Contribution of rare variants to heritability of a disease is much greater than conventionally estimated: modification of allele distribution model

Yoshiro Nagao[1,2]

[1]Department of Clinical Genetics, Tokai University Hospital, Shimokasuya 143, Isehara, Kanagawa, Japan

[2]Department of Laboratory Examination, Takashimadaira Chuo General Hospital, Itabashi, Tokyo, Japan

Correspondence and requests for materials should be addressed to Y.N. (email: yn-nagao@cool.odn.ne.jp)

Missing heritability is a current problem in human genetics. I previously reported a method to estimate heritability of a polymorphism (h_p^2) for a common disease without calculating the genetic variance under dominant and the recessive models. Here, I extended the method to the co-dominant model and carry out trial calculations of h_p^2. I also calculated h_p^2 applying the allele distribution model originally reported by Pawitan et al. for a comparison. Unexpectedly, h_p^2 calculated for rare variants with high odds ratios was much higher. I noticed that conventional methods use the allele frequency (AF) of a variant in the general population. However, this implicitly assumes that the unaffected are included among the phenotypes: an assumption that is inconsistent with case-control studies in which unaffected individuals belong to the control group. Therefore, I modified the allele distribution model by using the AF in the patient population. Consequently, the h_p^2 of rare variants was quite high. Recalculating h_p^2 of several rare variants reported in the literature with the modified allele distribution model, yielded results were 3.2 - 53.7 times higher than the original model. These results suggest that the contribution of rare variants to heritability of a disease has been considerably underestimated.

Genome-wide association studies (GWAS) for diseases are based on case-control studies[1]. GWAS have identified thousands of genetic polymorphisms associated with common diseases, however, every effort to account for more than a fraction of the heritability of the disease by the discovered variants has failed. This is called the missing heritability problem[2,3].

Heritability (h^2) is a concept that summarizes how much of the variation in a trait is due to variation in genetic factors[4]. The phenotypic variance in the trait (V_P) is the sum of genetic variance (V_G) and environmental variance (V_E) as follows:

$$V_P = V_G + V_E \tag{1}$$

h^2 is defined as:

$$h^2 = V_G / V_P \tag{2}$$

Considering a specific disease as a trait here, V_P and V_G represent the variations of phenotypes and genotypes of the disease in the population, respectively.

In conventional studies estimating the contribution of variants to h^2 a co-dominant model is commonly assumed, and various estimation methods have been proposed[5,6]. Among them, the method of Pawitan et al. is simple. According to this formulation, when n variants are associated with a disease the variance of the risk distribution, $V(g)_k$, for the k-th variant is represented by its allele frequency (AF) in the general population, p_k, and odds ratio, OR_k, by the following

equation:

$$V(g)_k = 2p_k (1 - p_k) \{\log_e (OR_k)\}^2 \quad (3)$$

$V(g)_k$ is considered as the genetic variance of the k-th variant, and there should be the following relation between $V(g)_k$ and V_G.

$$V_G = \sum_k V(g)_k \quad (4)$$

h^2 is evaluated by the following equation:

$$h^2 = V_G / (V_G + \pi^2/3) \quad (5)$$

Oliynyk called the method using the equations (3), (4) and (5) the "allele distribution model" [7].

In a previous report, I presented a method to estimate h^2 of a polymorphism (h_p^2) by probabilistically predicting how the risk genotype of a variant is inherited by the relatives of a patient under dominant and the recessive models[8]. In the present study, I extend the approach to the co-dominant model and do trial calculations of h_p^2 for variants with various ORs to assess the results compared to the standard method.

Results

Definitions and premises

- p is the AF of the risk allele of a variant in the general population

- q is the AF of the non-risk allele of a variant in the general population.
- u is the AF of the risk allele of a variant in the patient population.
- v is the AF of the non-risk allele of a variant in the patient population
- **P** is the Prevalence of a disease.
- **Q** is the recurrence risk of the disease among the first-degree relatives of a patient

Derivation of the allele frequency of a variant in the patient population

If the frequencies of the risk allele and the non-risk allele in the unaffected group are x and y, respectively, the following equation holds[8]:

$$\mathbf{P}u + (1 - \mathbf{P})x = p \quad (6)$$

$$\mathbf{P}v + (1 - \mathbf{P})y = q \quad (7)$$

When OR is represented as k, the following equation holds:

$$k = buy/(vx) \ (= OR) \quad (8)$$

Simultaneous equations (8), (9) and (10) cannot be solved analytically, but they can be solved numerically using the following approximation:

$$x \approx cp / (1 - \mathbf{P}) \quad (9)$$

In this relation, c is a constant with a value between 1 and 0. The approximate solution for u then becomes

$$u = 1 - [(\mathbf{P} - ckp - q) - \{ (\mathbf{P} + ckp + q)^2 - 4\mathbf{P}q \}^{1/2}] / (2\mathbf{P}) \qquad (10)$$

The value of c is estimated by checking the OR'. Allowing for an error of +/- 1 %, the value of OR'/OR should be between 0.99 and 1.01.

$$OR' = u(q - \mathbf{P}v) / \{v(p - \mathbf{P}u)\} \qquad (11)$$

Derivation of the equation for the "mutation model"

Assuming that a polymorphism (variant) is the only genetic factor for a disease, **Q** is represented, for a variant that acts in a co-dominant manner, by Equation (12):

$$\mathbf{Q} = \mathbf{P} + \mathbf{P}\,(1 - v/q)\,\{(p + u) / (2p) - 1\} \qquad (12)$$

Once **Q** is estimated, h_p^2 is calculated by the Falconer liability threshold model[8]. [Details of the calculation are given in Falconer[9].] Equation (12) assumes that the non-risk genotype has no genetic effect; in other words, the variant is regarded as a mutation. Indeed, substituting **P** = p and u = 1 into Equation (12) yields a **Q** of about 0.5, which is the incidence in first-degree relatives when the patient is homozygote for a Mendelian disease with autosomal co-dominant inheritance. Therefore, Equation (12) describes what can be called the "mutation model".

Application of the equation of the allele distribution model to calculate h_p^2

Equation (5) is a formula for to estimate the heritability of all variants taken together;

however, it can also be used to estimate the contribution of individual variants to h^2. The following equations are derived to estimate h_p^2 for a variant that has a MAF of p and an odds ratio of OR:

$$V(g) = 2p\,(1 - p)\,\{\log_e(OR)\}^2 \tag{13}$$

$$h_p^2 = V(g) / \{V(g) + \pi^2/3\} \tag{14}$$

Comparison of the h_p^2 calculated with the allele distribution model and with the mutation model

The calculated values of the h_p^2 by the two methods can now be compared. In Fig. 1, h_p^2 calculated by the allele distribution model (h_1^2) or by the mutation model (h_2^2), are shown for various values of OR (Figs.1a-1d). In the calculation of h_2^2, **P** was set to 0.01 (as in the case of schizophrenia, e.g., Schultz et al.[10]). h_1^2 generally has a greater value than h_2^2, however, when OR is high and AF is low, h_2^2 is greater than h_1^2. As shown in Fig. 1b, h_2^2 is always lower than h_1^2 when OR is 4, but when OR is 10, as shown in Fig. 1c, h_2^2 is higher than h_1^2 at low AF.

Why are the calculated values of h_p^2 by the two methods so different? Looking at Equation (3), we can see that $V(g)_k$ is calculated using the AF of a variant in the general population, p_k. Therefore, V_G defined by Equation (4) is also the genetic variance of a disease in the general population. If Equation (2) holds, V_P should also be the phenotypic variance of a disease in the

general population. But of course, the general population includes the large number of individuals who do not have the disease, and the conventional method has calculated $V(g)_k$ of k-th variant assuming that the phenotypes of a disease include those of all the unaffected individuals. This assumption is, however, inconsistent with the established rule that in a case-control study, unaffected individuals belong to the control (unaffected) group. I therefore propose here that V_P in Equation (4) should be the phenotypic variance of a disease in the patient population. Therefore, V_G should be the genotypic variance in the patient population and the $V(g)_k$ of the k-th variant should be calculated using the AF in the patient cohort.

Derivation of the modified equation for the allele distribution model

Pawitan et al. assume that the proportion of SNPs associated with a certain disease is normally distributed when plotted against the log scale value of OR for each SNP on the x-axis (Fig. 2)[5]. In their model the mean value of polygenic risk score (β_{mean}) in the general population is represented by the following equation[7]:

$$\beta_{mean} = 2p_1\log_e OR_1 + 2p_2\log_e OR_2 + \ldots\ldots\ldots + 2p_n\log_e OR_n \qquad (15)$$

In this equation, p_k and OR_k are minor allele frequency (MAF) and OR of the k-th SNP, respectively. What is the expected for β_{mean} in the general population? Under the assumption that $\log_e$ OR for each SNP is normally distributed, it should be 0.

In a similar way, the mean value of polygenic risk score for patients of a disease (β_{mean}) is

represented by the following equation:

$$\beta_{mean} = 2u_1\log_e OR_1 + 2u_2\log_e OR_2 + \ldots\ldots\ldots + 2u_n\log_e OR_n \qquad (16)$$

In this equation u_k and OR_k are MAF in the patient population and OR of the k-th SNP, respectively. In formula (16), for SNPs with positive $\log_e OR_k$, u_k is generally larger than p_k, and u_k is generally smaller than p_k for SNPs with negative $\log_e OR_k$. Consequently, β_{mean} becomes greater than 0. Schematic images of β_{mean} and β_{mean} are represented in Fig. 2. Comparing Equations (15) and (16), V(g) of a variant in a patient population is represented as

$$V(g) = 2u\,(1 - u)\,\{\log_e (OR)\}^2 \qquad (17)$$

Therefore, h_p^2 is calculated as

$$h_p^2 = V(g) / \{V(g) + \pi^2/3\} \qquad (18)$$

The method using Equations (17) and (18) can be termed the "modified allele distribution model". In Equation (17), u is a function of p because it can be calculated using p, **P** and OR. In Fig. 1, h_p^2 calculated with the modified allele distribution model (h_3^2) is shown for various ORs in comparison with the h_p^2 calculated with the mutation model (h_2^2) (Figs. 1e-1h**)**. h_3^2 is always higher than h_2^2, but the graphs are similar in shape. The curves for h_1^2 and h_3^2 look similar for low values of OR, but diverge as OR increases (Figs. 1a-1h).

What is the expected for β_{mean} in the general population? Under the assumption that $\log_e$ OR for each SNP is normally distributed, it should be 0.

Comparison of h_p^2 by three methods for variants from the literature

Table 1 shows a comparison of h_p^2 assessed by the three methods discussed here for a number of single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) cited from the literatures[11-24]. The values of h_3^2 for variants with common AFs and intermediate AFs ($p \geq 0.05$) were not so different from h_1^2, while h_2^2 was obviously lower. For so-called rare variants with low AFs ($p < 0.01$), h_3^2 was the highest, and the difference from h_1^2 was prominent, with a ratio ranging from 3.2 up to 53.7-fold. All variants whose risk alleles were not found in the control group were scored as rare variants, and h_p^2 was calculated as $p = \mathbf{P}u$ using the mutation model.

In Table 2, features of the three methods are summarized.

Discussion

The results show significant differences in the value of h_p^2 depending on the calculation method used. As shown in Fig. 1 and Table 2, the values of h_p^2 by three calculation methods (h_1^2, h_2^2 and h_3^2) show similar trends, although **P** is different. As mentioned in the derivation of the modified allele distribution model, the AF of the variant in the patient population is appropriately used to calculate the genetic variance of the variant. Accepting that formulation, the values of h_1^2 by the allele distribution model without the modification proposed here are

inaccurately low.

Why then does h_2^2 take a lower value than h_3^2? A possible cause is as follows: h_2^2 is calculated with **P** and **Q** using the liability threshold model supposing that only the variant is a genetic factor for a disease. **Q** is calculated with Equation (12), which assumes that the non-risk genotype has no genetic effect. On the other hand, h_1^2 is calculated using Equations (17) and (18), which are derived considering all genotypes of the variant[5]. Because h_3^2 is calculated using Equations (17) and (18), which are applications of Equations (13) and (14), all genotypes are also considered in the calculation of h_3^2. Therefore, because the mutation model does not consider the contribution of the non-risk genotype of the variant to h_p^2, h_2^2 is assessed as smaller than h_3^2.

As shown in Figs. 1e to 1h, h_2^2 becomes closer to h_3^2 as the OR increases. In other words, at high OR, the h_p^2 of a variant calculated with the modified allele distribution model becomes closer to h_p^2 calculated with the mutation model This reflects the transition of a rare variant to a rare allele causing Mendelian disease when the effect size becomes high[11].

h_3^2 calculated with the modified allele distribution model suggests that the contribution of rare variants with high ORs to h^2 of a common disease has been generally underestimated with the original allele distribution model. Well then, is h_2^2 which is calculated with the mutation model meaningless? As shown in Table 2, for a variant in which the risk allele is not found in

the control group, *V(g)* becomes infinite when the modified allele distribution model is used, and the non-risk allele of such a variant can be regarded as the wild type allele. I infer that reevaluation of the h^2 is required for all the variants already identified as associated with diseases – frequent as well as rare.

To carry out such reevaluation h^2 for a common disease using the modified allele distribution model, one must calculate *V(g)* for each variant and then estimate V_G as the sum of all *V(g)*s. Even when the mutation model is appropriate, the sum of (**Q** − **P**) must first be calculated for each variant, rather than using the sum of individual h_p^2 values (Nagao, 2015).

In conclusion, the analysis indicates that genetic variance of a variant affecting a disease should be calculated using the AF in the patient population, and as was seen for sample instances presented here, h_p^2 for rare variants with high ORs then is scored as significantly higher than h_p^2 calculated using AFs for the general population. Because rare variants show high ORs compared to common variants in average[25], the contribution of rare variants to the heritability of common diseases is likely much larger than previously thought. Therefore, for diseases and qualitative traits, one source of ”missing heritability” is likely to be the use in conventional methods of AF of a variant in the general population rather than in the patient population[5,6]. Reevaluation of h^2 for previously identified variants and discovery of novel rare variants would thereby significantly reduce the level of missing heritability for common

diseases and possibly for qualitative traits as well.

Methods

Derivation of the frequency of the risk genotype in the first-degree relatives of a patient

Let the frequency of the risk genotype(s) of a variant and its penetrance be X_1 and α in the general population, respectively. X_1 represents ($p^2 + 2pq$) for a dominant risk allele; p for a co-dominant risk allele; and p^2 for a recessive risk allele.

Suppose that this variant is the only genetic factor for a disease. Assuming that the non-risk genotype is not involved in onset, the cumulative population incidence of a disease, **I,** is represented by the following equation, including the incidence attributable to environmental factors as E:

$$\mathbf{I} = \alpha X_1 + E \qquad (19)$$

Derivation of population attributable risk

Population attributable risk (PAR) is the proportion of the incidence of a disease in the population that is due to exposure. Table 3 shows the four basic quantities, A, B, C, and D applicable to a case-control study. PAR is represented by the following formula:

$$PAR = (A+C)/(A + B + C+D) - C/(C+D) \quad (20)$$

In Equations (6) and (7), the ratios of **P**u, (1 − **P**)x, **P**v, and (1 − **P**)y correspond to the ratios of A, B, C, and D in Table 3, respectively. Therefore, PAR is represented as following:

$$PAR = \mathbf{P}(1 - v/q) \quad (21)$$

The following equation then holds:

$$\mathbf{P}\,(1 - v/q) = \alpha X_1 \quad (22)$$

Derivation of the frequency of the risk genotype in the first-degree relatives of a patient

Regarding the incidence in the first-degree relatives of a patient, **Q**, if the frequency of the risk genotype(s) is X_2 and the penetrance is α', **Q** is represented as following:

$$\mathbf{Q} = \alpha' X_2 + E \quad (23)$$

Because the relatives of the patient belong to the same population, α' is equal to α. Then **Q** is represented as

$$\mathbf{Q} = \alpha X_2 + E$$

$$= \mathbf{P}\,(1 - v/q)\,(X_2/X_1) + E \quad (24)$$

Using Equations [19], [21] and [24], **Q** is equivalent to

$$\mathbf{Q} = \mathbf{P}\,(1 - v/q)\,(X_2/X_1) + \mathbf{I} - \alpha X_1$$

$$= \mathbf{P}\,(1 - v/q)\,\{(X_2/X_1) - 1\} + \mathbf{I} \quad (25)$$

The cumulative population incidence **I** of each disease is approximated by the prevalence **P** for a chronic disease. Therefore, **Q** is represented by the following equation:

$$\mathbf{Q} = \mathbf{P} + \mathbf{P}\,(1 - v/q)\,\{(X_2/X_1) - 1\} \qquad (26)$$

When this equation is applied to a variant under the recessive model, the heterozygote is a non-risk genotype, and the population attributable risk should be qualified[8].

The frequency of the risk genotype for the variant under the co-dominant model

To assess the contribution of a variant to the incidence of a disease in a population, suppose a variant, Var1, is associated with the disease. Definitions and premises are then as follows:

- p_1 is the minor AF of risk alleles of Var1.
- q_1 is the major AF of non-risk alleles of Var1.
- M and N are the risk and non-risk alleles, respectively.
- a, b, and c are the respective penetrance of genotypes MM, MN, and NN.

I, is represented by

$$\mathbf{I} = ap_1^2 + 2bp_1q_1 + cq_1^2 \qquad (27)$$

Assuming that the genetic mode of Var1 is codominant and that heterozygotes show an additive effect, then 2b = a + c. Therefore, I can be represented as

$$\mathbf{I} = ap_1 + cq_1 \qquad (28)$$

As shown in Equation (26), for the variant under a co-dominant model, the sum of the frequency of homozygosity of the risk allele and half the frequency of heterozygosity is regarded as the "frequency of risk genotype in calculation"; it is p_1.

Calculation of genotype frequency of risk allele for the first-degree relatives

To calculate genotype frequencies of the risk allele for first-degree relatives,

- M and N represent risk and non-risk allele, respectively.
- β is the genotype frequency of MM for the proband.
- γ is the genotype frequency of MN for the proband.
- δ is the genotype frequency of NN for the proband.

Probabilities of each genotype for offspring is as follow[8]:

MM: $\beta p + \gamma p/2$

MN: $\beta q + \gamma/2 + \delta p$

NN: $\gamma q/2 + \delta q$

For the proband, β, γ, δ become u^2, $2uv$, and v^2, respectively. Therefore, the genotype frequency of MM for offspring is up and that of MN is $(u^2q + uv + v^2p)$. Then, for co-dominant inheritance, X_2 is represented as follow:

$$X_2 = up + (u^2q + uv + v^2p)/2 = (p + u)/2 \qquad (29)$$

As described above, the risk genotype(s) of a variant and its penetrance, X_1 represents p for a codominant risk allele. Therefore, Therefore, **Q** is represented by the following equation:

$$\mathbf{Q} = \mathbf{P} + \mathbf{P}\,(1 - v/q)\,\{(p + u) / (2p) - 1\} \qquad (12)$$

Data availability

All data generated or analysed during this study are included in this published article.

References

1. Wu, M. C, *et al*. Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* **86**, 929-942 (2010).
2. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
3. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111,** E455–464 (2014).
4. Wray, N. & Visscher, P. Estimating trait heritability. *Nature Education* **1**, 29 (2008)
5. Pawitan, Y., Seng, K. C. & Magnusson, P. K. How many genetic variants remain to be discovered? *PLoS One* **4,** e7969 (2009).

6. Lee, S. H, *et al*. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294-305 (2011).

7. Oliynyk, R. T. Age-related late-onset disease heritability patterns and implications for genome-wide association studies. *PeerJ* **7**, e7168 (2019).

8. Nagao, Y. Copy number variations play important roles in heredity of common diseases: a novel method to calculate heritability of a polymorphism. *Sci. Rep*. **5**, 17156 (2015).

9. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet. Lond*. **29,** 51–76 (1965).

10. Schultz, S. H., North, S. W. & Shields, C. G. Schizophrenia: a review. *Am. Fam. Physician.* **75,** 1821–1829 (2007).

11. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

12. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism *N. Engl. J. Med*. **358,** 667–675 (2008).

13. Weiss, L. A. & Arking, D. E. The Gene Discovery Project of Johns Hopkins the Autism Consortium. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461,** 802–808 (2009).

14. Ronai, Z. *et al.* Glycogen synthase kinase 3 beta gene structural variants as possible risk

factors of bipolar depression. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **165B,** 217–222 (2014).

15. McMahon, F. J. *et al.* Bipolar Disorder Genome Study (BiGS) Consortium. Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. *Nat. Genet.* **42,** 128–131 (2010).
16. Chen, J., *et al.* A pilot study on commonality and specificity of copy number variants in schizophrenia and bipolar disorder *Transl. Psychiatry* **6**, 824 (2016).
17. Rees, E., *et al*. Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204**, 108-114 (2014).
18. Stefansson, H., et al. Large recurrent microdeletions associated with schizophrenia. *Nature*. **455**, 232-236 (2008).
19. Dow, D. J. *et al.* ADAMTSL3 as a candidate gene for schizophrenia: gene sequencing and ultra-high density association analysis by imputation. *Schizophr. Res.* **127,** 28–34 (2011).
20. Walitza, S. *et al.* Pilot study on HTR2A promoter polymorphism, -1438G/A (rs6311) and a nearby copy number variation showed association with onset and severity in early onset obsessive-compulsive disorder. *J. Neural. Transm.* **119,** 507–515 (2012).
21. Kato, T. *et al.* Segmental copy-number gain within the region of isopentenyl diphosphate isomerase genes in sporadic amyotrophic lateral sclerosis. *Biochem. Biophys. Res. Commun.*

402, 438–442 (2010).

22. van Es, M. A. *et al.* Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **41,** 1083–1087 (2009).

23. Kudo, H. *et al.* Frequent loss of genome gap region in 4p16.3 subtelomere in early-onset type 2 diabetes mellitus. *Exp. Diabetes Res.* **2011,** 498460 (2011).

24. SIGMA Type 2 Diabetes Consortium *et al.* Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311,** 2305–2314 (2014).

25. Walter, B. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet*. **40**, 695-701. (2008).

Acknowledgements

I am grateful to Dr. David Schlessinger of National Institute of Aging for valuable advice.

Contributions

Y.N. designed the study. Y.N. is responsible for the assessment and discussion of the obtained results and wrote the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

Figure Legends

FIGURE 1. h_p^2 of variants calculated by three methods for various odds ratios.

(a) to (d) h_p^2 of variants are shown for odds ratio (OR) of 1.05, 4, 10 and 100 calculated with the allele distribution model (blue line) and with the mutation model (green line). Vertical axis, value of h_p^2; horizontal axis, allele frequency in the general population.

(e) to (h) h_p^2 of variants are shown for OR of 1.05, 4, 10 and 100 calculated with the modified allele distribution model (red line) and with the mutation model (green line). Vertical axis, value of h_p^2; horizontal axis, allele frequency in the general population.

FIGURE 2. Distribution of the proportion of variants that are associated with a disease

The proportions of variants are normally distributed in the general population (blue line), whereas the distribution of variant proportions in the patient population becomes asymmetric (red line). β_{mean} measures the deviation of the mean value of $\log_e$ OR from y-axis in the general population and β_{mean} measures the deviation of the mean value of $\log_e$ OR from y-axis in the

patient population.

Horizontal axis: log scaled OR of variant

Vertical axis: proportion of variant

Table 1. Results of a trial to calculate h^2 by the allele distribution model (h_1^2), the mutation model (h_2^2) and the modified allele distribution model (h_3^2) of CNVs and SNPs using published data.

Disease	**Variants**	**Population or source**	**OR**	**P**	**p**	**h_1^2**	**h_2^2**	**h_3^2**
Autism	CNV* (16p11.2 del)[11,12]	Americans	100	0.006	0.00016	0.0021	0.0666	0.1128
	CNV (16p11.2 dup) [11,12]	Americans	16	0.006	0.00034	0.0016	0.0077	0.0226
	SNP (rs4307059)[13]	Americans	1.19	0.0067	0.61	0.00439	0.00087	0.00419
Depression	CNV (3q13.33 dup)[14]	Hungarians	5.27	0.085	0.013	0.0211	0.0617	0.0725
	SNP (rs2251219)[15]	(Meta-analysis)	0.87	0.2	0.4	0.00282	0.00151	0.00275
Schizophrenia	CNV (22q11.2 del)[16]	European	∞	0.01	0.000066	†	0.0801	†
	CNV (16p11.2 dup)[17]	Several sources	∞	0.01	0.000039	†	0.0498	†
	CNV (NRXN1 del)[17]	Several sources	∞	0.01	0.000016	†	0.0214	†
	CNV (AS/PWS dup)[17]	Several sources	∞	0.01	0.000012	†	0.0161	†
	CNV (15q13.3 del)[18]	Several sources	8.27	0.01	0.00021	0.00057	0.00132	0.00436
	CNV (1q21.1 del)[18]	Several sources	12.0	0.01	0.00022	0.00061	0.00199	0.00607
	SNP (ADAMTSL3)[19]	(HapMap)	0.68	0.01	0.29	0.01828	0.00353	0.01517
Obsessive-compulsive disorder	CNV (13q14.2 del)[20]	Swiss	6.23	0.023	0.01	0.01974	0.03265	0.09330
	SNP (rs6311, located on HTR2A promoter)[20]	Swiss	1.69	0.023	0.44	0.03901	0.01160	0.03948
Sporadic ALS	CNV (10q15.3 dup)[21]	Japanese	10.8	0.0001	0.101	0.2382	0.0970	0.4605
	SNP (rs10260404)[22]	Dutch	1.3	0.0001	0.27	0.00818	0.00097	0.00918
Type 2	CNV (4p16.3 del)[23]	Japanese	14.8	0.1	0.022	0.0867	0.1613	0.3331

diabetes mellitus	SNP (missense variant of HNF1A gene)[24]	Mexicans and US Latinos	5.48	0.14	0.006	0.0104	0.0130	0.0334

Odds ratio (OR), allele frequency (p), and prevalence of disease (**P**) of each variant are cited from the literatures[11-24].

Table 2. Features of three different methods to estimate h_p^2

	Allele distribution model	Mutation model	Modified allele distribution model
Published year	2009	2015	This year
Symbol of h_p^2 in this paper	h_1^2	h_2^2	h_3^2
Parameters in calculation	p and OR	p, u, and **P**	u and OR, or p, **P**, and OR
Target of calculation	variance	deviation	variance
For common variant	slightly underestimating	underestimating	thought to be appropriate
For rare variant	highly underestimating	underestimating	thought to be appropriate
For variant with OR of +∞	not applicable	applicable	not applicable

p: MAF of the risk allele in the general population

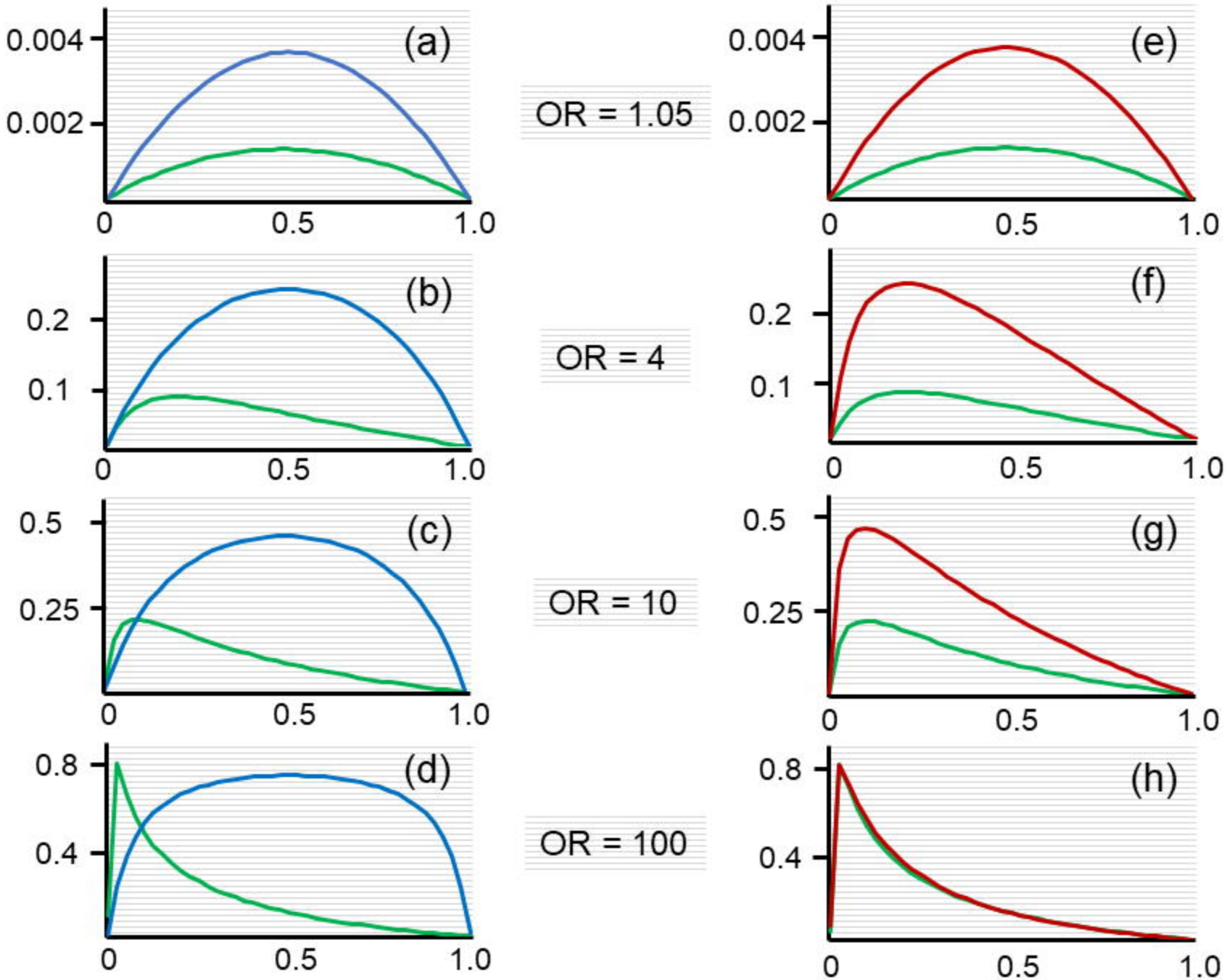
u; MAF of the risk allele in the patient population

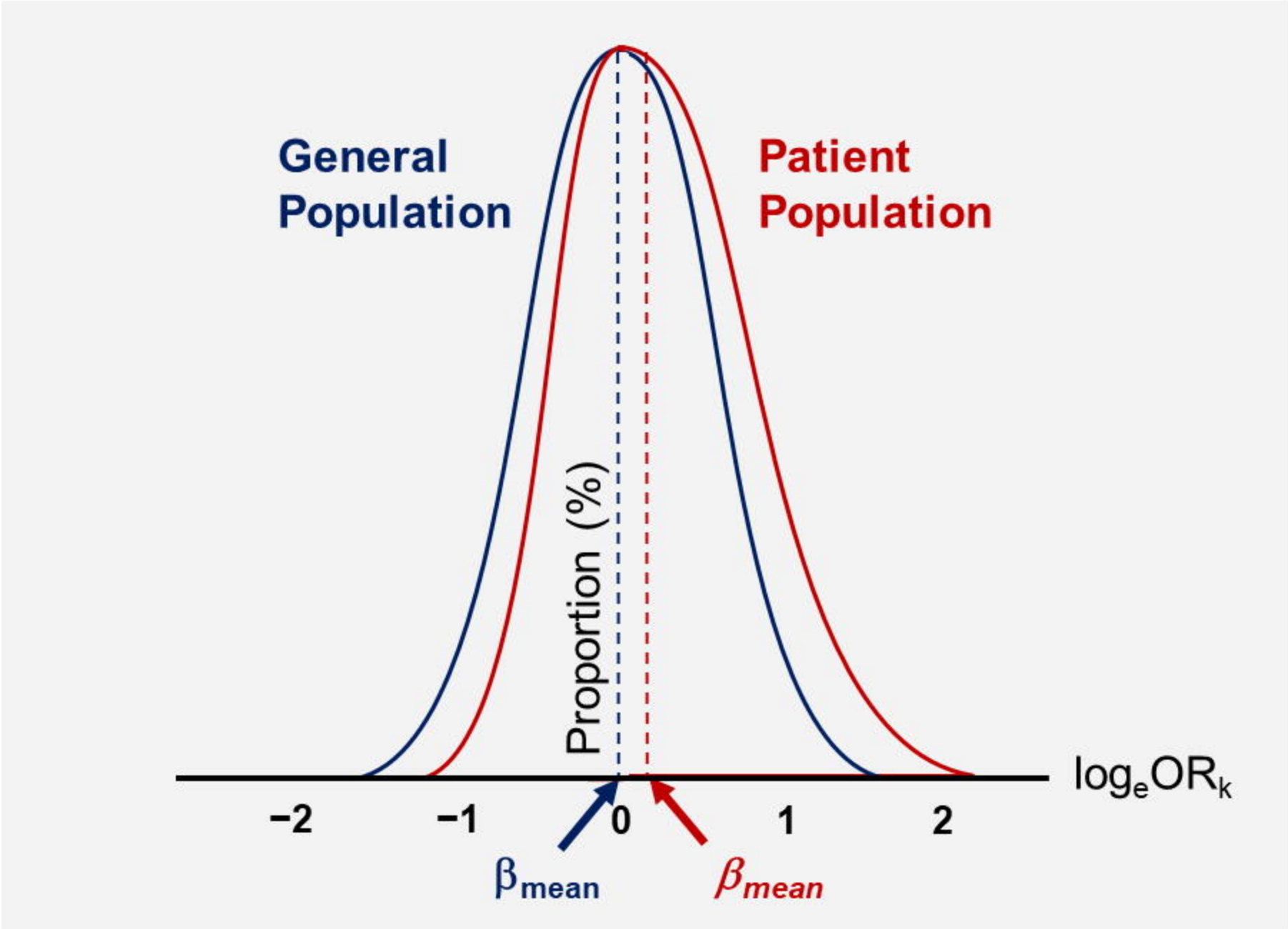
P: prevalence of a disease

OR: odds ratio

Table 3. Four basic quantities in the case-control study

	Patient group	Unaffected group	Total
Exposed	A	B	A + B
Unexposed	C	D	C + D
Total	A + C	B + D	A + B + C + D


(a)
0.004
0.002
0
0.5
1.0
OR = 1.05
(e)
0.004
0.002
0
0.5
1.0
(b)
0.2
0.1
0
0.5
1.0
OR = 4
(f)
0.2
0.1
0
0.5
1.0
(c)
0.5
0.25
0
0.5
1.0
OR = 10
(g)
0.5
0.25
0
0.5
1.0
(d)
0.8
0.4
0
0.5
1.0
OR = 100
(h)
0.8
0.4
0
0.5
1.0


General Population
Patient Population
Proportion (%)
$\log_e OR_k$
−2
−1
0
1
2
β_{mean}
β_{mean}