SCIENTIFIC REPORTS

OPEN

Optimal Trend Tests for Genetic Association Studies of Heterogeneous Diseases

Received: 03 November 2015
Accepted: 24 May 2016
Published: 09 June 2016

Wen-Chung Lee

The Cochran-Armitage trend test is a standard procedure in genetic association studies. It is a directed test with high power to detect genetic effects that follow the gene-dosage model. In this paper, the author proposes optimal trend tests for genetic association studies of heterogeneous diseases. Monte-Carlo simulations show that the power gain of the optimal trend tests over the conventional Cochran-Armitage trend test is striking when the genetic effects are heterogeneous. The easy-to-use R 3.1.2 software (R Foundation for Statistical Computing, Vienna, Austria) code is provided. The optimal trend tests are recommended for routine use.

Genetic factors contribute to many human diseases, conferring susceptibility or resistance. Unlike simple Mendelian disorders, more common complex diseases may have many genes involved in their pathogenesis[1–3]. The association of candidate genes (or markers across the genome) with a disease can be efficiently evaluated by a case-control design, in which genotype frequencies are compared for diseased cases and unaffected controls. Genetic association studies are the important first step of gene characterization. Candidate genes or markers found to be statistically significant are then subject to further studies, to identify causal variants, to quantify genetic effects, to examine possible gene-environment or gene-gene interactions, and so on[4–7]; results from different studies can also be pooled for a meta-analysis[8–10]. The Cochran-Armitage trend test[11–15] has become a standard procedure in this crucial first-step study of complex diseases. It is a directed test most sensitive to detecting genetic effects that follow the gene-dosage model.

However, a disease may comprise more than one disease entity, each with a different etiology, clinical picture and prognosis. Examples of such heterogeneous diseases are Alzheimer's disease[16], breast tumors[17], B-cell lymphoma[18], acute lymphoblastic leukemia[19], primary thyroid lymphoma[20], otosclerosis[21], rheumatoid arthritis[22], and autism spectrum disorder[1]. The effect of a gene associated with a heterogeneous disease can be variable, depending on which disease entity one is considering; and if the distinct disease entities themselves, often obscure and subtle, are not recognized and taken into account, the genetic effect associated with the heterogeneous disease at large may vary from person to person.

Genetic heterogeneity can complicate our association study of complex diseases even further. The following hypothetical example should highlight this issue. Consider the disease occurrences in a population of one million people (250,000 people with genotype *aa*; 500,000 people with genotype *Aa*; 250,000 people with genotype *AA*). Assume that the disease under study has two distinct subtypes (which are unknown to researchers). Further assume that both subtypes conform strictly to the gene-dosage model. For Subtype I, the disease risk is 0.0001 for the *aa* genotype, and the risk increases ten-fold per *A* allele; for Subtype II, the disease risk is 0.0020 for the *aa* genotype, and the risk decreases two-fold per *A* allele. A simple calculation shows that the majority (73%) of the diseased subjects in this population are of Subtype I (where the risk increases ten-fold per *A* allele), so the *A* allele should be regarded as a risk allele rather than a protective one. Yet, ignoring the subtypes, we observe disease risks of 0.0021 (*aa* genotype), 0.0020 (*Aa* genotype), and 0.0105 (*AA* genotype), respectively. This is nothing like a gene-dosage model, and moreover, the *A* allele now appears protective, when comparing the *Aa* and the *aa* genotypes. Obviously, applying the standard Cochran-Armitage trend test[11–15] to this setting will result in power loss.

In this paper, we propose optimal trend tests for genetic association studies of heterogeneous diseases.

Research Center for Genes, Environment and Human Health and Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan. Correspondence and requests for materials should be addressed to W.-C.L. (email: wenchung@ntu.edu.tw)

	aa	*Aa*	*AA*	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Table 1. Genotype distribution for case-control studies.

Methods

Notation. For a marker with two alleles *a* and *A*, each individual in a case-control study is genotyped with one of three genotypes, *aa*, *Aa* and *AA* (indexed by $i = 0, 1, 2$, respectively). Assume that the case-control study consists of a total of $n = r + s$ subjects (r cases and s controls). These n subjects can be classified into a 2×3 table based on each subject's genotype and disease status as shown in Table 1.

Let $(x_0, x_1, x_2) = (0, c, 1)$ where the coefficient c can assume any value. Under the null hypothesis of no genetic association, the following test statistic is distributed asymptotically as a chi-square distribution with one degree of freedom:

$$Z^2(c) = \frac{r \times s}{n} \times \left[\sum_{k=0}^{2} x_k \times \left(\frac{r_k}{r} - \frac{s_k}{s}\right)\right]^2 \Bigg/ \left[\left(\sum_{k=0}^{2} x_k^2 \times \frac{n_k}{n}\right) - \left(\sum_{k=0}^{2} x_k \times \frac{n_k}{n}\right)^2\right]. \tag{1}$$

The test with a coefficient of 0.5, $Z(0.5)$, is the familiar Cochran-Armitage trend test[11–15].

Optimal Trend Test. Assume that the non-diseased population is in Hardy-Weinberg equilibrium with an allele frequency (for the *A* allele) of q. The expected genotype frequencies for the controls are then, respectively,

$$\begin{cases} q_0 = (1 - q)^2, \\ q_1 = 2 \times q \times (1 - q), \\ q_2 = q^2. \end{cases} \tag{2}$$

Further assume that the genetic effect is heterogeneous; the allele relative risk (relative risk per *A* allele) is not a constant value but may vary from person to person. Let the expected value of this relative risk be denoted as RR, its coefficient of variation (standard deviation divided by mean; a measure of heterogeneity), as $\mathrm{CV}_{\mathrm{RR}}$. The expected allele frequency for the cases is then

$$p = \frac{q \times \mathrm{RR}}{1 - q + q \times \mathrm{RR}}, \tag{3}$$

and its variance, calculated by a Taylor approximation (S1 Exhibit), is then

$$\mathrm{Var}(p) = [p \times (1 - p) \times \mathrm{CV}_{\mathrm{RR}}]^2 \tag{4}$$

This variance is also the Hardy-Weinberg disequilibrium coefficient in the diseased population, and therefore, the expected genotype frequencies for the cases are, respectively,

$$\begin{cases} p_0 = (1 - p)^2 + \delta, \\ p_1 = 2 \times p \times (1 - p) - 2 \times \delta, \\ p_2 = p^2 + \delta, \end{cases} \tag{5}$$

where $\delta = \mathrm{Var}(p)$.

In the above calculations, we assumed Hardy-Weinberg equilibrium for the non-diseased population and a gene-dosage genetic model (a constant increase or decrease in risk per *A* allele). We now alleviate these assumptions. In general, the expected genotype frequencies for the controls are, respectively,

$$\begin{cases} q_0 = (1 - q)^2 + \Delta, \\ q_1 = 2 \times q \times (1 - q) - 2 \times \Delta, \\ q_2 = q^2 + \Delta, \end{cases} \tag{6}$$

where Δ is the Hardy-Weinberg disequilibrium coefficient in the non-diseased population. The expected genotype relative risks are, respectively,

$$\begin{cases} \mathrm{RR}^{2\gamma}\ (Aa \text{ vs. } aa), \\ \mathrm{RR}^2\ (AA \text{ vs. } aa), \end{cases} \tag{7}$$

	Val/Val	Val/Ala	Ala/Ala	Total
Cases	307	509	184	1000
Controls	359	522	137	1018
Total	666	1031	321	2018

Table 2. Association between the *adenosine diphosphate ribosyltransferase* (*ADPRT*) gene (Val762Ala polymorphism) and lung cancer risk (data taken from ref. 23).

where γ is a genetic model parameter. $\gamma = 0$ corresponds to an autosomal recessive model, $\gamma = 0.5$, a gene-dosage model, and $\gamma = 1$, an autosomal dominant model. As before, we allow the parameter RR to have a coefficient of variation $\mathrm{CV}_{\mathrm{RR}}$, and the parameter p (though here it may not be interpreted as the expected allele frequency for the cases) to have a variance as prescribed in Equation (4). Under these conditions, the expected genotype frequencies for the cases (p_0, p_1 and p_2) can be derived from a Taylor expansion. The formulas are rather cumbersome and are relegated to S2 Exhibit.

With the p_i and q_i calculated for $i = 0$, 1 and 2, simple algebra shows that the following optimal coefficient will maximize the test statistic in Equation (1):

$$c^{\mathrm{optimal}} = \left(\frac{p_1 - q_1}{f_1} - \frac{p_0 - q_0}{f_0}\right) \Big/ \left(\frac{p_2 - q_2}{f_2} - \frac{p_0 - q_0}{f_0}\right), \tag{8}$$

where

$$f_i = \frac{1}{n} \times (r \times p_i + s \times q_i), \tag{9}$$

for $i = 0$, 1 and 2, respectively, are the expected genotype frequencies in the total case-control sample. $Z(c^{\mathrm{optimal}})$ is our proposed optimal trend test.

An Example.

We use published case-control data to demonstrate our method. Zhang *et al.*[23] examined the association between the *adenosine diphosphate ribosyltransferase* (*ADPRT*) gene (Val762Ala polymorphism) and lung cancer risk. The data (1000 cases and 1018 controls) are shown in Table 2.

For simplicity, we assume Hardy-Weinberg equilibrium for the non-diseased population (with an allele frequency of $q = 0.4$) and a gene-dosage genetic model for the *ADPRT* gene (with a weak association of RR = 1.25 and a moderate heterogeneity of $\mathrm{CV}_{\mathrm{RR}} = 0.4$). Using [2]~[5], we then calculate $q_0 = (1 - 0.4)^2 = 0.36$, $q_1 = 2 \times 0.4 \times (1 - 0.4) = 0.48$, $q_2 = 0.4^2 = 0.16$, $p = \frac{0.4 \times 1.25}{1 - 0.4 + 0.4 \times 1.25} = 0.45$, $\delta = \mathrm{Var}(p) = [0.45 \times (1 - 0.45) \times 0.4]^2 = 0.0098$, $p_0 = (1 - 0.45)^2 + 0.0098 = 0.31$, $p_1 = 2 \times 0.45\ (1 - 0.45) - 2 \times 0.0098 = 0.48$ and $p_2 = 0.45^2 + 0.0098 = 0.22$, respectively.

Using [9], we calculate the expected genotype frequencies in the total case-control sample as $f_0 = \frac{1000 \times 0.31 + 1018 \times 0.36}{2018} = 0.33$, $f_1 = \frac{1000 \times 0.48 + 1018 \times 0.48}{2018} = 0.48$, and $f_2 = \frac{1000 \times 0.22 + 1018 \times 0.16}{2018} = 0.19$, respectively. Using [8], we calculate the optimal coefficient for this example as $c^{\mathrm{optimal}} = \left(\frac{0.48 - 0.48}{0.48} - \frac{0.31 - 0.36}{0.33}\right) \Big/ \left(\frac{0.22 - 0.16}{0.19} - \frac{0.31 - 0.36}{0.33}\right) = 0.33.$

Using [1], we then calculate

$$\begin{aligned} Z^2(0.33) \;=\; & \frac{1000 \times 1018}{2018} \times \left[0.33 \times \left(\frac{509}{1000} - \frac{522}{1018}\right) + \left(\frac{184}{1000} - \frac{137}{1018}\right)\right]^2 \\ & \div \left[\left(0.33^2 \times \frac{1031}{2018} + \frac{321}{2018}\right) - \left(0.33 \times \frac{1031}{2018} + \frac{321}{2018}\right)^2\right] = 10.9. \end{aligned}$$

From this, we obtain a very small p-value of 0.00095. By comparison, the conventional Cochran-Armitage trend test for this example results in a higher p-value of 0.00164. Zhang *et al.*[23] used a chi-square test with two degrees of freedom, which resulted in an even higher p-value of 0.00420. Such differences in p-values should not be taken lightly, considering that a severe multiple-testing penalty often has to be made before declaring significance in a genetic association study.

Simulation Study.

We perform a simulation study to examine the statistical properties of the optimal trend test. The non-diseased population is assumed to be in Hardy-Weinberg equilibrium ($\Delta = 0$), with an allele frequency of $q = 0.4$. We assume a gene-dosage genetic model ($\gamma = 0.5$), and we consider situations where the *A* allele is a risk allele (RR = 2, 1.5, and 1.25, respectively) and a protective allele (RR = 0.5, 0.67, 0.8, respectively), in turn. For each scenario, we use a sample-size formula for the Cochran-Armitage trend test[13] to calculate the respective sample size needed for a case-control study (assuming an equal number of cases and controls) to achieve a power of 0.8 at a significance level of 0.05.

We consider various values of $\mathrm{CV}_{\mathrm{RR}}$: 0.0 (no heterogeneity), 0.1, 0.2,…, 1.0 (profound heterogeneity). For each value of q, RRand $\mathrm{CV}_{\mathrm{RR}}$, we use Equation (8) to calculate the optimal coefficient. We then perform Monte-Carlo simulations (a total of 1,000,000 simulations for each scenario) to calculate the empirical power of the optimal trend test (at the sample sizes described above). For comparison, we also calculate the empirical power of the Cochran-Armitage trend test.

Figure 1 presents the results when the *A* allele is a risk allele (panels A, C, and E for the coefficients; panels B, D and F for the empirical powers). When the genetic effect is homogeneous ($\mathrm{CV}_{\mathrm{RR}} = 0$), the optimal coefficients as

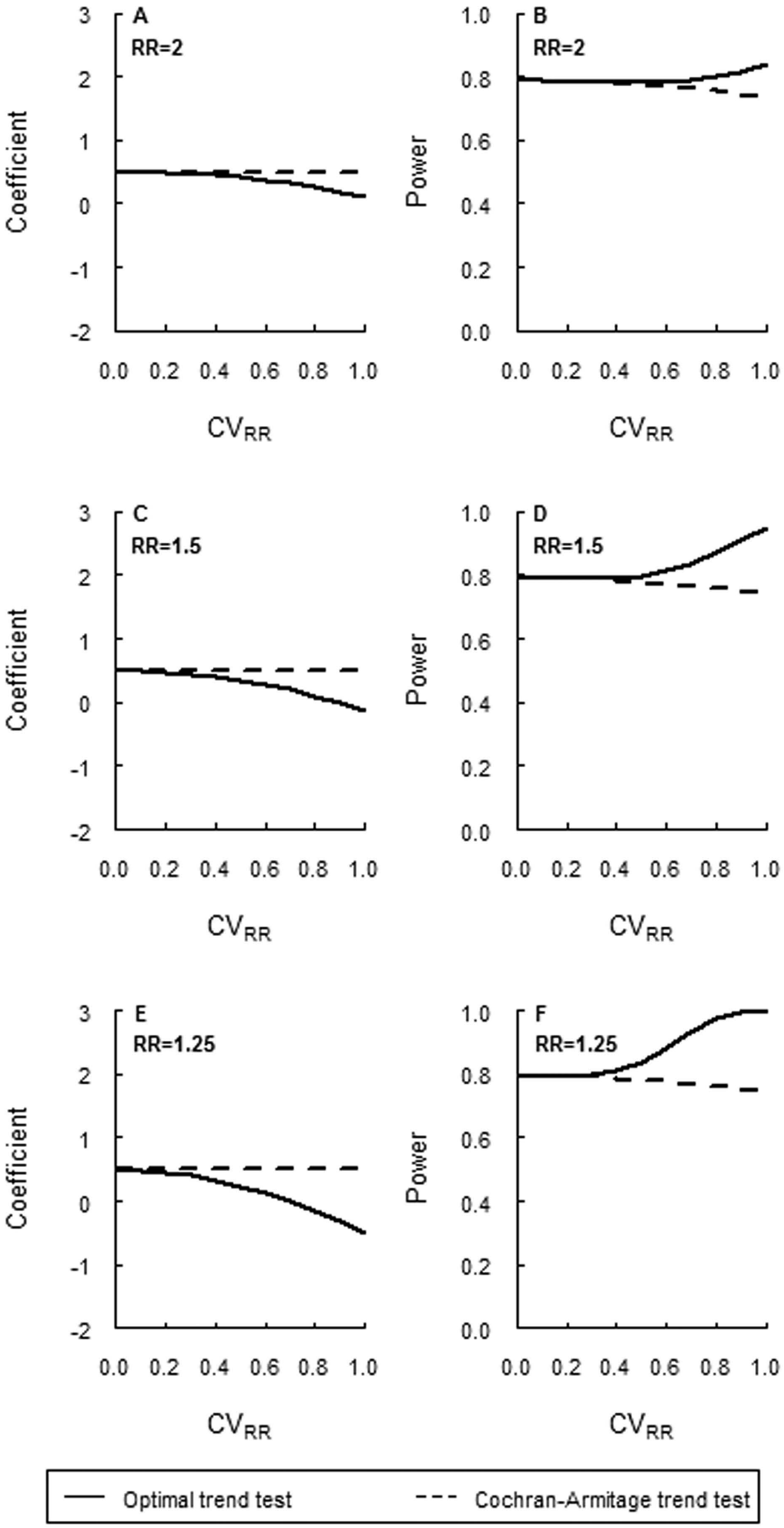

Figure 1. Simulation results for a risk allele ((**A,B**): RR = 2; (**C,D**): RR = 1.5; (**E,F**): RR = 1.25; solid lines: the optimal trend test; dash lines: Cochran-Armitage tend test).

calculated from Equation (8) are very close to the coefficient of the Cochran-Armitage trend test, namely, 0.5. As a result, the powers of the optimal trend test and the Cochran-Armitage trend test are very similar. As the genetic effect becomes more heterogeneous (larger CV_{RR}), the optimal coefficient decreases (down to below zero), and the power of the optimal trend test increases (up to ~100%). The rates of the coefficient decrease/power increase are more striking for a weaker genetic effect (RR = 1.25; panels E and F) than for a stronger genetic effect (RR = 2;

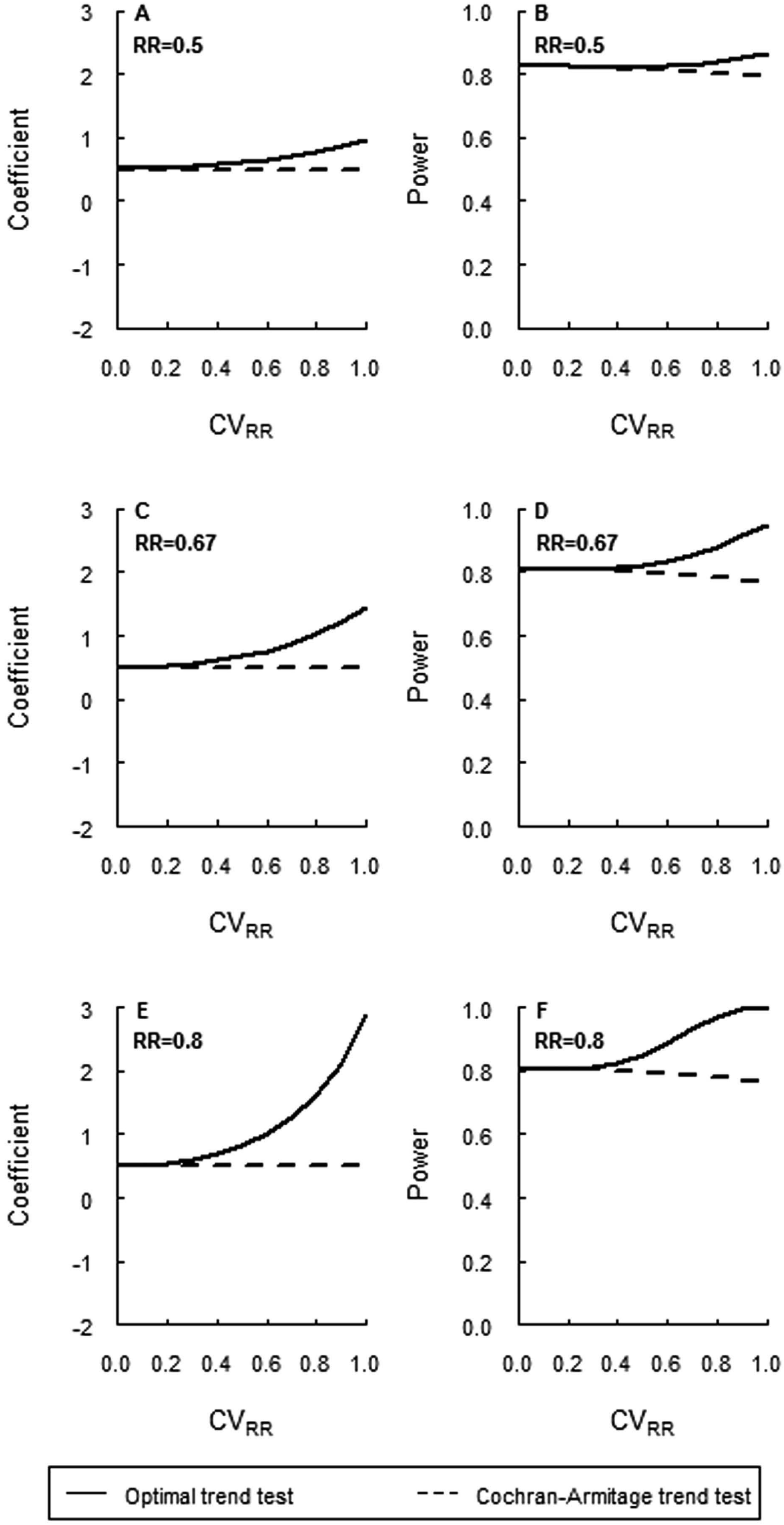

Figure 2. Simulation results for a protective allele ((**A,B**): RR = 0.5; (**C,D**): RR = 0.67; (**E,F**): RR = 0.8; solid lines: the optimal trend test; dash lines: Cochran-Armitage tend test).

panels A and B). By comparison, the Cochran-Armitage trend test uses a constant coefficient of 0.5, and its power decreases gradually with greater heterogeneity.

Figure 2 presents the results when the *A* allele is a protective allele. Similar findings can be seen in Fig. 1 when *A* is a risk allele, except that as the genetic effect becomes more heterogeneous, the optimal coefficient deviates away from 0.5 in the other direction, increasing up to beyond 1.0 rather than decreasing.

We consider different values of q, Δ andγ, and the results (S3 Exhibit) all show a superiority of the optimal trend test over the conventional Cochran-Armitage trend test.

Discussion

The optimal trend test as proposed in this paper is a directed test that is most sensitive for a particular specified alternative. The optimal coefficient depends on the effect of the study gene (mean RR, variability CV_{RR} and genetic model γ) and on the underlying population (allele frequency q, and Hardy-Weinberg disequilibrium coefficient Δ). This *a priori* information is to be supplied by researchers, either by a literature search or an educated guess. As shown in this study, the power gain over the conventional Cochran-Armitage trend test is striking when the genetic effects are very heterogeneous.

Sometimes, to pinpoint exactly one set of RR, CV_{RR}, γ, q and Δ, calculating the optimal coefficient can be difficult, but suggesting a list of possible sets of parameter values may be easier. Assuming that a researcher comes up with a total of m sets of parameter values, he/she can input these into our Equation (8) to calculate a total of m optimal coefficients, $c_1^{\text{optimal}}, \ldots, c_m^{\text{optimal}}$ and then input these into our Equation (1) for a total of m optimal trend tests. Next, a summary test can be performed based on a weighted sum of these m test statistics:

$$Z_{\text{summary}}^2 = w_1 \times Z^2(c_1^{\text{optimal}}) + \ldots + w_m \times Z^2(c_m^{\text{optimal}}), \tag{10}$$

where $w_1, \ldots, w_m$ are the weights given to reflect the plausibility of each set of parameter values. The multiple testing problem should not concern us here, because we make one and only one summary test. Under the null hypothesis of no genetic association, Z_{summary}^2 is distributed asymptotically as a mixture of chi-square variables (detailed in S4 Exhibit). (The test reduces to the optimal trend test in this paper when $m = 1$)

The proposed optimal trend tests (and the summary test) are easy to calculate. S5 Exhibit presents the R 3.1.2 software (R Foundation for Statistical Computing, Vienna, Austria) code and a number of worked examples. The R program also allows for the direct input of the optimal coefficients. For example, if one suspects a gene-dosage model with heterogeneous effects, one can input one coefficient slightly above 0.5, say $c_1 = 0.8$, another coefficient slightly below 0.5, say $c_2 = 0.2$ and $w_1 = w_2 = 1$, to the R program to test $Z_{\text{summary}}^2 = Z^2(0.8) + Z^2(0.2)$. As another example, if one is uncertain about the genetic model, one can input $c_1 = 0.5$ (gene dosage), $c_2 = 1$ (autosomal dominant), $c_3 = 0$ (autosomal recessive), and $w_1 = w_2 = w_3 = 1$ into the R program to test $Z_{\text{summary}}^2 = Z^2(0.5) + Z^2(1) + Z^2(0)$.

References

1. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156,** 872–877 (2014).
2. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153,** 17–37 (2013).
3. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–424 (2015).
4. Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6,** 287–298 (2005).
5. Le Marchand, L. & Wilkens, L. R. Design considerations for genomic association studies: importance of gene-environment interactions. *Cancer Epidemiol. Biomarkers Prev.* **17,** 263–267 (2008).
6. Lewis, C. M. & Knight, J. Introduction to genetic association studies., *Cold Spring Harb. Protoc.* **2012,** 297–306 (2012).
7. Rava, M. *et al.* Selection of genes for gene-environment interaction studies: a candidate pathway-based strategy using asthma as an example. *Environ. Health* **12,** 56 (2013).
8. Thompson, J. R., Attia, J. & Minelli, C. The meta-analysis of genome-wide association studies. *Brief Bioinform.* **12,** 259–269 (2011).
9. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14,** 379–389 (2013).
10. Pharoah, P. D. P. *et al.* GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat. Genet.* **45,** 362–370 (2013).
11. Cochran, W. G. Some methods for strengthening the common chi-square tests. *Biometrics* **10,** 417–451 (1954).
12. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11,** 375–386 (1955).
13. Slager, S. L. & Schaid, D. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum. Hered.* **52,** 149–153 (2001).
14. Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J. L. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* **53,** 146–152 (2002).
15. Zheng, G. & Gastwirth, J. L. On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies. *Stat. Med.* **25,** 3150–3159 (2006).
16. Corder, E. H. & Woodbury, M. A. Genetic heterogeneity in Alzheimer's disease: a grade of membership analysis. *Genet. Epidemiol.* **10,** 495–499 (1993).
17. Perou, C. M. *et al.* Molecular portraits of human breast tumors. *Nature* **406,** 747–752 (2000).
18. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene-expression profiling. *Nature* **403,** 503–511 (2000).
19. Yeoh, E. J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene-expression profiling. *Cancer Cell* **1,** 133–143 (2002).
20. Thieblemont, C. *et al.* Primary thyroid lymphoma is a heterogeneous disease. *J. Clin. Endocrinol. Metab.* **87,** 105–111 (2002).
21. Van der Bogaert, K. *et al.* Otosclerosis: a genetically heterogeneous disease involving at least three different genes. *Bone* **30,** 624–630 (2002).
22. van der Pouw Kraan, T. C. *et al.* Rheumatoid arthritis is a heterogeneous disease: evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues. *Arthritis Rheum.* **48,** 2132–2145 (2003).
23. Zhang, X. *et al.* Polymorphisms in DNA base excision repair genes *ADPRT* and *XRCC1* and risk of lung cancer. *Cancer Res.* **65,** 722–726 (2005).

Acknowledgements

This paper is partly supported by grants from Ministry of Science and Technology, Taiwan (NSC 102-2628-B-002-036-MY3) and National Taiwan University, Taiwan (NTU-CESRP-102R7622-8). No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The author declares no competing financial interest.

How to cite this article: Lee, W.-C. Optimal Trend Tests for Genetic Association Studies of Heterogeneous Diseases. *Sci. Rep.* **6**, 27821; doi: 10.1038/srep27821 (2016).