Annals of human genetics

Short communication

doi: 10.1111/j.1469-1809.2011.00700.x

Blindly Using Wald's Test Can Miss Rare Disease-Causal Variants in Case-Control Association Studies

Guan Xing[1], Chang-Yun Lin[2], Stephen P. Wooding[2] and Chao Xing[2,3]*

[1]*Bristol-Myers Squibb Company, Pennington, NJ*

[2]*McDermott Center of Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX*

[3]*Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX*

Summary

There are four tests – the likelihood ratio (LR) test, Wald's test, the score test and the exact test – commonly employed in genetic association studies. On comparison of the four tests, we found that Wald's test, popular in genome-wide screens due to its low computational demands, exhibited a paradoxical behaviour in that the test statistic decreased as the effect size of the variant increased, resulting in a loss of power. The LR test always achieved the most significant *P*-values, followed by the exact test. We further examined the results in a real data set composed of high- and low-cholesterol subjects from the Dallas Heart Study (DHS). We also compared the single-variant LR test with two multi-variant analysis approaches – the burden test and the C-alpha test – in analysing the sequencing data by simulation. Our results call for caution in using Wald's test in genome-wide case-control association studies and suggest that the LR test is a better alternative in spite of its computational demands.

Keywords: Case-control study, Wald's test, low-frequency variants

Introduction

To test for association between genotype and phenotype in a case-control study, one can either employ the exact test or fit a logistic regression model to the data and test whether the coefficient for genotype is zero or not. There are three asymptotic tests that are commonly employed: the likelihood ratio (LR) test, Wald's test and the score test. The LR test compares the null and alternative hypotheses on an equal basis, while Wald's test starts at the alternative and considers movement towards the null and the score test begins with the null and asks whether movement towards the alternative could be an improvement. The three tests have equivalent asymptotic power for testing local alternatives (Cox & Hinkley, 1974). From a computational standpoint, the LR test is most demanding because it requires both the restricted and unrestricted estimates of parameters, whereas Wald's test uses only the unrestricted estimates and the score test uses only the restricted estimates. Besides hypothesis testing, investigators are also interested in estimating a variant's odds ratio $e^{\hat{\beta}}$, where $\hat{\beta}$ is the unrestricted maximum likelihood estimate (MLE) of the coefficient for genotype. Computer programs often produce $\hat{\beta}$ and its estimated variance $\hat{var}(\hat{\beta})$, which makes it convenient to compute the Wald's test statistic $(\hat{\beta} - \beta_{\text{null}})^2/\hat{var}(\hat{\beta})$ to test the null hypothesis $\beta = \beta_{\text{null}}$. Thus, Wald's test is often the default option – for example, *–logistic* command in PLINK (Purcell et al., 2007) – in a genome-wide scan; in particular, when covariates are present.

However, we notice an anomalous behaviour of Wald's test; if a variant is mainly present in cases or controls, which means large effect sizes under the alternative hypothesis, Wald's test generates an insignificant *P*-value. On the contrary, the other two tests produce significant *P*-values. This abnormal phenomenon of Wald's test may have been observed by many researchers, but its theoretical interpretation is less understood; in a binary logit model, as the distance between the parameter estimate and the null value increases, the test statistic decreases to zero and the power of the test diminishes to the test size (Hauck & Donner, 1977). This aberrant behaviour of Wald's test is particularly pertinent to low-frequency variants. Suppose a causal variant with high penetrance is present at low frequency in the cases and nearly absent from the controls; the power of Wald's test will be minimal even if the effect size

*Corresponding author: Chao Xing, Ph.D., McDermott Center of Human Growth and Development, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas TX 75390, USA. Tel: 214-648-1695; Fax: 214-648-1666; E-mail: chao.xing@utsouthwestern.edu

Table 1 Distribution of cases and controls by genotype.

	Genotype		
Phenotype	*AA*	*Aa*	Total
Case	r_0	r_1	R
Control	s_0	s_1	S
Total	n_0	n_1	N

estimate of the variants is large. Were Wald's test employed, the causal variant would not show statistically significant association with the disease, and one could miss the association by only screening a list of *P*-values. Thus, alternative tests – the LR test, the score test and the exact test – should be considered in this situation. In this paper, we compared the four tests in terms of both validity when a variant is at low frequency, and power when a low-frequency variant is highly penetrant.

Methods

Consider a test for association between a low-frequency single-nucleotide polymorphism (SNP) and the disease affection status in a case-control study. Denote by A and a, the major and minor alleles of the site, respectively, with the frequency of a sufficiently low that we only observe genotypes AA and Aa, but not aa, in the sample. Data can then be summarised into a 2×2 contingency table (Table 1). Denote by Y_i, the affection status of individual i, and $Y_i = 1$ or 0 indicates individual i being a case or control. Denote by X_i, the genotypic value of individual i, and $X_i = 1$ or 0 indicates individual i being Aa or AA. To test for association between the genotype and phenotype, we fit to the data a logistic regression model $\ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X$, where $\pi = P(Y = 1) = \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$. The hypotheses to be tested are $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Denote by $l(\cdot)$ the log-likelihood function, by $\hat{\beta}_1$ the MLE of β_1, by $S(\cdot)$ the score statistic and by $I(\cdot)$ the information matrix. The LR test statistic is defined as $LRT = 2[l(\hat{\beta}_1) - l(\beta_1 = 0)]$. Wald's test statistic is defined as $WT = \hat{\beta}_1^2 / \hat{var}(\hat{\beta}_1)$, where $\hat{var}(\hat{\beta}_1) = \hat{I}_{22}(\cdot)^{-1}$ is the estimated variance of $\hat{\beta}_1$. The score test is defined as $ST = S(\beta_1 = 0)^T I(\beta_1 = 0)^{-1} S(\beta_1 = 0)$. Note that the widely used Armitage's trend test (Armitage, 1955; Sasieni, 1997) and Pearson's χ^2 test (Pearson, 1900) in genetic studies are score tests, and for data in Table 1, both tests lead to the same test statistic $X^2 = \frac{N(Nr_1 - Rn_1)^2}{RSn_0n_1}$ as ST. All three statistics – LRT, WT and ST – follow an asymptotic distribution of χ_1^2, and therefore they should have approximately equivalent power, though a systemic inequality $WT \geq LRT \geq ST$ exists (Berndt & Savin, 1977). Given a sample, when β_1 is far from 0, we desire adequate power from all three tests to reject the null hypothesis. However, it was shown that $\lim_{|\hat{\beta}_1| \to \infty} WT = 0$ (Hauck & Donner, 1977), and therefore the power of Wald's test diminishes to the test level in this situation. Another commonly used test is Fisher's exact test (Fisher, 1925), which calculates the exact significance level by assuming a hypergeometric distribution for contingency tables. When a sample size is small, Fisher's exact test is often preferred to the other three tests that rely on asymptotic theories to evaluate significance levels.

Although all four tests are assumed to maintain proper test sizes in general, in case of low-frequency exposure variables their properties are unclear. We first compared their type I error rates in testing for association of low-frequency variants by large-scale simulations. We considered a balanced design with an equal number (500, 1000 and 2000) of cases and controls, and a SNP with the minor allele frequency (MAF) equal to 0.005. In each setting, 100,000 replicates, in each of which the SNP was polymorphic, were generated. We performed the four tests on all the datasets. Except in the case of the exact test, we calculated *P*-values assuming an asymptotic distribution of χ_1^2. The empirical type I error rates at four nominal levels ($\alpha \in \{0.05, 0.01, 0.001, 0.0001\}$) were calculated as the proportion of the 100,000 replicates for which the *P*-value was less than or equal to α.

Second, we compared the significance levels attained by the four tests as the effect size of a low-frequency variant increased. In this paper, we define the effect size as relative risk, i.e. $\frac{r_1 n_0}{r_0 n_1}$. Consider a balanced design with 2000 cases and 2000 controls, and a SNP with the MAF at most equal to 0.005 in cases and even less frequent in controls, i.e. $r_1 \leq 20$ and $s_1 < r_1$ in Table 1. Given r_1, as s_1 decreased, the effect size of allele a increased. We compared the significance levels that the four tests could attain for each combination of r_1 and s_1.

Third, we considered a special scenario in which the frequency of a highly penetrant disease-causal variant was constrained to be low under selection pressure such that in a sample of a case-control study this variant appeared only scarcely in cases, but not at all in controls, i.e. $s_1 = 0$ and $0 < r_1 << r_0$ in Table 1. We investigated three sample sizes $R = S \in \{500, 1000, 2000\}$. In each setting, we fixed s_1 to be 0, constrained the MAF of the variant in cases less than or equal to 0.005, and enumerated all the possibilities, i.e. $1 \leq r_1 \leq \frac{R}{100}$. We compared the significance levels that the four tests could attain in all of the 35 possible situations.

In a genetic study, it is often necessary to adjust for covariates such as known risk factors and confounding factors (Xing & Xing, 2010). Therefore, besides the genetic variant, we also simulated a binary covariate mimicking an environmental factor independent of the genetic variant with the exposure rates of 0.3 and 0.2 in cases and controls, respectively. Wald's test, the LR test and the score test are readily

applicable to multiple logistic regression models. As to the counterpart of Fisher's exact test, we performed the conditional exact inference by enumerating the exact distributions of sufficient statistics for the parameter of interest conditional on the remaining parameters in a logistic regression model (Hirji et al., 1987; Cox & Snell, 1989), as implemented in the SAS procedure LOGISTIC (Derr, 2009).

Fourth, as a proof-of-principle example, we performed Wald's test, the LR test and the score test in a genome-wide association (GWA) study of plasma levels of low-density lipoproteins (LDL) cholesterol with nonsynonymous SNPs in the African Americans (AAs) of the Dallas Heart Study (DHS), and compared their performance in testing for low-frequency variants present only in the upper quintile of the population. The DHS is a multiethnic, population-based cohort in Dallas County (Victor et al., 2004). In this study, we focused on the AAs. There were 1722 individuals with complete phenotypes – LDL, body mass index (BMI), age, sex – after deleting those taking cholesterol-lowering medicine. We sampled the upper and lower quintiles ($N = 345$) as cases and controls, respectively. There were a total of 8968 nonsynonymous SNPs assayed across the autosomes. We first filtered out singletons and those out of Hardy–Weinberg equilibrium (P-value $< 1.0 \times 10^{-5}$) in the whole population, then filtered out monomorphic ones in the case-control sample, resulting in 8263 SNPs for further analysis. The data were fit by a logistic regression model $\ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \times G + \beta_2 \times Age + \beta_3 \times Sex + \beta_4 \times BMI + \beta_5 \times Ancestry$, where π was probability of being affected, G was genotypic value coded in an additive genetic model with 0, 1 and 2 denoting major allele homozygote, heterozygote and minor allele homozygote, respectively, and ancestry of each individual was inferred using ancestry-informative markers as described elsewhere (Romeo et al., 2008). We tested whether $\beta_1 = 0$ by Wald's test, the LR test and the score test. The exact test was not performed because of memory constraints when multiple covariates were present.

Fifth, we evaluated the performance of the single-variant LR test in analyzing sequencing data by simulation studies based on the results of a pooled sequencing study. Neale et al. (2011, see their table 1) reported results of pooled sequencing of the *ApoB* gene in 96 individuals with high triglyceride levels exceeding the 5% upper tail of the population distribution and 96 individuals with low triglyceride levels below the 5% lower tail. There were a total of 27 nonsynonymous variants detected, among which 10 were singletons with 6 in the upper tail and 4 in the lower tail. The distribution of singletons between the two tails was relatively balanced, which contributed little information in testing association between the gene and phenotype; therefore, we simulated data based on the distribution of the remaining 17 variants. Five sample sizes – $n \times 96$, $n \in \{1, 2, 3, 4, 5\}$, cases and controls, respectively – were considered. For each variant, we fixed the total number of counts as n times the number observed in the original data and decided its distribution between cases and controls by a binomial trial. The binomial parameter p equalled 0.5 if the variant was designated as neutral; otherwise it equalled its MLE in the observed data, and the MLE was set to 0.005 (or 0.995) when a variant was only observed in one tail. Two types of simulation schemes were carried out (Table 2). In the first type (A1–A3), a mixture of risk, protective and neutral variants were simulated. In particular, in scheme A1 all 17 variants were simulated at $p =$ MLE. In scheme A2 there were nine variants simulated at $p =$ MLE and others at $p =$ 0.5; the variant with the largest effect size was N1914S, whose occurrence counts ratio (CR) between the upper and lower tails was $0 : 5n$. In scheme A3, P1143S replaced N1914S as the variant with the largest effect size ($CR = 0 : 6n$). In the second type (B1–B3), in addition to neutral variants, only variants acting in the same direction were simulated. The difference among schemes was the variant with the largest

Table 2 Simulation schemes based on pooled sequencing of ApoB in the upper and lower 5% tails of the distribution of triglyceride levels[a].

	Number of counts		Simulation scheme[b]					
Variant	Upper tail	Lower tail	A1	A2	A3	B1	B2	B3
A4481T	2	5	√	√	√	√	√	√
I4314V	3	0	√	√	√	×	×	×
R4270T	6	3	√	√	√	×	×	×
V4128M	1	7	√	×	×	×	×	×
T3388K	2	1	√	×	×	×	×	×
S3203Y	6	0	√	×	×	×	×	×
L2404I	2	3	√	×	×	×	×	×
E2391D	2	2	√	×	×	×	×	×
T2373N	2	2	√	×	×	×	×	×
V2313I	2	1	√	×	×	×	×	×
H1923R	6	12	√	√	√	√	√	√
N1914S	0	5	√	√	×	×	√	×
D1871N	2	0	√	√	√	×	×	×
P1143S	0	6	√	×	√	×	×	√
R1128H	0	3	√	√	√	√	×	×
D1113H	1	3	√	√	√	√	√	√
T498N	2	0	√	√	√	×	×	×

[a]The original result was reported in Table 1 of Neale et al. (2011), in which there were 96 individuals sequenced in each tail. In the simulation study, five sample sizes – $n \times 96$, $n \in (1, 2, 3, 4, 5)$, cases and controls, respectively – were considered.

[b]For each variant the total number of counts was fixed at n times the number observed in the original data. Symbols √ and × denote a variant was simulated with the binomial parameter being the MLE of the observed data and a half, respectively.

effect size – $CR = 0:3n$, $0:5n$ and $0:6n$ for B1, B2 and B3, respectively. In each scheme 1000 replicates were generated and analyzed by the single-variant LR test, a version of burden tests (Morris & Zeggini, 2010) and the C-alpha test (Neale et al., 2011). For both the burden test and the *C*-alpha test, a single *P*-value for the gene could be obtained; for the LR test, the minimum of the 17 single-variant Bonferroni-corrected *P*-values was designated as the *P*-value for the gene. The empirical power at α level was calculated as the proportion of the 1000 replicates for which the *P*-value was less than or equal to α.

Results

The empirical type I error rates of the four tests in case of low-frequency variants are summarised in Table 3. The LR test showed a slightly supranominal test size when the sample size was 500; as the sample size increased to 2000, it maintained a proper test size. In contrast, the other three tests all showed an infranominal test size even when the sample size was as large as 2000. The score test and the exact test had similar type I error rates, and the former was more conservative when there was no covariate, but the latter was more conservative when there was a covariate. Compared with these two tests, Wald's test became more and more conservative when the nominal level turned more stringent.

As the effect size of a low-frequency variant increased, Wald's test showed an aberrant behaviour of attaining less significant levels (Table 4). Given r_1, the effect size of *a* allele increased as s_1 decreased, and we would expect a test to attain more significant levels. However, when $r_1 < 15$, the *P*-values for Wald's test increased as s_1 decreased from 1 to 0; when $r_1 \geq 15$, the *P*-values increased as s_1 decreased from 2 to 1 and then to 0. In particular, when $s_1 = 0$, the effect size of *a* allele approached infinity, and Wald's test led to *P*-values ≥ 0.9. This anomalous behaviour of Wald's test can be visualised in Figure 1. Given $r_1 = 20$, when $s_1 = 19$, allele *a* had no detectable effect on the trait. As s_1 decreased, the effect size of allele *a* increased, as did its estimate $\hat{\beta}_1$ and the standard error of $\hat{\beta}_1$. WT also increased and reached its maximum when $s_1 = 2$. As s_1 further decreased, the increase of $\hat{\beta}_1$ was slower than that of its the standard error, and WT started decreasing. In the extreme case of $s_1 = 0$, WT was approaching zero, too. In the extreme situations where a variant appeared only scarcely in cases but not at all in controls, with the increase of allele *a* effect size, i.e. the number of cases with genotype *Aa* given a sample size, and the increase of sample size given a certain effect size, Wald's test always remained at levels ≥ 0.9 (Table 5).

In contrast to Wald's test, the LR test, the score test and the exact test all attained more significant levels as the effect size of allele *a* increased (Tables 4 and 5). Among the four tests, the LR test always outperformed the others. Without covariates, the exact test produced *P*-values slightly smaller than those of the score test; but this trend was not obvious when a covariate was included, possibly because of random noises in simulation.

In the DHS there were 75 variants present in the cases but not in the controls, out of which two variants appeared eight times, six variants appeared five times, four variants appeared four times, 14 variants appeared three times, 21 variants appeared twice and the remaining 28 variants appeared only once. All of the 12 SNPs manifesting four times or more in

Table 3 Empirical type I error rates[a] of Wald's test, likelihood ratio (LR) test, score test and exact test for a variant with the minor allele frequency (MAF) equal to 0.005.

		Without covariates[b]				With a covariate			
Sample size	Test size	Wald's test	LR test	Score test	Exact test	Wald's test	LR test	Score test	Exact test
500	0.05	0.019	0.070	0.021	0.022	0.018	0.059	0.048	0.030
	0.01	0.000	0.015	0.002	0.003	0.000	0.014	0.007	0.005
	0.001	0.0000	0.0022	0.0001	0.0003	0.0000	0.0016	0.0003	0.0004
	0.0001	0.00000	0.00020	0.00000	0.00000	0.00000	0.00020	0.00000	0.00001
1000	0.05	0.040	0.052	0.025	0.028	0.040	0.054	0.050	0.038
	0.01	0.003	0.012	0.004	0.005	0.004	0.011	0.009	0.007
	0.001	0.0000	0.0015	0.0003	0.0006	0.0000	0.0012	0.0006	0.0006
	0.0001	0.00000	0.00018	0.00001	0.00005	0.00000	0.00019	0.00001	0.00002
2000	0.05	0.046	0.052	0.034	0.035	0.045	0.051	0.049	0.041
	0.01	0.008	0.011	0.006	0.006	0.007	0.011	0.010	0.008
	0.001	0.0003	0.0011	0.0004	0.0006	0.0004	0.0011	0.0009	0.0008
	0.0001	0.00000	0.00011	0.00004	0.00004	0.00000	0.00014	0.00007	0.0006

[a]Based on 100,000 replicates, in each of which the variant was polymorphic and the number of cases/controls equalled that in the first column.
[b]The covariate was binary with exposure rates of 0.3 and 0.2 in cases and controls, respectively.

Table 4 Comparison of significance levels[a] attained by Wald's test, likelihood ratio (LR) test, score test and exact test as the effect size of a rare variant increases[b].

No. cases with *Aa*	No. controls with *Aa*	Without covariates				With a covariate[c]			
		Wald's test[d]	LR test	Score test	Exact test	Wald's test[d]	LR test	Score test	Exact test
10	0	***9.36E–01***	1.94E–04	4.38E–03	1.93E–03	***9.35E–01***	2.24E–04	1.76E–03	1.49E–03
	1	2.79E–02	3.42E–03	1.57E–02	1.16E–02	2.87E–02	3.85E–03	7.34E–03	9.65E–03
	2	3.74E–02	1.57E–02	4.30E–02	3.83E–02	3.88E–02	1.70E–02	2.23E–02	3.12E–02
	3	6.67E–02	4.58E–02	9.56E–02	9.18E–02	6.89E–02	4.81E–02	5.42E–02	7.24E–02
	4	1.20E–01	1.03E–01	1.81E–01	1.79E–01	1.25E–01	1.07E–01	1.13E–01	1.15E–01
	5	2.05E–01	1.92E–01	3.01E–01	3.01E–01	2.10E–01	1.98E–01	2.02E–01	2.10E–01
12	0	***9.30E–01***	4.44E–05	1.47E–03	4.80E–04	***9.35E–01***	5.36E–05	6.20E–04	3.85E–04
	1	1.67E–02	9.11E–04	5.47E–03	3.37E–03	1.72E–02	1.06E–03	2.59E–03	2.67E–03
	2	1.87E–02	4.81E–03	1.60E–02	1.28E–02	1.98E–02	5.45E–03	8.35E–03	1.06E–02
	3	3.13E–02	1.60E–02	3.85E–02	3.48E–02	3.25E–02	1.71E–02	2.12E–02	2.92E–02
	4	5.63E–02	4.04E–02	7.95E–02	7.62E–02	5.91E–02	4.30E–02	4.79E–02	6.18E–02
	5	9.90E–02	8.42E–02	1.45E–01	1.43E–01	1.03E–01	8.78E–02	9.25E–02	1.33E–01
15	0	***9.49E–01***	4.96E–06	2.93E–04	5.94E–05	***9.48E–01***	6.13E–06	1.27E–04	4.93E–05
	1	***8.43E–03***	1.23E–04	1.13E–03	5.07E–04	***8.56E–03***	1.43E–04	5.21E–04	4.17E–04
	2	7.30E–03	7.77E–04	3.54E–03	2.30E–03	7.64E–03	9.05E–04	1.82E–03	1.89E–03
	3	1.07E–02	3.06E–03	9.36E–03	7.41E–03	1.12E–02	3.41E–03	5.07E–03	6.11E–03
	4	1.85E–02	9.05E–03	2.15E–02	1.89E–02	1.95E–02	9.91E–03	1.25E–02	1.29E–02
	5	3.27E–02	2.19E–02	4.36E–02	4.09E–02	3.55E–02	2.43E–02	2.76E–02	3.27E–02
18	0	***9.44E–01***	5.63E–07	5.92E–05	7.34E–06	***9.44E–01***	7.08E–07	2.61E–05	5.95E–06
	1	***4.80E–03***	1.63E–05	2.34E–04	7.37E–05	***4.91E–03***	2.01E–05	1.12E–04	6.39E–05
	2	3.11E–03	1.20E–04	7.72E–04	3.90E–04	3.28E–03	1.46E–04	4.02E–04	3.40E–04
	3	3.94E–03	5.48E–04	2.19E–03	1.45E–03	4.16E–03	6.35E–04	1.18E–03	1.22E–03
	4	6.34E–03	1.86E–03	5.45E–03	4.24E–03	6.79E–03	2.12E–03	3.14E–03	3.12E–03
	5	1.10E–02	5.10E–03	1.21E–02	1.04E–02	1.19E–02	5.79E–03	7.40E–03	7.68E–03
20	0	***9.41E–01***	1.33E–07	2.05E–05	1.82E–06	***9.40E–01***	1.77E–07	2.59E–05	1.60E–06
	1	***3.37E–03***	4.22E–06	8.21E–05	2.01E–05	***3.49E–03***	5.62E–06	4.22E–05	1.77E–05
	2	1.84E–03	3.41E–05	2.79E–04	1.17E–04	1.93E–03	4.23E–05	1.46E–04	1.01E–04
	3	2.08E–03	1.69E–04	8.20E–04	4.72E–04	2.24E–03	2.02E–04	4.46E–04	4.16E–04
	4	3.20E–03	6.22E–04	2.13E–03	1.50E–03	3.42E–03	7.26E–04	1.22E–03	9.50E–04
	5	5.38E–03	1.85E–03	4.97E–03	3.97E–03	5.77E–03	2.09E–03	2.94E–03	2.73E–03

[a]Except the exact test, *P*-values were calculated under the asymptotic distribution of χ_1^2; when there was a covariate, the mean *P*-values of 1000 replicates were reported.

[b]The structure of data was as Table 1 with $R = S = 2000$ and the numbers in the first and second columns corresponding to r_1 and s_1, respectively. The maximal minor allele frequency (MAF) was 0.005.

[c]The covariate was binary with exposure rates of 0.3 and 0.2 in cases and controls, respectively.

[d]The aberrant *P*-values by Wald's test were in bold and italic font.

the cases attained a nominal *P*-value less than 0.05 by the LR test, and we listed their *P*-values along with the ranks for all three tests in Table 6. By the LR test, two SNPs ranked in the top 20 observed *P*-values, four ranked in the top 50 and seven ranked in the top 100 in the genome scan; by the score test two ranked in the top 100; in contrast, by Wald's test all 12 SNPs remained at levels ≥ 0.9.

We compared the power of the burden test, C-alpha test and LR test in analyzing sequencing data under different sample sizes at an arbitrary level of 0.0001 (Fig. 2), which was chosen such that the power of tests in all simulation situations spread from 0 to 1. In scheme A1 data was simulated mimicking the results of the original sequencing study, in which both risk and protective variants were present. At the original sample size, the power of the C-alpha test was 0.343, whereas that of the LR test and burden test was close to 0; when the sample size was doubled, the power of the C-alpha test rapidly increased to 1, whereas that of the other two tests only had a mild increase; when the sample size was tripled, the power of the LR test increased to 0.997, whereas that of the burden test was still less than 0.1. Although both risk and protective variants were also present in schemes A2 and

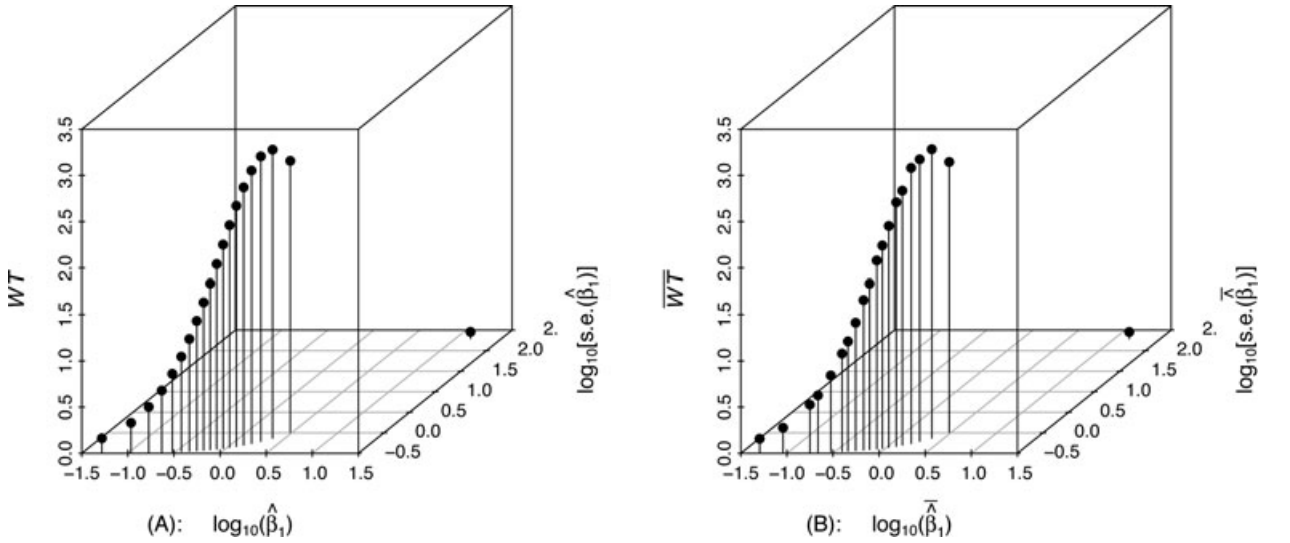

Figure 1 Maximum likelihood estimate (MLE) of coefficient, its estimated standard deviation and Wald's test statistic in a logistic regression model.
(A): There was no additional covariate. (B): A binary covariate with exposure rates of 0.3 and 0.2 in cases and controls, respectively, was generated in 1000 replicates, and the mean estimates were plotted. The sample size was 2000 cases and controls, respectively. The number of cases with genotype *Aa* was fixed to be 20. The number of controls with genotype *Aa* decreased by 1 from 19 to 0, corresponding to the 20 points from left to right along the *x*-axis. Note that when the number of controls with genotype *Aa* equalled 0, the MLE of coefficient was infinity. In the figure, we substituted it with a value that met the convergence tolerance criterion (10^{-8}) by Fisher's scoring algorithm as implemented in R (R Development Core Team, 2011).

Table 5 Comparison of significance levels[a] attained by Wald's test, likelihood ratio (LR) test, score test and exact test when rare variants appear only in cases[b].

		Without covariates				With a covariate[c]			
Sample size	No. cases with *Aa*	Wald's test	LR test	Score test	Exact test	Wald's test	LR test	Score test	Exact test
500	1	9.69E–01	2.39E–01	1.00E–00	1.00E–00	9.69E–01	2.47E–01	3.26E–01	6.34E–01
	3	9.65E–01	4.12E–02	2.48E–01	2.49E–01	9.68E–01	4.37E–02	8.71E–02	1.99E–01
	5	9.71E–01	8.35E–03	7.29E–02	6.19E–02	9.70E–01	8.91E–03	4.27E–02	2.64E–02
1000	1	9.69E–01	2.39E–01	1.00E–00	1.00E–00	9.69E–01	2.45E–01	3.23E–01	6.23E–01
	5	9.55E–01	8.41E–03	7.33E–02	6.22E–02	9.54E–01	9.15E–03	4.35E–02	2.71E–02
	10	9.58E–01	1.91E–04	4.33E–03	1.91E–03	9.58E–01	2.27E–04	1.78E–03	1.53E–03
2000	1	9.53E–01	2.39E–01	1.00E–00	1.00E–00	9.53E–01	2.47E–01	3.26E–01	6.35E–01
	5	9.55E–01	8.44E–03	7.35E–02	6.23E–02	9.54E–01	9.08E–03	2.70E–02	4.32E–02
	10	9.36E–01	1.94E–04	4.38E–03	1.93E–03	9.35E–01	2.24E–04	1.76E–03	1.49E–03
	15	9.49E–01	4.96E–06	2.93E–04	5.94E–05	9.48E–01	6.13E–06	1.27E–04	4.93E–05
	20	9.41E–01	1.33E–07	2.05E–05	1.82E–06	9.40E–01	1.77E–07	2.59E–05	1.60E–06

[a]Except the exact test, *P*-values were calculated under the asymptotic distribution of χ_1^2; when there was a covariate, the mean *P*-values of 1000 replicates were reported.
[b]The structure of the table is as Table 1, with $s_1 = 0$, $1 \le r_1 \le \frac{R}{100}$ and the numbers in the first column corresponding to R, and also S.
[c]The covariate was binary with exposure rates of 0.3 and 0.2 in cases and controls, respectively.

A3, the distribution of variants was less dispersed than that of scheme A1, which weakened the power of the C-alpha test at certain sample size ranges. At the original sample size, the power of all three tests was close to 0; when the sample size was doubled, the power of the C-alpha test increased to ∼0.6, whereas that of the other two tests was nearly unchanged; when the sample size was tripled, the power of the C-alpha test was close to 1 and that of the LR test rapidly increased to ∼0.9, whereas the power of the burden test was still less than 0.1.

Table 6 Significant nonsynonymous variants present only in cases at a nominal level of 0.05 by the likelihood ratio (LR) test in a case-control study of low-density lipoproteins (LDL) cholesterol in the Dallas Heart Study (DHS) African Americans[a].

Identification[b]			Wald's test		LR test		Score test	
SNP	Count[c]	MAF[d]	*P*-value	Rank[e]	*P*-value	Rank[e]	*P*-value	Rank[e]
p4759986	8	0.0043	9.75E–01	7920	1.37E–03	11	7.35E–03	58
p1694771	8	0.0042	9.76E–01	7931	1.40E–03	12	7.48E–03	59
p3419794	5	0.0033	9.79E–01	8007	4.36E–03	34	1.35E–02	111
p4779548	5	0.0020	9.70E–01	7831	5.92E–03	49	1.78E–02	160
p4222011	5	0.0054	9.80E–01	8031	9.00E–03	81	2.65E–02	248
p4652837	5	0.0053	9.78E–01	7980	1.14E–02	98	4.23E–02	374
p4004951	5	0.0051	9.71E–01	7845	1.39E–02	130	4.13E–02	362
p4776211	5	0.0027	9.71E–01	7858	1.82E–02	184	5.27E–02	468
p4129917	4	0.0038	9.72E–01	7878	1.07E–02	94	2.58E–02	238
p4752843	4	0.0031	9.73E–01	7902	1.26E–02	118	3.00E–02	281
p1629497	4	0.0039	9.73E–01	7901	2.74E–02	279	6.70E–02	584
p4229120	4	0.0031	9.73E–01	7904	3.31E–02	323	7.89E–02	678

[a]The upper and lower quintiles of the DHS African American population by LDL were treated as cases and controls ($N = 345$), respectively.
[b]Perlegen identification.
[c]Copies of variants in cases. All carriers were heterozygotes except for one minor allele homozygote for p4652837.
[d]Minor allele frequency (MAF) in the DHS African American population.
[e]Out of 8263 variants in ascending order.

In schemes B1–B3 multiple variants acted in the same direction. The burden test was more powerful than in schemes A1–A3, and it was more powerful than the other two tests when the sample size was small. With the increase of sample size, both the C-alpha test and the LR test outperformed the burden test. The C-alpha test suffered power loss when the effects were less dispersed compared to those in schemes A1–A3. The power of all three tests increased as the effect sizes of variants increased from scheme B1 to B2 and then to B3; however, unlike the power of the other two tests that depends on the collective effects of multiple variants, for a given sample size, the power of the LR test could dramatically increase when the largest effect size of variants increased to a certain level.

Discussion

A conventional recommendation in testing for small-sample categorical data, in particular, when the smallest expected counts in a cell is less than five (Cochran, 1952), is to use the exact test instead of the score test. However, both tests are conservative (Campbell, 2007). Small observed counts have more impact on the LR test statistic than on the score test statistic such that the latter is recommended in small-sample datasets because the former has inflated type I error rate (Larntz, 1978). In this paper, we consider a situation of small expected/observed counts in a large-sample dataset, which is the case for low-frequency variants in genetic studies. We find that both the score test and the exact test are still conservative, whereas the LR test statistic maintains a proper size. Wald's test is very conservative at stringent test levels. Therefore, regardless of the aberrant behaviour of Wald's test under the extreme alternative hypothesis that we report in this paper, the LR test is preferable in large-scale genetic studies of low-frequency variants.

Our results suggest that blindly using Wald's test in association studies carries unrecognised risks of failing to identify low-frequency disease-causal variants. This issue concerns both GWA and sequencing data. The power of a genetic test depends on the sample size, a variant's MAF, its effect size and its degree of linkage disequilibrium with the disease-casual variation (Chapman et al., 2003). A conventional wisdom in analyzing sequencing data, which are mainly composed of low-frequency variants, is that a single-variant analysis has insufficient power due to low MAFs and that one has to aggregate multiple rare variants to detect significant associations. Implicit assumptions behind this rationale include: the sample size is limited, a variant is extremely rare, or its effect size is moderate to small. However, in the case of low-frequency variants with large effect sizes, the single-variant analysis can still achieve genome-wide significance, in particular, when the disease-causal variant, either chip-genotyped or *in silico*-imputed, is examined. For example, in three recently published Icelandic population-based whole-genome sequencing/imputation studies, a rare missense variant c.2161C>T (MAF = 0.0038) in *MYH6* was

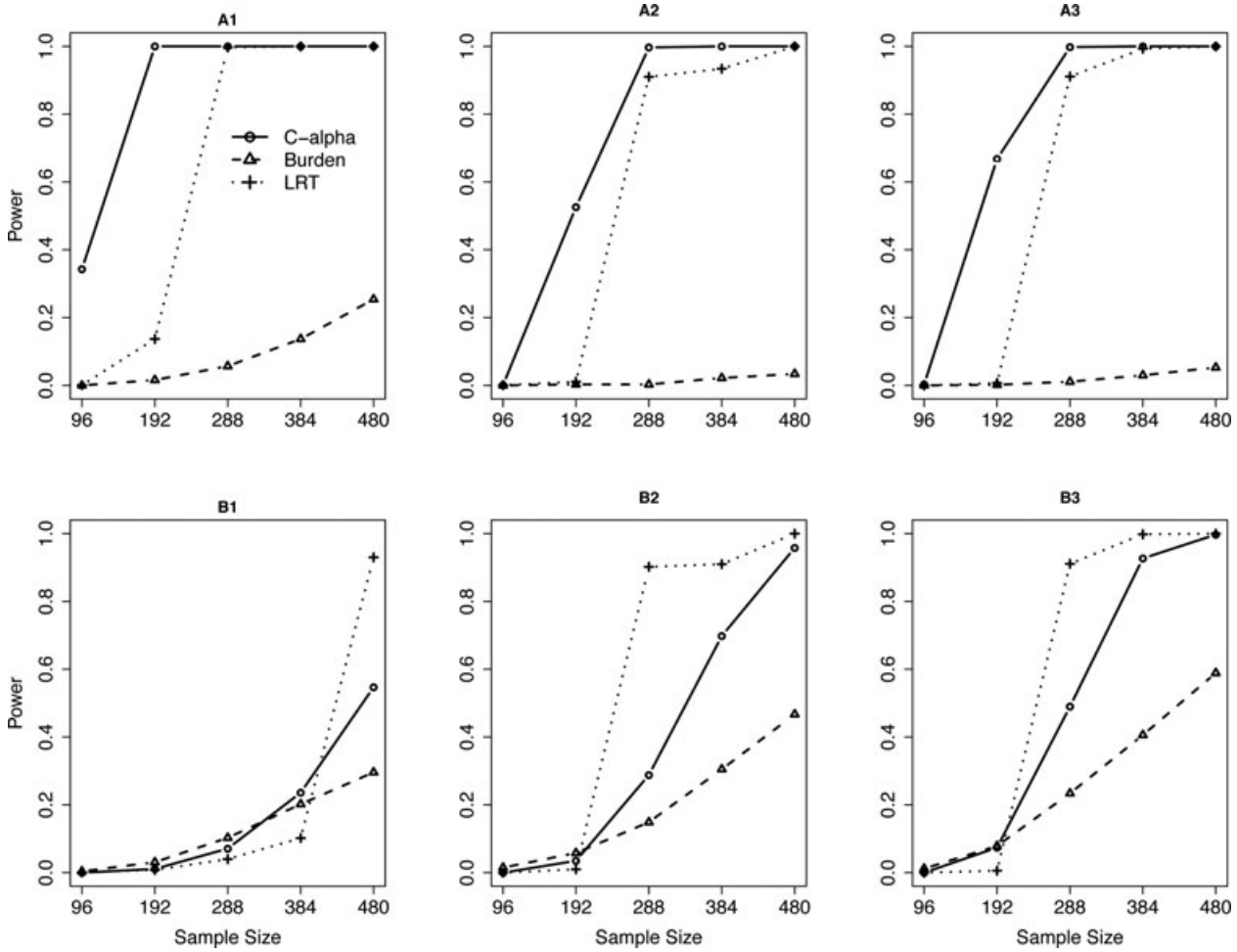

Figure 2 Empirical power comparison of the burden test, C-alpha test and single-variant likelihood ratio (LR) test in analyzing sequencing data. x-axis denotes the sample size at each tail, and y-axis denotes the empirical power at an arbitrary level of 0.0001, which was calculated as the proportion of the 1000 replicates for which the P-value was less than or equal to 0.0001. For both the burden test and the C-alpha test, a single P-value for the gene was obtained; for the LR test, the minimum of the 17 single-variant Bonferroni-corrected P-values was designated as the P-value for the gene.

found to be associated with sick sinus syndrome (OR = 12.53, P-value = 1.5×10^{-29}) (Holm et al., 2011); a rare frameshift mutation c.2040_2041insTT (MAF = 0.0041) in *BRIP1* was found to be associated with ovarian cancer (OR = 8.13, P-value = 2.8×10^{-14}) (Rafnar et al., 2011); and a low-frequency missense variant c.1580C > G (MAF = 0.019) in *ALDH16A1* was found to be associated with gout (OR = 3.12, P-value = 1.5×10^{-16}) and serum uric acid levels (P-value = 4.5×10^{-21}) (Sulem et al., 2011). Note that all these three novel variants were identified by the single-variant analysis.

Given a small sample size, the power to detect association for a single low-frequency variant is limited even though the LR test is employed. However, with large consortia set up nowadays, true association, in particular, for those disease-causal variants with large effect sizes, is still likely to be distinctive. For example, the *ApoB* sequencing study on the 192 individuals, referred to in the Methods, was based on the cardiovascular cohort of the Malmö Diet and Cancer Study consisting of ~6000 individuals (Kathiresan et al., 2008). There were two variants (S3203Y and P1143S) having six copies in one tail but not in the other, and the single-variant LR test produced a P-value of 3.53×10^{-3}. Suppose they were causal variants with large effect sizes and their expressivity was stable, then a joint analysis of three, four and five similar studies would generate Fisher's combined P-values (Fisher, 1925) of 7.1×10^{-6}, 3.4×10^{-7} and 1.7×10^{-8}, respectively. Thus, even without a single cohort as large

as the Icelandic studies, genome-wide significance can still be achieved for low-frequency disease-causal variants by the consortia already established. Note, however, that such signals could be missed if Wald's test were employed.

Based on the data from a pooled sequencing study, we performed proof-of-principle simulations to demonstrate the characteristics of two multi-variant analysis approaches – the burden test and the C-alpha test – and the single-variant LR test. The simulation settings were chosen not for the intention of undermining the power of multivariant analysis approaches, but for the purpose of characterizing each method. The power of the burden test is sensitive to the mean effects of multiple variants, whereas the power of the C-alpha test is sensitive to the dispersion of the effects; on the contrary, the single-variant LR test is robust to the distribution of multiple variants' effects, but is sensitive to the largest effect size among the variants. Simulation studies (Neale et al., 2011; Wu et al., 2011) had shown that multivariant analysis on sequencing data required a sample size of at least thousands to attain the genome-wide significance level (10^{-6}) and that the power was dependent on the distribution of effects of multiple variants. This might be the reason why no novel gene/locus has yet been reported for complex traits as a result of analysing sequencing data using the multivariant analysis approaches. Even for the extremely large gene *ApoB*, which is composed of 4536 amino-acid residues and is known to be associated with lipid levels, selective genotyping of 192 individuals with extreme phenotypes from a cohort of ~6000 and analysis using the C-alpha method only generated a mean *P*-value at the level of 10^{-3} based on our simulation of 1000 replicates. We speculate that, similar to the GWA studies, analysis of a large sample size, made possible by the formation of large consortia, will still be the key to success in the sequencing era. Meanwhile, as long as there is a variant with a large effect size and stable expressivity, the single-variant analysis approach has the potential to identify it, as demonstrated by the Icelandic studies (Holm et al., 2011; Rafnar et al., 2011; Sulem et al., 2011).

In summary, in this study, we have shown that the statistical weaknesses of Wald's test are not merely a side note, but are likely to be a significant issue in many realistic situations. Our results support the use of alternative approaches, particularly LR tests, which are not susceptible to such problems, especially as ongoing advances in computational capabilities continue to reduce obstacles to intensive analyses.

Acknowledgements

We thank the reviewers for constructive comments which tremendously improved the manuscript, as well as Dr. Nathan Morris for critically reading, commenting and editing an early version of the manuscript. We also thank Dr. Helen Hobbs for granting us permission to use the DHS data. This study is supported by the American Heart Association Scientist Development Grant (No. 10SDG4220051) to C.X.

References

Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

Berndt, E.R. & Savin, N.E. (1977) Conflict among criteria for testing hypotheses in the Multivariate Linear Regression Model. *Econometrica* **45**, 1263–1278.

Campbell, I. (2007) Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Stat Med* **26**, 3661–3675.

Chapman, J.M., Cooper, J.D., Todd, J.A., & Clayton, D.G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18–31.

Cochran, W.G. (1952) The chi-squared test of goodness of fit. *Ann Math Stat* **25**, 315–345.

Cox, D.R. & Hinkley, D.V. (1974) *Theoretical statistics.* London: Chapman & Hall/CRC.

Cox, D.R. & Snell, E.J. (1989) *Analysis of binary Data*, 2nd ed. London: Chapman & Hall/CRC.

Derr, R.E. (2009) Performing exact logistic regerssion with the SAS system – revised 2009. http://support.sas.com/rnd/app/papers/exaclogistic2009.pdf.

Fisher, R.A. (1925) *Statistical methods for research workers*, 14th ed., 1970. Edinburgh: Oliver & Boyd.

Hauck, W.W. & Donner, A. (1977) Wald's test as applied to hypotheses in logit analysis. *J Am Stat Assoc* **72**, 851–853.

Hirji, K.F., Mehta, C.R., & Patel, N.R. (1987) Computing distributions for exact logistic regression. *J Am Stat Assoc* **82**, 1110–1117.

Holm, H., Gudbjartsson, D.F., Sulem, P., Masson, G., Helgadottir, H.T., Zanon, C., Magnusson, O.T., Helgason, A., Saemundsdottir, J., Gylfason, A., Stefansdottir, H., Gretarsdottir, S., Matthiasson, S.E., Thorgeirsson, G.M., Jonasdottir, A., Sigurdsson, A., Stefansson, H., Werge, T., Rafnar, T., Kiemeney, L.A., Parvez, B., Muhammad, R., Roden, D.M., Darbar, D., Thorleifsson, G., Walters, G.B., Kong, A., Thorsteinsdottir, U., Arnar, D.O., & Stefansson, K. (2011) A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat Genet* **43**, 316–320.

Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C., Hirschhorn, J.N., Berglund, G., Hedblad, B., Groop, L., Altshuler, D.M., Newton-Cheh, C., & Orho-Melander, M. (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events. *New Engl J Med* **358**, 1240–1249.

Larntz, K. (1978) Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J Am Stat Assoc* **73**, 253–263.

Morris, A.P. & Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* **34**, 188–193.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., & Daly, M.J. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet* **7**, e1001322.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* **50**, 157–175.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575.

R Development Core Team (2011) *A language and environment for statistical computing*. Vienna, Austria: Foundation for Statistical Computing. http://www.R-project.org.

Rafnar, T., Gudbjartsson, D.F., Sulem, P., Jonasdottir, A., Sigurdsson, A., Besenbacher, S., Lundin, P., Stacey, S.N., Gudmundsson, J., Magnusson, O.T., Le Roux, L., Orlygsdottir, G., Helgadottir, H.T., Johannsdottir, H., Gylfason, A., Tryggvadottir, L., Jonasson, J.G., De Juan, A., Ortega, E., Ramon-Cajal, J.M., Garcia-Prats, M.D., Mayordomo, C., Panadero, A., Rivera, F., Aben, K.K., Van Altena, A.M., Massuger, L.F., Aavikko, M., Kujala, P.M., Staff, S., Aaltonen, L.A., Olafsdottir, K., Bjornsson, J., Kong, A., Salvarsdottir, A., Saemundsson, H., Olafsson, K., Benediktsdottir, K.R., Gulcher, J., Masson, G., Kiemeney, L.A., Mayordomo, J.I., Thorsteinsdottir, U., & Stefansson, K. (2011) Mutations in *BRIP1* confer high risk of ovarian cancer. *Nat Genet* **43**, 1104–1107.

Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L.A., Boerwinkle, E., Cohen, J.C., & Hobbs, H.H. (2008) Genetic variation in *PNPLA3* confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40**, 1461–1465.

Sasieni, P.D. (1997) From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Sulem, P., Gudbjartsson, D.F., Walters, G.B., Helgadottir, H.T., Helgason, A., Gudjonsson, S.A., Zanon, C., Besenbacher, S., Bjornsdottir, G., Magnusson, O.T., Magnusson, G., Hjartarson, E., Saemundsdottir, J., Gylfason, A., Jonasdottir, A., Holm, H., Karason, A., Rafnar, T., Stefansson, H., Andreassen, O.A., Pedersen, J.H., Pack, A.I., De Visser, M.C., Kiemeney, L.A., Geirsson, A.J., Eyjolfsson, G.I., Olafsson, I., Kong, A., Masson, G., Jonsson, H., Thorsteinsdottir, U., Jonsdottir, I., & Stefansson, K. (2011) Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet* **43**, 1127–1130.

Victor, R.G., Haley, R.W., Willett, D.L., Peshock, R.M., Vaeth, P.C., Leonard, D., Basit, M., Cooper, R.S., Iannacchione, V.G., Visscher, W.A., Staab, J.M., & Hobbs, H.H. (2004) The Dallas Heart Study: A population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol*, **93**, 1473–1480.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82–93.

Xing, G., & Xing, C.. (2010) Adjusting for covariates in logistic regression models. *Genet Epidemiol* **34**, 769–771.

Received: 5 July 2011
Accepted: 16 December 2011