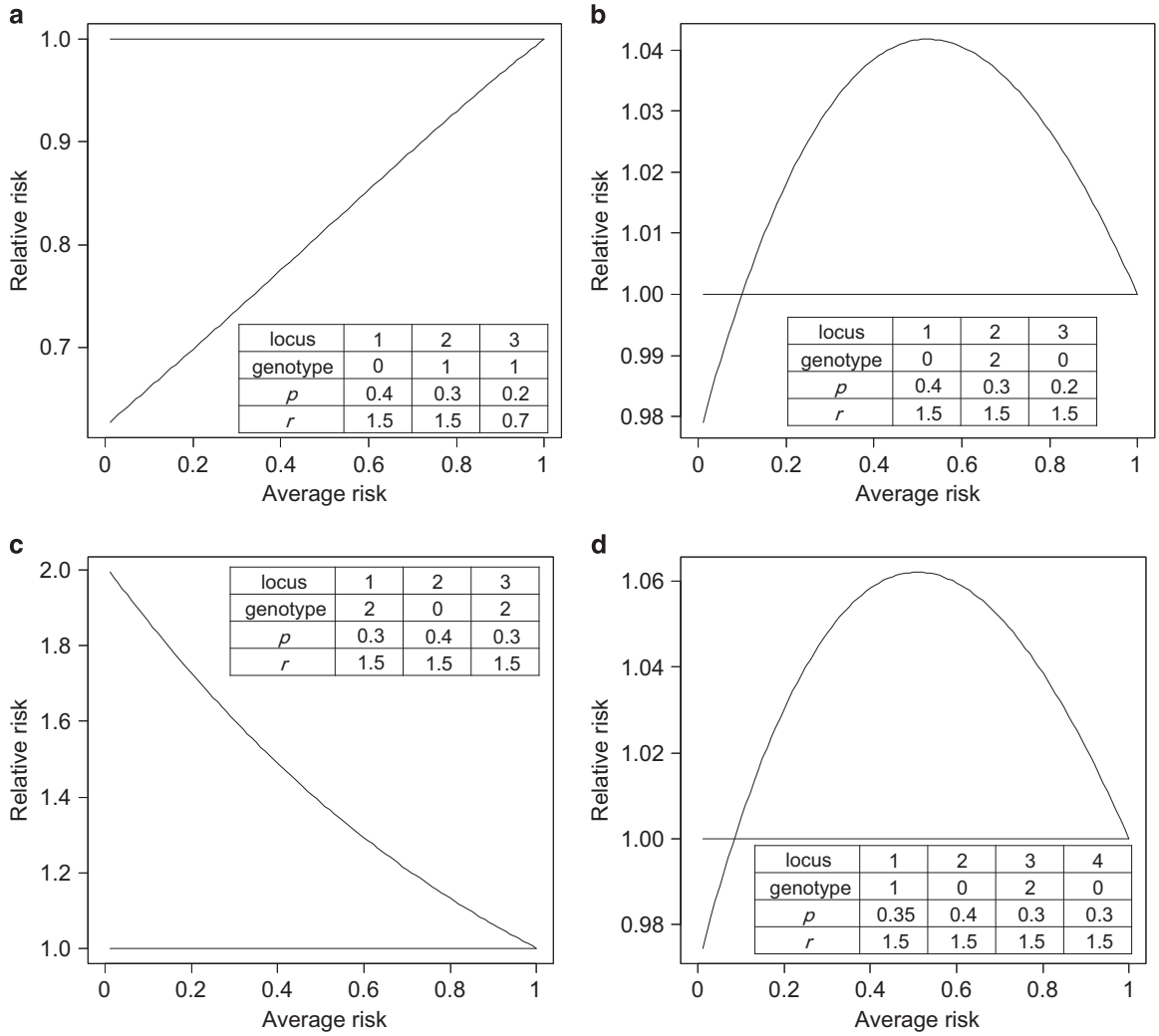


Figure 6 Relative risk of subjects with different genotype frequencies at multiple loci calculated using the R script Multilocus Curve. R with varying average population risk values, *m*. All parameters are the same as those described for Figure 5, except that three or four loci were assumed in these cases.

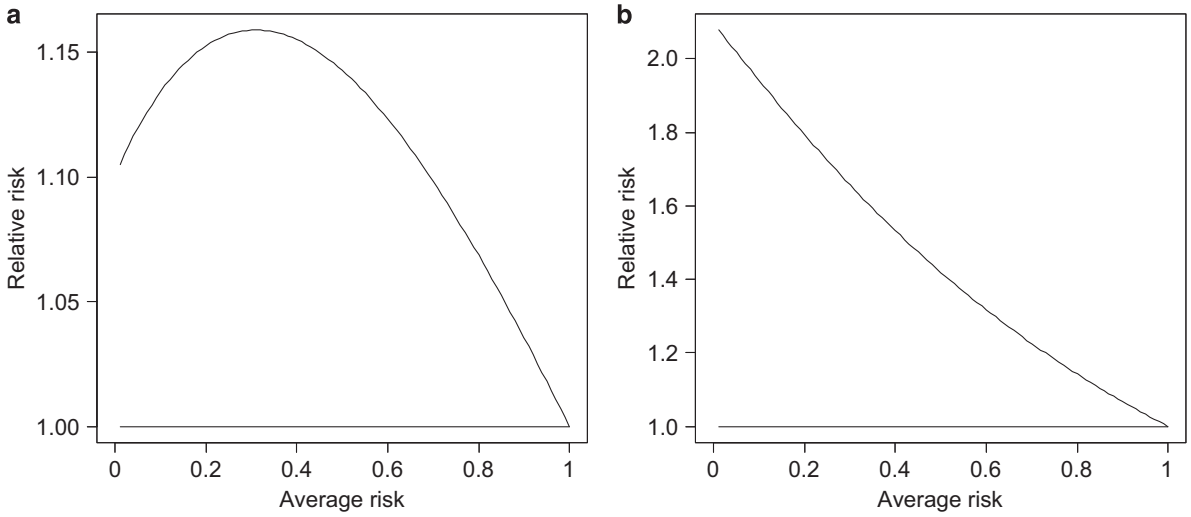
considered. The graph tended to cross the line of $y=1$ within the interval $0<m<1$ when some of the genotypes were *Aa*; that is, $x_i=1$. However, even when none of the loci had the genotype *Aa*; that is, $x_i=1$, the relative risk still crossed the line of $y=1$ (Figure 6b).

We implemented the algorithm to determine the specific value at which the relative risk crosses the line $y=1$; that is, the accurate value of *m* where the relative risk of an individual is equal to 1, using the bisection method[11] in MultilocusCurve.R.

DISCUSSION

In this study, we used the R environment (R version 2.15.0 The R Foundation for Statistical Computing, ISBN 3-900051-07-0 Platform: i386-pc-mingw32/i386) to successfully implement a system for estimating the risk of a subject given known allele frequencies, odds ratios and genotypes of the subject at multiple loci, in addition to the average population risk *m*. For estimation of the risks based on the genotypes at multiple loci, we assumed the additive model for the effect of the allele of interest. We found that the individual relative risk of the genotype *Aa* crosses the line of $y=1$ when the allele frequency *p* changes. This is not expected to cause a major problem in interpretation or analysis because the estimation of the allele frequency is often accurate. However, we also found that the estimated relative risk can cross the line of $y=1$ in some rare cases when the average population risk changes. This may cause a problem because estimating the individual relative risk is often more important than the absolute risk, and the average population risk is sometimes obtained as an interval or an approximate value. Therefore, we propose that the relative risk should be estimated for an interval of average risk values *m*, followed by an examination of whether the risk becomes lower or higher compared with the average within the interval. If the relative risk crosses the line of $y=1$ within the interval, we recommend that the interpretation should be reported as either 'the relative risk cannot be estimated' or 'the relative risk becomes higher or lower than the average risk of the population at the value *x*'.

A limitation of this study is that a non-consistent message may be acceptable for a hererozygote at a single locus; however, similar messages based on multiple loci may not be acceptable for some subjects. Another limitation is that the risk of a subject is largely influenced by the factors other than the observed genotypes.



locus	1	2	3	4	5
genotype	1	0	2	0	1
p	0.35	0.4	0.3	0.3	0.3
r	1.5	1.5	1.5	1.5	1.5

locus	1	2	3	4	5
genotype	1	1	2	0	1
p	0.3	0.2	0.3	0.3	0.3
r	1.5	0.7	1.5	0.7	1.5

Figure 7 Relative risk of subjects with different genotype frequencies at multiple loci calculated using the R script Multilocus Curve. R with varying average population risk values, *m*. All parameters are the same as those described for Figure 5, except that five loci were assumed in these cases.

Therefore, the usefulness of the estimation of the risk based on the genotypes of limited numbers of loci is limited.

CONFLICT OF INTEREST

NK received employment fee from StaGen, fee as an advisor from SRL, fee as an advisor from Teijin Pharma and have stocks of StaGen and SmartMed. TK received employment fee from Riken Genesis, SK received employment fee from StaGen and MA received employment fee from MTI.

ACKNOWLEDGEMENTS

We thank Kenji Suzuki for useful advice.

1 Helgason, A. & Stefansson, K. The past, present, and future of direct-to-consumer genetic tests. *Dialog. Clin. Neurosci.* **12**, 61–68 (2010).
2 Kalf, R. R., Mihaescu, R., Kundu, S., de Knijff, P., Green, R. C. & Janssens, A. C. Variations in predicted risks in personal genome testing for common complex diseases. *Genet. Med.* **16**, 85–91 (2014).
3 United States Government Accountability Office. *Direct-to-Consumer Genetic Tests: Misleading Test Results are Further Complicated by Deceptive Marketing and other Questionable Practices*, 2010. Available at: www.gao.gov/assets/130/125079.pdf (last, Accessed 1 January 2012.
4 Ng, P. C., Murray, S. S., Levy, S. & Venter, J. C. An agenda for personalized medicine. *Nature* **461**, 724–726 (2009).
5 Imai, K., Kricka, L. J. & Fortina, P. Concordance study of 3 direct-to-consumer genetic-testing services. *Clin. Chem.* **57**, 518–521 (2011).
6 Kido, T., Kawashima, M., Nishino, S., Swan, M., Kamatani, N. & Butte, A. J. Systematic evaluation of personal genome services for Japanese individuals. *J. Hum. Genet.* **58**, 734–741 (2013).
7 Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
8 Moonesinghe, R., Liu, T. & Khoury, M. J. Evaluation of the discriminative accuracy of genomic profiling in the prediction of common complex diseases. *Eur. J. Hum. Genet.* **18**, 485–489 (2010).
9 Miyake, K., Yang, W., Hara, K., Yasuda, K., Horikawa, Y., Osawa, H. *et al.* Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. *J. Hum. Genet.* **54**, 236–241 (2009).
10 Burton, D. *The History of Mathematics: An Introduction* 7th edn. (McGraw-Hill, 2010).
11 Monahan, J. F. in *Handbook of Computational Statistics* (eds Gentle J. E., Hardle W. K. & Mori Y.) 30 (Springer, 2012).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)