Progress in Neuro-Psychopharmacology & Biological Psychiatry 38 (2012) 302–309

Contents lists available at SciVerse ScienceDirect

Progress in Neuro-Psychopharmacology & Biological Psychiatry

journal homepage: www.elsevier.com/locate/pnp

How many genes are involved in schizophrenia? A simple simulation

Myung Jae Paek [a], Ung Gu Kang [b,c,d,*]

[a] Department of Psychiatry, The Armed Forces Capital Hospital, Seongnam, Republic of Korea
[b] Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea
[c] Department of Psychiatry and Behavioral Science, Seoul National University College of Medicine, Seoul, Republic of Korea
[d] Institute of Human Behavioral Medicine, Medical Research Center, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form 16 April 2012
Accepted 17 April 2012
Available online 24 April 2012

Keywords:
Genetic epidemiology
Odds ratio
Polygenic threshold model
Schizophrenia
Simulation

ABSTRACT

We attempted to estimate how many genes are involved in schizophrenia using a simulation based on the polygenic threshold model. The basic assumptions were as follows: (1) All genes involved are transmitted independently; (2) every locus is composed of two alleles — one pathogenic and the other non-pathogenic; (3) all pathogenic alleles are dominant; (4) the two alleles at any locus are in Hardy–Weinberg Equilibrium (HWE) in the general population (GP) but not within the patient (PP) or non-patient (NP) subpopulations; (5) the number of affected loci determines the disease genetically; and (6) only a fraction of genetically determined individuals actually becomes ill. A range of the total number of disease-related genes (N) and threshold genetic load (T) was set for the simulation. Assuming that the number of affected loci follows a binomial distribution, the mean gene frequencies satisfying a disease prevalence of 1.12% in the GP were sought for various N and T combinations. Based on these gene frequencies, the odds ratio and the incidence rate in relatives under random mating were calculated. These results were then compared with real genetic epidemiologic data to obtain best-fit estimates for N and T. The results indicated that a polygenic threshold model with an N greater than 100 and a T in the range of 0.3–0.8 fits the empirical data. It was estimated that at least several hundreds of study subjects are required to yield a statistically significant frequency difference for a single gene between the patient and the control groups.

1. Introduction

Although schizophrenia has a strong genetic component, its mode of inheritance remains unclear. Some recent approaches stress the importance of rare copy number variations (CNVs) (Grozeva et al., 2012; International Schizophrenia Consortium, 2008). Although these cytogenetic variations have a substantial effect size, they are found only in a minority of the patients (Grozeva et al., 2012) and more traditional approaches underscore the concerted smaller effects of a large number of genes. For example, the polygenic threshold model (Falconer, 1965) could successfully explain the risk of schizophrenia morbidity in relatives (McGue et al., 1985). However, many of the genes implicated as candidates in well-designed studies have not been confirmed as candidates in other well-designed studies or meta-analyses, and even when confirmed, they typically have low odds ratios (ORs). Studies cited in a recent review show this point (Girard et al., 2011). If the risk attributable to any single gene is low, the sample size required to obtain statistically stable results becomes large (Wang et al., 2002) and may be practically unattainable.

To address the questions “How many genes are involved in schizophrenia?” and “How large a sample is required for stable detection of a true genetic difference?” a primitive model for the genetics of schizophrenia was constructed based on the binomial distribution. This model was tested by simulations using various parameters: the penetrance, the number of schizophrenia-related genes (N), and the threshold proportion of genes that must be pathogenic for an individual to manifest the illness (T). The simulation results were then compared to real world data to obtain the best-fit values for these parameters. To make the model as simple as possible, following basic assumptions were made:

(1) The genes related to schizophrenia are dispersed across the genome and are inherited independently. The maximum N was set at 160, a number approaching 1% of the total number of human genes (International Human Genome Sequencing Consortium, 2004). As a 1% chance of genetic recombination in humans corresponds to an average of 7.5×10^5 bps (Scott et al., 2004) and the total length of human genome is 2.85×10^9 bps (International Human Genome Sequencing Consortium, 2004), one hundred sixty genes dispersed in this wide range have little chance of co-segregation.
(2) Each gene (locus) has only two, one pathogenic and one non-pathogenic alleles. This model combines all of the pathogenic

Abbreviations: CNV, copy number variation; HWE, Hardy–Weinberg Equilibrium; GP, general population; MAF, minor allele frequency; NP, non-patient population; OR, odds ratio; PP, patient population; SN, sample number; $p_{G\_}$, genotype frequency (in designated population); $p_{A\_}$, allele frequency (in designated population).

* Corresponding author at: Department of Psychiatry and Behavioral Science, Seoul National University College of Medicine, Yongon-dong, Chongno-gu, Seoul, 110-744, Republic of Korea. Tel.: +82 2 2072 2296; fax: +82 2 744 7241.
E-mail address: kangug@plaza.snu.ac.kr (U.G. Kang).

0278-5846/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.pnpbp.2012.04.015

alleles and treats them as a single pathogenic allele. Similarly, all non-pathogenic alleles are treated as a single non-pathogenic allele.

(3) All pathogenic alleles are dominant. Although pathogenic alleles may have dominant, recessive or dose-dependent effects, the model makes the assumption that the schizophrenic process is manifested by gain-of-function mutations that result in the novel emergence of pathologic behaviors. This type of mutation tends to have dominant effects (Wilkie, 1994).

(4) The frequencies of the two alleles at any locus are in Hardy–Weinberg Equilibrium (HWE) in the general population (GP). Although the frequency of pathogenic allele is higher in the patient population (PP) than in the GP, the homozygote/heterozygote ratios in the PP and the GP are the same. This assumption holds because the patients are selected from the GP, and homozygous and heterozygous individuals have the same chance of being selected if the gene effect is dominant. Because the patients do not create offspring within a segregated group, HWE is not likely to be established in the PP. The same logic applies to the non-patient population (NP).

(5) The manifestation of the disease depends solely on the number of affected loci regardless of the identity. This assumption makes it easy to apply the polygenic threshold model of liability proposed by Falconer (1965) and this model is related to the spectrum concept of the disease. Although current diagnostic criteria such as DSM-IV-TR define schizophrenia as a distinct disease that can be differentiated from the normal state, the spectrum concept was proposed early on (Siever and Davis, 2004; Weller, 1987). The concept of clinical treatment for the spectrum disorder is widely accepted and is the basis for suggestions that non-schizophrenic subjects (Meltzer, 1999) or the relatives of schizophrenic patients (Tsuang et al., 1999) might benefit from treatment. Recent studies on genetically high-risk populations also invoke the spectrum concept. In familial disease, the spectrum concept includes both the genetic load and the threshold.

(6) Finally, among the individuals who have supra-threshold genetic loads, only a fraction will become ill (incomplete penetration). Although Procopio (2005) suggested that a random effect influences the summed liability that is determined by the number of loci that have pathogenic alleles, the present model assumes that the less-than-perfect penetration is attributable to non-genetic effects and the penetrance was set as a constant. The concept of penetration is considered on the level of the disease phenotype, not on the level of gene effect.

Using these basic assumptions, an algorithm was constructed that shows the effects of N and T using a binomial distribution with N-independent trials.

2. Methods

2.1. Basic parameters

The number of schizophrenia-related genes (N) was set to range from 20 to 160. The number of proposed schizophrenia-related genes is now well over ten; this number was scaled up by doubling it until it approached 1% of the total number of human genes (20,000–25,000) (International Human Genome Sequencing Consortium, 2004). The threshold T was arbitrarily allowed to range from 0.05 to 0.9. The prevalence of schizophrenia in the GP was defined as 1.12% (Gottesman et al., 2010). A default penetrance of 0.5 was chosen based on the assumption that the proband-wise monozygotic twin concordance rate is a valid estimate for the penetrance as defined in the present study. For simplicity, a single averaged frequency was assigned for all pathogenic alleles.

2.2. Allele frequency and genotype frequency

The assumption that all pathogenic alleles are dominant and that all loci are independent means that the distribution of genetic load in the GP should follow a binomial distribution determined by N and the average genotype frequency (summed homozygous and heterozygous, designated as p_{G_GP}). In the B(N, p_{G_GP}) distribution of genetic load, a load equal to or greater than the product N•T passes the threshold. Individuals who pass the threshold become ill with a chance equal to the penetrance. From the above parameters, the upper 2.24% of the distribution becomes genetic candidates. Therefore p_{G_GP} values that satisfied that condition were sought for each N and T combination (Fig. 1).

The pathogenic allele frequency in the GP (p_{A_GP}) under HWE was calculated from p_{G_GP}. For details, see the Appendix A. The mean genetic load and the probability of each locus having the pathogenic genotype in the PP (p_{G_PP}) were calculated as shown in the Appendix A. Under assumption (4) above, the pathogenic allele frequency in the PP (p_{A_PP}) was then calculated. The allele frequencies in the NP (p_{A_NP}) were calculated from the equation below:

$$p_{A\_GP} = p_{A\_PP} \cdot \text{prevalence} + [p_{A\_NP} \cdot (1 - \text{prevalence})]$$

The genotype frequencies in the NP (p_{G_GP}) were calculated similarly. All calculations were performed using Microsoft Excel software. All subsequent calculations depend on these values.

2.3. Odds ratio

The ORs were defined as follows:

$$\text{OR} = [p_{A\_PP}/(1-p_{A\_PP})]/[p_{A\_NP}/(1-p_{A\_NP})] \quad (allele-wise)$$
$$\text{OR} = [p_{G\_PP}/(1-p_{G\_PP})]/[p_{G\_NP}/(1-p_{G\_NP})] \quad (genotype-wise)$$

2.4. Prevalence in offspring and other relatives

To obtain the prevalence in a patient's parents, the probability that the patient had a specific genotype at a particular locus was calculated. Next, the probability that the alleles were inherited from either the mother or the father was calculated and finally, the probability that the mother or the father had the specific genotype was calculated, assuming that the parents were from the GP.

For PP × PP, PP × NP and NP × NP offspring and parent × parent offspring (*i.e.*, sibling), the genotype frequency was calculated from the allele frequencies in the parental germ cells. Then, the proportion of offspring having a genetic load of at least N•T was calculated using

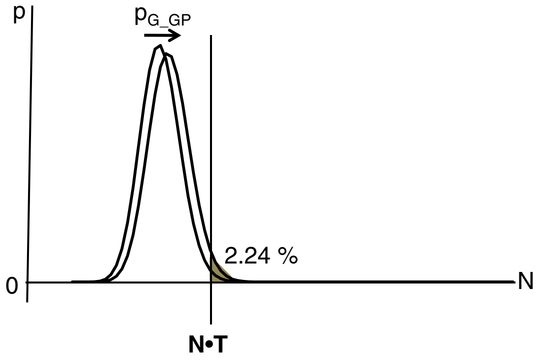


Fig. 1. Assigning p_{G_GP} values according to N and T in the binomial distribution B(N, P_{G_GP}). For each N and T combination, the appropriate genotype frequency in the general population (P_{G_GP}) was sought to satisfy the condition that the shaded area is 0.024.

the binomial distribution of the corresponding N. This value was adjusted by the penetrance to obtain the prevalence.

For second-degree relatives, the above algorithms were applied recursively. These relatives were defined as follows: half-sibling, offspring of parent × GP; uncle, parent's sibling; nephew, offspring of sibling × GP; grandparent, parent's parent; grandchild, GP × (offspring of PP × GP).

2.5. Number of samples required for statistically significant differences

Next, the number of samples (SN) required was estimated to obtain statistically significant differences at $P<0.05$. For a case–control study with a sample size of SN for each group, the distribution of allele frequency in the sample can be approximated to a normal distribution with the following parameters:

$$\begin{aligned}&\mathrm{PP}: \mathrm{mean} = p_{A\_PP};\ \sigma^2 = p_{A\_PP}\cdot\left(1-p_{A_pP}\right)/\mathrm{SN}\\&\mathrm{NP}: \mathrm{mean} = p_{A_NP};\ \sigma^2 = p_{A\_NP}\cdot\left(1-p_{A_NP}\right)/\mathrm{SN}\end{aligned}$$

The SN values were sought that would separate the mean $+\,1.96\sigma$ of the NP from the mean $-\,1.96\sigma$ of the PP. SN values were plotted against N and T, which determine p_{A_PP} and p_{A_NP}. A similar calculation was used to determine the SN values for genotype frequency.

3. Results

3.1. Gene frequencies according to N and T in the PP and NP

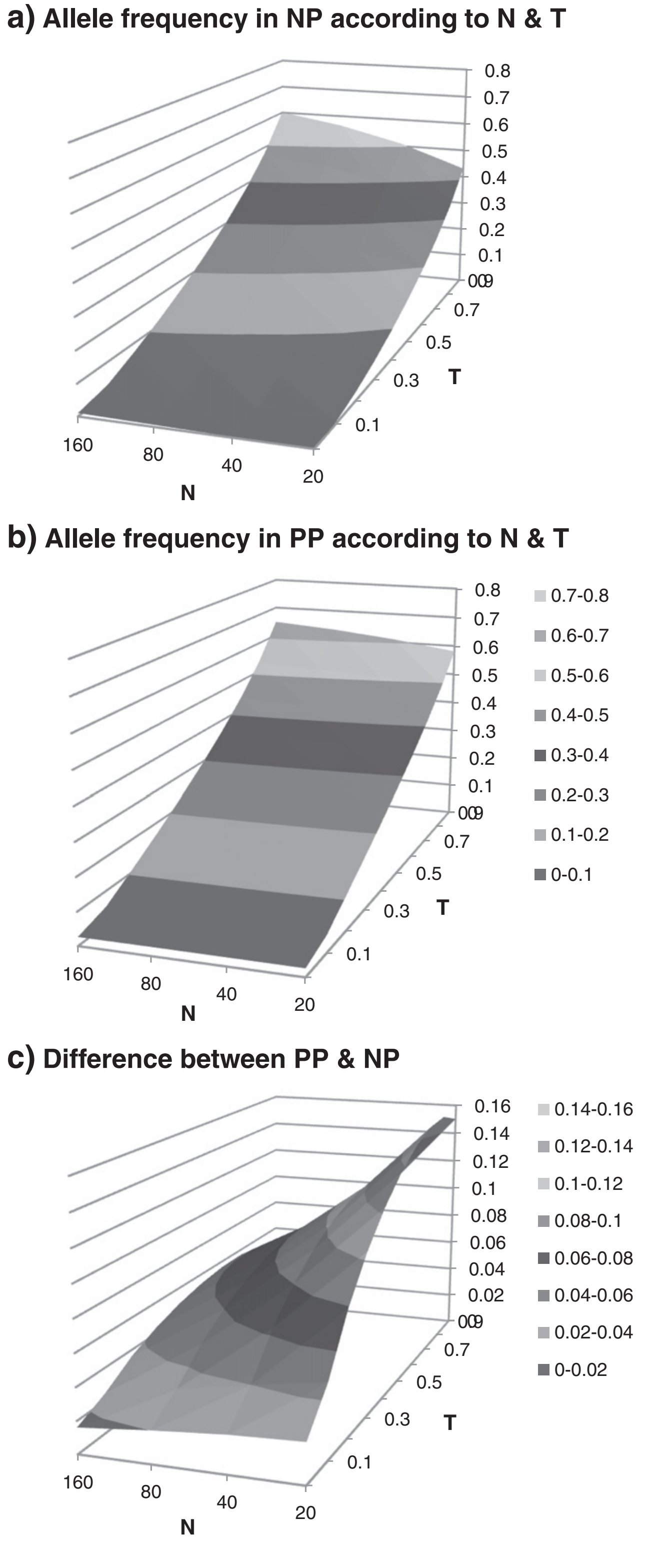
The mean allele frequencies in the NP (p_{A_NP}) and PP (p_{A_PP}) increased with increasing N and T but were more critically dependent on T than on N. The difference in allele frequency between the PP and NP decreased with increasing N. At any given N, the difference reached a maximum at T = 0.7 or 0.8. The maximum difference (15.5%) was obtained for N = 20 and T = 0.8 (Fig. 2). The allele frequency in the GP (p_{A_GP}) was very close to p_{A_NP} (not shown). The genotype frequency showed similar patterns. The difference between the PP (P_{G_PP}) and NP (P_{G_NP}) reached a maximum at N = 20 and T = 0.7 (Fig. 3). Regardless of the specific values of N and T, the mean genetic load divided by N (mean disease genotype frequency) in the PP was only slightly greater than T, the threshold, owing to the characteristics of binomial distribution (Fig. 3d).

3.2. Odds ratio

The allele-wise OR of the PP against the NP increased with decreasing N and T (Fig. 4a). Since N•T signifies the minimum number of genes required for the expression of disease, the OR was plotted against N•T. It increased with decreasing N•T. The typical OR obtained from genetic studies is less than 1.5, and at this level N•T was greater than 40. When plotted against the GP, the result was virtually the same (Fig. 4b). Compared to the allele-wise OR, the genotype-wise OR was higher. The genotype-wise OR increased with decreasing N and increased at both extremes of T (Fig. 4c). For OR < 2, N ≥ 80 and 0.3 ≤ T ≤ 0.8. When the OR was calculated for varying values of penetrance, it increased with increasing penetrance. But the effect was not robust when N was large (Fig. 4a and c).

3.3. Prevalence in offspring

The prevalence in first-degree relatives was sensitive to the penetrance because of the way it was defined. Therefore, varying penetrance values were applied ranging from 0.25 to 1, using 0.5 as the default. The prevalence in the PP × PP offspring increased with decreasing N and T. The effect of N was not as robust as the effect of T, and the effect of N was minimal when T was low. At the default penetrance, the PP × PP offspring prevalence was 10.9–32.0%. When the penetrance was 0.25, the prevalence decreased to 6.7–16.1%. When the perfect penetration was


a) Allele frequency in NP according to N & T

b) Allele frequency in PP according to N & T

c) Difference between PP & NP

Fig. 2. Pathogenic allele frequency according to N and T (penetrance = 0.5). (a) In the non-patient population (NP), the allele frequency increased with increasing N and T. Nearly identical results were obtained for the general population (GP). (b) In the patient population (PP), the allele frequency increased with increasing N and T. (c) The difference between the PP and NP increased with decreasing N and reached maximum when T was 0.7 to 0.8. N, number of genes that is hypothetically involved in schizophrenia (20–160); T, hypothetical threshold (the fraction of N genes that is required for manifestation of the illness; 0.05–0.9).

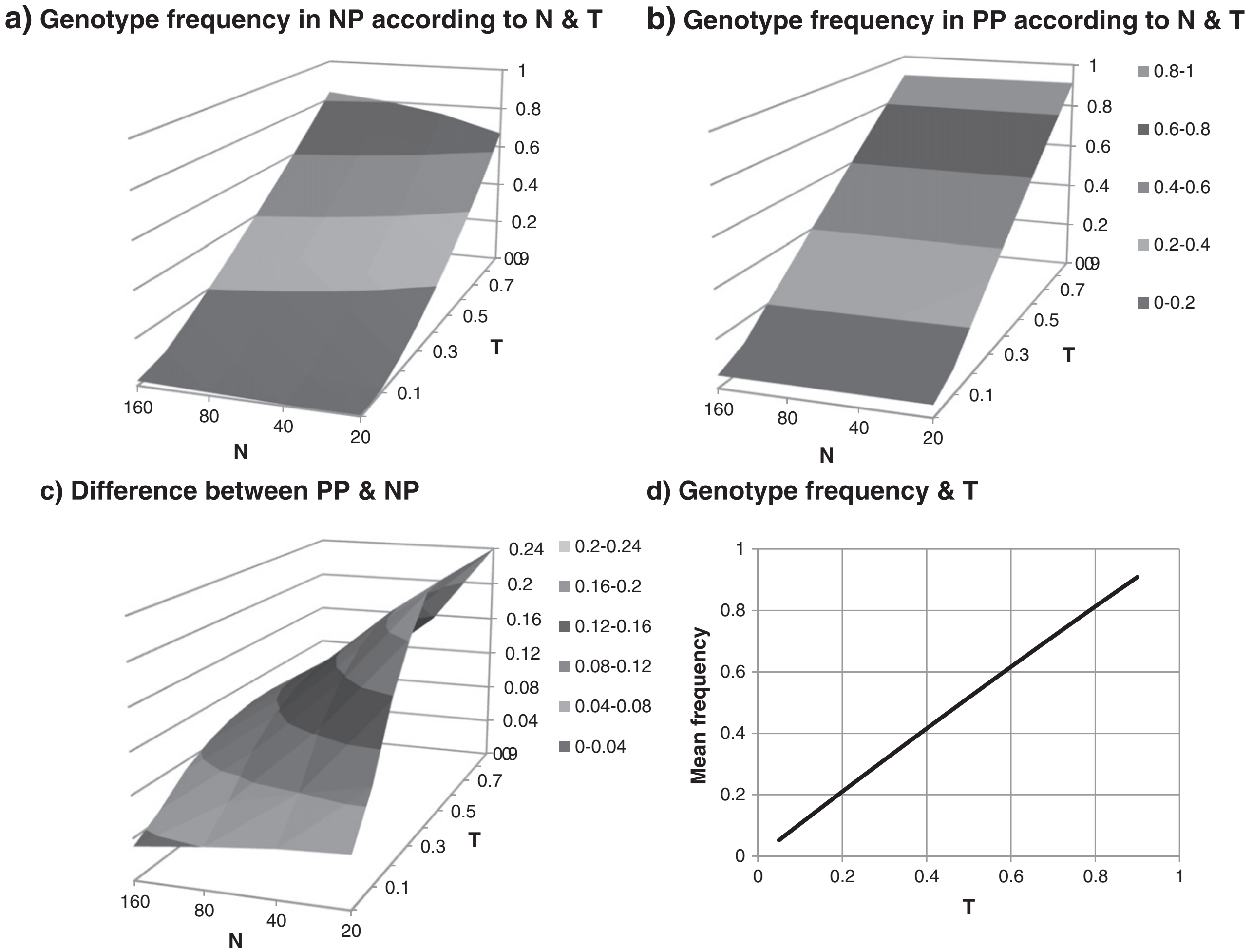


Fig. 3. Pathogenic genotype frequency according to N and T (penetrance = 0.5). (a) For the NP, the frequency of pathogenic genotype increased with increasing N and T. Nearly identical results were obtained for the GP. (b) For the PP, gene frequencies increased linearly with T. The value of N had little effect on the genotype frequency. (c) The difference between the PP and NP increased with decreasing N and reached maximum when T was 0.6 to 0.7. (d) Relationship between genotype frequency in the PP (averaged across all N values) and the threshold value. The averaged frequency was slightly greater than T. N, T, GP, NP and PP are defined as in Fig. 2.

assumed, it was 17.9–63.9% (Fig. 5a). The prevalence in the PP × NP offspring also increased with decreasing N and T. The effect of N was more robust when T was small. At the penetrance values of 0.25, 0.5, and 1, the prevalence ranges were 3.1–10.4, 4.2–20.2, and 5.6–39.9%, respectively (Fig. 5b). The prevalence in the NP × NP offspring decreased with decreasing N and T. When T was greater than 0.2, the prevalence was 1.0–1.1%, regardless of N. Variation in the penetrance value did not significantly affect the result. At the lowest T value, the prevalence decreased sharply with decreasing N (Fig. 5c). At the most extreme values of N (20), T (0.05), and penetrance (1), the disease is entirely determined by a single gene and NP group has a genetic load of 0, so the prevalence in the offspring should be 0.

3.4. Prevalence in other relatives

In parents and siblings, prevalence increased with decreasing N and T, but the effect of T was stronger. Prevalence was slightly higher in parents than in siblings. At T = 0.9, the parental prevalence range was 4.2–5.6%, whereas the sibling prevalence range was 4.1–5.1%. At T = 0.05, the parental and sibling prevalence ranges were 12.3–20.4 and 12.3–20.2%, respectively (Fig. 5a and b). In second-degree relatives, the hypothetical prevalence ranges were 2.2–2.6% at T = 0.9 and 5.0–11.9% at T = 0.05 (Fig. 5c). Prevalence values were the same for half-sibling, grandparents, grandchildren, and nephews, and were slightly lower in uncles.

3.5. Sample number required for a statistically significant result

In terms of allele-wise assessment, the SN was far more dependent on N than on T. For N values of 20 and 40, the SN ranges were 66–167 and 229–363, respectively, per group. For N values of 80 or more, the SN was greater than 482 (Fig. 7a). When assessed genotype-wise, the SN ranges were 33–53 and 335–469 for N values of 20 and 160, respectively (Fig. 7b). When the GP rather than the NP was used as a control, SN was somewhat larger, particularly at large N values, but the difference was not robust (Fig. 7a and b).

3.6. Effects of penetrance

It was found that the penetrance only slightly affected the allele or genotype frequencies in the PP, NP and GP (not shown). The OR increased slightly with increasing penetrance (Fig. 4). However, the variation in penetrance markedly affected the prevalence in the PP × PP and PP × NP offspring (Fig. 5a and b). The effect on NP × NP offspring was minimal (Fig. 5c).

4. Discussion

In this study, the polygenic threshold model of liability (Falconer, 1965) was adopted to simulate the genetics of schizophrenia. Procopio (2005) previously used the same model with a probabilistic

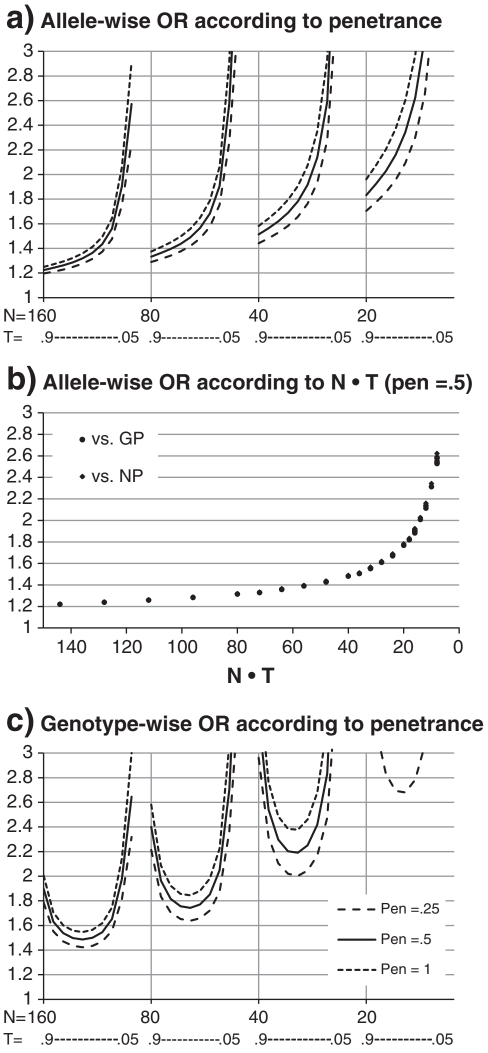


Fig. 4. Dependence of the odds ratio (OR) on N, T and the penetrance. (a) The allele-wise OR increased with decreasing N and T. For N≥80, OR was not markedly affected by the penetrance. (b) The allele-wise OR increased as the product N•T decreased (penetrance = 0.5). The ORs were nearly identical regardless of control group (GP or NP). (c) The genotype-wise OR increased with decreasing N and increased at both extremes of T. pen, penetrance; N, T, GP and NP are defined as in Fig. 2.

modification but did not provide numerical estimates. In the present study, a more deterministic model was used to provide some numerical estimates.

4.1. Assumption of non-HWE in the PP

An important assumption in the present study was that the genotypes in the PP were not in HWE. As the patients are not from a segregated population but selected from the GP because they exhibit deviant characteristics determined by the genetic locus of concern, this assumption is reasonable. When HWE is assumed within the PP, many of the results varied dramatically, generally deviating from the empirical data (not shown).

4.2. Odds ratios

The OR becomes small when N is large. If more genes are involved, then the contribution of any single gene should be diminished. For any given N and T, genotype-based ORs are higher than those based on allele frequency. This is due to the assumption of dominance.

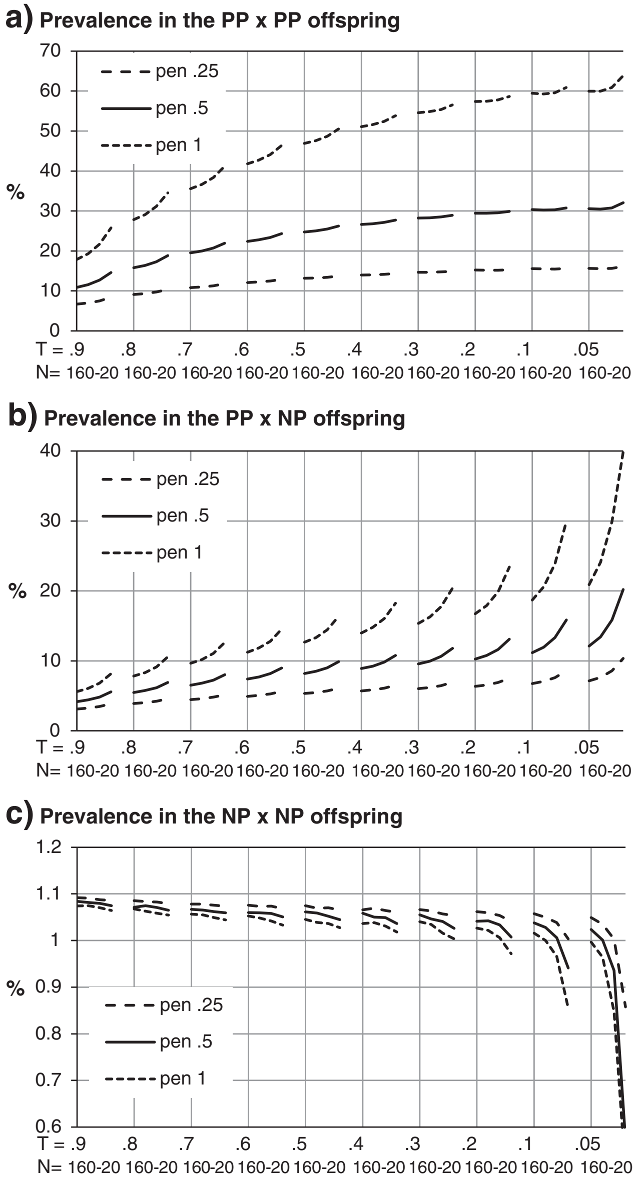


Fig. 5. Prevalence in offspring. (a, b) In both PP × PP offspring (a) and PP × NP offspring (b), prevalence increased with decreasing N and T but it was markedly affected by the penetrance. (c) In NP × NP offspring, prevalence decreased with decreasing N and T and was not affected by the penetrance. N, T, NP and PP are defined as in Fig. 2.

Some recent studies, including meta-analyses, have provided ORs for presumed schizophrenia-related genes. For example, the OR of the NRG1 haplotype is 1.22 (Li et al., 2006), and three significant markers on the CMYA5 gene have ORs ranging from 1.17 to 1.31 (Chen et al., 2011). A meta-analysis on ppp3c rs10108011 indicated an OR of 1.12 (Kyogoku et al., 2011), and another meta-analysis on the interleukin-1 gene complex locus indicated a pooled allele-wise OR of 0.88 (Xu and He, 2010). The maximum OR of 1.36 (inverse of 0.88) in these reports corresponds to an N•T greater than 64. Thus, if schizophrenia is caused entirely by genes with this range of OR, more than 64 genes would have to be simultaneously affected for the disease to be expressed, and the number of genes possibly related to schizophrenia (N) might be greater than 100. If the OR is much higher as in the cases with rare CNVs and if it is 5, only 3 genes with similar effect size need to be simultaneously pathogenic to become ill. Theoretically, “rare genes with large OR” corresponds to the “large N and very small T” situation in the model if de novo mutations are not considered.

The low ORs in association studies suggest that the effect of any single disease gene is minimal. In that case, even if schizophrenic individuals have a low reproductive fitness, each disease gene would work as a near-neutral gene and not be extinguished by purifying selection during human evolution. Genes with a large effect size presented in the early human evolution should have been eliminated had it not been for the new mutations (Hughes, 2008).

4.3. Prevalence in offspring and other relatives

In a recent report (Gottesman et al., 2010), the cumulative incidence of schizophrenia in the GP was 1.12% [95% confidence interval (CI), 1.09–1.14], a value adopted for this study. It also reported the prevalence in PP × PP offspring as 27.3% (95% CI, 18.3–36.2). The present data fit this value when penetrance = 0.5 and $T \leq 0.7$ (best with $T = 0.5$) regardless of N, although if N is large then T may be smaller. The reported value for PP × NP offspring, 7.0% (95% CI, 6.4–7.7), corresponds to the present data with penetrance = 0.5 and T = 0.5 to 0.8. However, for NP × NP offspring, the reported value was 0.86% (95% CI, 0.83–0.88), whereas results of the present study were higher than that (Fig. 4c). Although schizophrenia is a genetic disease, both genetic and non-genetic environmental factors are required for the development of the illness (Maynard et al., 2001). The penetrance in offspring might be affected by non-genetic familial factors. Considerations of shared environmental factors improve the fitness of genetic model (McGue et al., 1985). It can be reasonably assumed that a stable family environment without members affected by the illness would favor non-penetrance of the disease. Moreover if it is assumed that penetrance differs with the genetic load, the prevalence in NP × NP offspring decreases because the mean genetic load of the genetic candidates from NP × NP offspring is lower than that from other populations, including the GP (not shown).

Early studies of familial risk in schizophrenia indicated that the relative risk (λ) of relatives compared to the GP as follows: parents, 17.65; children, 15.4; siblings, 10.6; half-siblings, 4; grandchildren, 3.5; nephews, 2.8; and uncles, 2.5 (Strachan and Read, 1999). Although the baseline prevalence in the GP in these early studies differed from that in the present study, these data are fairly compatible with the present results obtained with T ranging from 0.2 to 0.4 and a penetrance of 0.5.

The λ hierarchy from Strachan (1999) is as follows:

$$Parents \geq Offspring > Siblings \gg Half-siblings \geq Grandchildren > Nephews \geq Uncles$$

The present study yielded a similar hierarchy:

$$Parents \geq Offspring\ (PP \times NP) > Siblings \gg Half-siblings = Grandchildren = Nephews > Uncles$$

4.4. Effects of penetrance

The prevalence for relatives was significantly affected by the penetrance; the best estimates were obtained with a penetrance value of 0.5. However, for other indices, such as gene frequency and OR, variation in the penetrance had little effect. This is because the disease manifests only at the extreme end of the genetic load distribution. At the extreme end of the binomial distribution, a significant increase in the supra-threshold proportion (decreased penetrance) can occur with a minimal change in the event frequency.

Although the default penetrance was simply defined as the monozygotic twin concordance rate, a reasonable *a priori* value of penetrance is difficult to obtain. However the present results suggest that the basic parameter, allele frequency, is not critically dependent upon the hypothetical penetrance. It can also be suggested that, the results might not differ significantly even if some patients were to be misclassified as normal controls. This misclassification would decrease the apparent penetrance, which would not adversely affect the gene frequency.

4.5. Estimated N and T value

The OR (Fig. 4) and the prevalence in relatives (Figs. 5 and 6) indicate a moderate range for T (0.3–0.8). In this model, the genotype frequency in the PP (p_{G_PP}) can serve as a simple estimate for T (Fig. 3d), and the moderate value of T suggests that the disease genes would not be rare ones.

N can be estimated using the OR. When assessed on the basis of allele frequency, an OR of 2 corresponds to an N•T of 14 (Fig. 4b) and if T is as above, N is in the low double digits. When the OR is 1.2, N•T is greater than 150 and N should be several hundred. In the wide range of non-extreme T values, the genotype-wise OR can also be used to estimate N. If the OR is 2 N would be greater than 40 (Fig. 3c) and if it 1.2, N is greater than the maximum value simulated (160). Thus, even estimated conservatively, N would be greater than one hundred.

4.6. Number of subjects required to detect statistically significant differences

As expected, the sample number required to detect statistically significant differences (SN) increases with increasing N. It is because the genetic heterogeneity increases with N, which dilutes the effect of any single gene. If N = 20, the SN is less than 200 per group, a number achievable in relatively small studies. However if N > 100, SN approaches one thousand (Fig. 7a). Thus, even if true differences exist, studies using only a few hundred subjects may not have sufficient power to show statistical significance. Real world GWA studies indicate that SN over thousands is insufficient and this suggests that N may be at least several hundred. In this model, the minor allele frequency (MAF) in the PP (P_{A_PP}) also affects the SN. However, MAF is not a major determinant because T has only limited effect on the SN and P_{G_PP} has nearly the same value as T (Fig. 3d). P_{A_PP} can be estimated from P_{G_PP} as shown in the Appendix A.

The SN did not vary significantly regardless of whether the control group was the GP or those selected by the absence of the disease (NP). Most of the present results were also not affected by the choice of control group (GP or NP), except for the comparison between NP × NP and GP × GP offspring.

4.7. Genetic heterogeneity and the phenocopy phenomenon

The issue of genetic heterogeneity in schizophrenia is not whether heterogeneity exists but *how much* heterogeneity exists (Tsuang and Faraone, 1995). This issue poses a great hindrance to genetic study. In the present model, it was assumed that only a portion of the disease-related genes are required for the expression of the disease phenotype, which means that different members of the PP have different genetic make-ups. Thus, the model is compatible to the genetic heterogeneity. Lower T indicates more heterogeneity and if T is less than 0.5, any two patients could become ill without sharing a common disease gene. The present estimate for T (0.3–0.8) favors the existence of genetic heterogeneity.

Diseases phenotypically similar to schizophrenia, such as bipolar disorder, have been proposed to share disease-related genes with schizophrenia (Berrettini, 2003). If bipolar disorder is similarly determined by mixed influences of genes, mixed-up results of genetic studies of the two disorders can be explained. There may not be any specific “schizophrenia gene” that can be reliably separated from genes causing other phenotypically similar diseases.

4.8. Relevance to the animal model

Many genetic animal studies of schizophrenia are based on mutations of a single gene (Carlson et al., 2011; Fatemi, 2001; Hikida et al.,

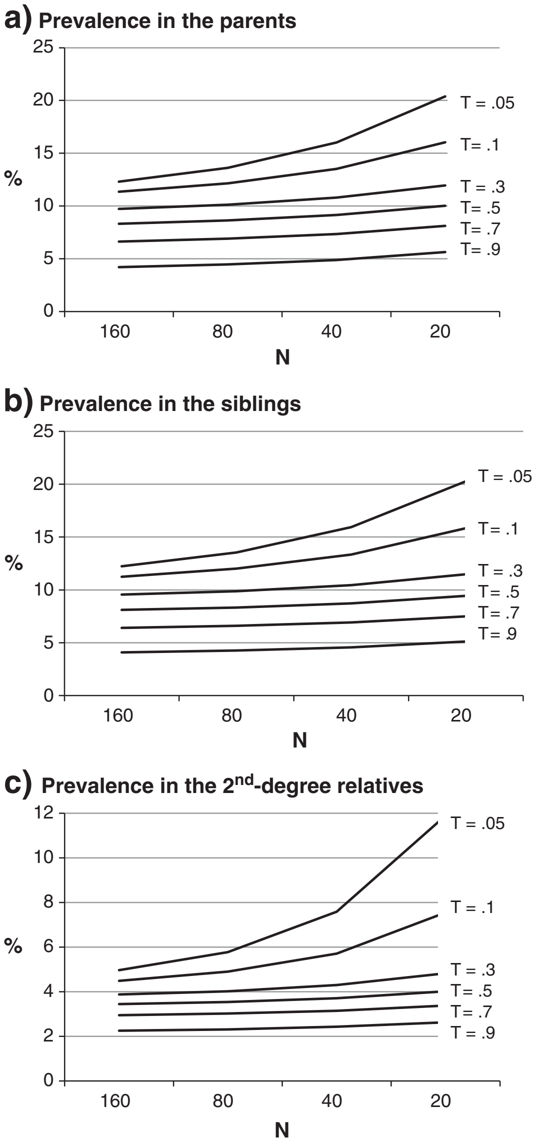


Fig. 6. Prevalence in parents, siblings, and second-degree relatives. (a–c) At a penetrance value of 0.5, the prevalence in parents (a), siblings (b) and second-degree relatives (c) increased with decreasing T. Prevalence was slightly higher in parents than in siblings. Half-siblings, grandchildren, grandparents, and nephews have the same prevalence, which is slightly higher than that of the uncle. N and T are defined as in Fig. 2.

2007). The effect of a single mutation is measured in terms of the animal model of schizophrenia. However schizophrenia is not determined by a single gene. A more plausible experimental approach related to the polygenic threshold model of liability is the quantitative trait loci mapping of a disease trait in a large population of animals as shown in Samocha et al. (2010).

4.9. Conclusions and limitations

(1) A model for the genetic determinants of schizophrenia based on the polygenic threshold model grossly fits the empirical genetic epidemiologic data when the number of genes possibly related to schizophrenia (N) is at least a hundred and the threshold (T) is moderate (0.3–08).
(2) This model can treat genetic heterogeneity.
(3) The gene frequency in the PP, the OR, and the prevalence in relatives can be used to estimate N, T, and penetrance.

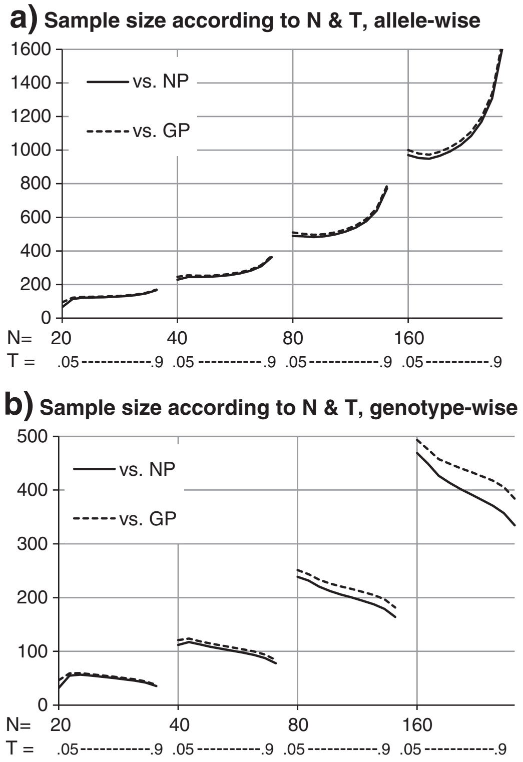


Fig. 7. Sample number (SN) required for statistically significant differences. (a) In the allele-wise assessment, SN increased with increasing N and T, and whether the control group was NP or GP made no difference. (b) In the genotype-wise assessment, SN increased with increasing N but decreased with increasing T. When N was large, NP control group required less samples compared to GP control group. N, T, GP and NP are defined as in Fig. 2.

(4) The effect of a single gene is small, even when it is a real determinant of schizophrenia.
(5) Many patients become ill at the threshold level of genetic load.
(6) Penetrance does not significantly affect the estimated gene frequency in patients or controls.
(7) Obtaining statistically stable results in a case–control study may require several hundred subjects per sample group.

To simplify the model, many unrealistic assumptions were made. The most critical ones were that every disease gene has the same contribution for the disease and that all disease genes have the same frequency. Other unrealistic assumptions were that all disease genes are dominant and that the multiple genes have simple additive effects. Moreover, any de novo mutations was simply ignored. In the future, it may be possible to construct more sophisticated models treating these factors as modifiable variables, but reasonable default values for these variables have not yet been established.

Acknowledgment

This study was supported by KT&G, Republic of Korea.

Appendix A

A.1. Distribution of pathogenic alleles and genotypes in the general population

Assuming Hardy–Weinberg equilibrium (HWE) in the general population (GP) with allele frequency p_{A_GP}, the probability of having pathogenic genotype (p_{G_GP}) is:

$$p_{G\_GP} = 1-(1-p_{A\_GP})^2$$

Since all loci are independent, the distribution of pathologic genes (genetic load) follows a binomial distribution B(N, p_{G_GP}), and the probability that an individual from the GP has a genetic load of k is:

$$p(k)_{\_GP} = {}_{n}C_{k} \cdot p_{G\_GP}{}^{k} \cdot (1 - p_{G\_GP})^{N-k}$$

When the threshold for disease manifestation is T and the penetration is incomplete, the prevalence of disease in the GP is:

$$\text{Prevalence}_{\_GP} = \text{penetrance} \cdot \sum_{k=NT}^{N} p(k)_{\_GP}$$

This value depends upon N, T, p_{G_GP}, and the penetrance. After $\text{Prevalence}_{\_GP}$ was fixed to 0.0112 and the default penetrance was set to 0.5, the value of p_{G_GP} was sought for every value of N and T (Fig. 1). The p_{A_GP} value was calculated from p_{G_GP} assuming that HWE is maintained in the GP.

In this population, the mean genetic load is calculated as:

$$\text{Load}_{\_GP} = \sum_{k=0}^{N} (p(k)_{\_GP} \cdot k)$$

This value is actually the same as N•p_{G_GP}

A.2. Genetic load and allele frequency in the patient population (PP)

Since every individual in the patient population (PP) is selected from the GP and has a genetic load equal to or greater than the product N•T, their summed genetic load can be calculated as:

$$\text{Summed} - \text{load}_{\_PP} = \sum_{k=NT}^{N} (p(k)_{\_GP} \cdot k)$$

This sum is represented by the portion of the GP composed only of potential patients. Thus, to obtain the mean genetic load in individuals from the PP, this sum should be divided by the proportion of potential patients ($\text{Prevalence}_{\_GP}$/penetrance). Then:

$$\begin{aligned}\text{Load}_{\_PP} &= \sum_{k=NT}^{N} (p(k)_{\_GP} \cdot k) / (\text{Prevalence}_{\_GP}/\text{penetrance}) \\ &= \sum_{k=NT}^{N} (p(k)_{\_GP} \cdot k) / \sum_{k=NT}^{N} p(k)_{\_GP}\end{aligned}$$

Because it was assumed that all loci are equal, the probability that a specific locus has pathologic genotype (p_{G_PP}) is:

$$p_{G\_PP} = \text{Load}_{\_PP}/N$$

Finally, because it was assumed that the homozygous/heterozygous ratios in the PP is the same as that in the GP:

$$p_{A\_PP} = p_{G\_PP} \cdot p_{A\_GP}/p_{G\_GP}$$

References

Berrettini W. Evidence for shared susceptibility in bipolar disorder and schizophrenia. Am J Med Genet C Semin Med Genet 2003;123C:59–64.

Carlson GC, Talbot K, Halene TB, Gandal MJ, Kazi HA, Schlosser L, et al. Dysbindin-1 mutant mice implicate reduced fast-phasic inhibition as a final common disease mechanism in schizophrenia. Proc Natl Acad Sci U S A 2011;108:E962–70.

Chen X, Lee G, Maher BS, Fanous AH, Chen J, Zhao Z, et al. GWA study data mining and independent replication identify cardiomyopathy-associated 5 (CMYA5) as a risk gene for schizophrenia. Mol Psychiatry 2011;16:1117–29.

Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann Hum Genet 1965;29:51–76.

Fatemi SH. Reelin mutations in mouse and man: from reeler mouse to schizophrenia, mood disorders, autism and lissencephaly. Mol Psychiatry 2001;6:129–33.

Girard SL, Xiong L, Dion PA, Rouleau GA. Where are the missing pieces of the schizophrenia genetics puzzle? Curr Opin Genet Dev 2011;21:310–6.

Gottesman II, Laursen TM, Bertelsen A, Mortensen PB. Severe mental disorders in offspring with 2 psychiatrically ill parents. Arch Gen Psychiatry 2010;67:252–7.

Grozeva D, Conrad DF, Barnes CP, Hurles M, Owen MJ, O'Donovan MC, et al. Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. Schizophr Res 2012;135:1–7.

Hikida T, Jaaro-Peled H, Seshadri S, Oishi K, Hookway C, Kong S, et al. Dominant-negative DISC1 transgenic mice display schizophrenia-associated phenotypes detected by measures translatable to humans. Proc Natl Acad Sci U S A 2007;104:14,501–6.

Hughes AL. Near neutrality: leading edge of the neutral theory of molecular evolution. Ann N Y Acad Sci 2008;1133:162–79.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45.

International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 2008;455:237–41.

Kyogoku C, Yanagi M, Nishimura K, Sugiyama D, Morinobu A, Fukutake M, et al. Association of calcineurin A gamma subunit (PPP3CC) and early growth response 3 (EGR3) gene polymorphisms with susceptibility to schizophrenia in a Japanese population. Psychiatry Res 2011;185:16–9.

Li D, Collier DA, He L. Meta-analysis shows strong positive association of the neuregulin 1 (NRG1) gene with schizophrenia. Hum Mol Genet 2006;15:1995–2002.

Maynard TM, Sikich L, Lieberman JA, LaMantia AS. Neural development, cell–cell signaling, and the “two-hit” hypothesis of schizophrenia. Schizophr Bull 2001;27:457–76.

McGue M, Gottesman II, Rao DC. Resolving genetic models for the transmission of schizophrenia. Genet Epidemiol 1985;2:99-110.

Meltzer HY. Treatment of schizophrenia and spectrum disorders: pharmacotherapy, psychosocial treatments, and neurotransmitter interactions. Biol Psychiatry 1999;46:1321–7.

Procopio M. Does god play dice with schizophrenia? A probabilistic model for the understanding of causation in mental illness. Med Hypotheses 2005;64:872–7.

Samocha KE, Lim JE, Cheng R, Sokoloff G, Palmer AA. Fine mapping of QTL for prepulse inhibition in LG/J and SM/J mice using F(2) and advanced intercross lines. Genes Brain Behav 2010;9:759–67.

Scott MP, Matusdaria P, Lodish H, Darnell J, Zipursky L, Kaiser CA, et al. Molecular cell biology. 5th ed. San Francisco: W. H. Freeman; 2004. p. 396.

Siever LJ, Davis KL. The pathophysiology of schizophrenia disorders: perspectives from the spectrum. Am J Psychiatry 2004;161:398–413.

Strachan T, Read AP. Human molecular genetics. 2nd ed. New York: Wiley-Liss; 1999.

Tsuang MT, Faraone SV. The case for heterogeneity in the etiology of schizophrenia. Schizophr Res 1995;17:161–75.

Tsuang MT, Stone WS, Seidman LJ, Faraone SV, Zimmet S, Wojcik J, et al. Treatment of nonpsychotic relatives of patients with schizophrenia: four case studies. Biol Psychiatry 1999;45:1412–8.

Wang H, Chow SC, Li G. On sample size calculation based on odds ratio in clinical trials. J Biopharm Stat 2002;12:471–83.

Weller MP. The spectrum of schizophrenia. Postgrad Med J 1987;63:1021–4.

Wilkie AO. The molecular basis of genetic dominance. J Med Genet 1994;31:89–98.

Xu M, He L. Convergent evidence shows a positive association of interleukin-1 gene complex locus with susceptibility to schizophrenia in the Caucasian population. Schizophr Res 2010;120:131–42.