Computing Power and Sample Size for Case-Control Association Studies with Copy Number Polymorphism: Application of Mixture-Based Likelihood Ratio Test

Wonkuk Kim[1], Derek Gordon[2]*, Jonathan Sebat[3], Kenny Q. Ye[4], Stephen J. Finch[5]

1 Department of Mathematics and Statistics, University of South Florida, Tampa, Florida, United States of America, **2** Department of Genetics, Rutgers University, Piscataway, New Jersey, United States of America, **3** Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **4** Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York, United States of America, **5** Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, United States of America

Abstract

Recent studies suggest that copy number polymorphisms (CNPs) may play an important role in disease susceptibility and onset. Currently, the detection of CNPs mainly depends on microarray technology. For case-control studies, conventionally, subjects are assigned to a specific CNP category based on the continuous quantitative measure produced by microarray experiments, and cases and controls are then compared using a chi-square test of independence. The purpose of this work is to specify the likelihood ratio test statistic (*LRTS*) for case-control sampling design based on the underlying continuous quantitative measurement, and to assess its power and relative efficiency (as compared to the chi-square test of independence on CNP counts). The sample size and power formulas of both methods are given. For the latter, the CNPs are classified using the Bayesian classification rule. The *LRTS* is more powerful than this chi-square test for the alternatives considered, especially alternatives in which the at-risk CNP categories have low frequencies. An example of the application of the *LRTS* is given for a comparison of CNP distributions in individuals of Caucasian or Taiwanese ethnicity, where the *LRTS* appears to be more powerful than the chi-square test, possibly due to misclassification of the most common CNP category into a less common category.

Citation: Kim W, Gordon D, Sebat J, Ye KQ, Finch SJ (2008) Computing Power and Sample Size for Case-Control Association Studies with Copy Number Polymorphism: Application of Mixture-Based Likelihood Ratio Test. PLoS ONE 3(10): e3475. doi:10.1371/journal.pone.0003475

Editor: Peter Heutink, Vrije Universiteit Medical Centre, Netherlands

Received April 1, 2008; **Accepted** September 15, 2008; **Published** October 22, 2008

Funding: National Institutes of Health (grant P41 HG004222-01), The Simons Foundation

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gordon@biology.rutgers.edu

Introduction

Large-scale copy number polymorphisms (CNPs) are a recently discovered feature of human genomic architecture [1]. As reported by Sebat et al. [1], large-scale copy number polymorphisms (CNPs) (about 100 kilobases and greater) contribute substantially to genomic variation among normal humans. These authors documented CNPs of 70 different genes within CNP intervals, including genes involved in neurological function, regulation of cell growth, regulation of metabolism, and several genes known to be associated with disease. For example, investigators have documented that copy number variation of the region encompassing the *CCL3L1* gene [MIM 601395] is associated with HIV/AIDS susceptibility [2] [MIM 609423]. Other investigators have documented that copy number variation of the orthologous rat and human *FCGR3* genes [MIM 146740] is a determinant of susceptibility to immunologically mediated glomerulonephritis [3,4] [MIM 610665]. Additional recent publications suggest that CNPs may play a role in cardiovascular disease [5], lipoprotein and metabolic phenotypes [6], nervous system disorders [7], age-related macular degeneration [8,9] [MIM 610149], autism [10], cancer [1,11], and schizophrenia [12]. More generally, CNPs may play an important role in disease etiology for common, complex traits. Additionally, CNPs, like SNPs and microsatellite markers, may have different distributions for populations with different ethnicities [13,14].

Case-control genetic association designs can be a powerful way to map disease susceptibilty genes, particularly for diseases with smaller effect sizes [15,16,17,18,19]. In such designs, unrelated cases (with the phenotype of interest) and controls (who do not have the phenotype) are genotyped usually for thousands to hundreds of thousands of single nucleotide polymophisms (SNPs) across the human genome. Standard statistical analyses include the chi-square test of independence or the linear trend test [20,21] applied to the individual SNP genotype counts from cases and controls. These genotypes are usually determined through use of clustering algorithms applied to underlying quantitative measurements (e.g., see [22]).

Compared to SNP genotyping technologies, procedures for calling CNPs are less developed and less accurate [23]. Earlier CNP studies focused on discovery [24], with copy number changes being called using data from a single array, comparing DNA from an individual with a reference DNA sample. Recently developed methods classify known CNPs using array data collected from a large group of individuals. One method classifies known CNPs from the distribution of a univariate quantitative measure (C. Yoon; manuscript in preparation). Such a quantitative measure is either an average log fluorescent intensity ratio (between sample and reference DNA) over multiple probes representing the CNP, or the log-intensity ratio of the best probe within the CNP region.

One reason for the relative difficulty of CNP classification is that such classification is determined by relative intensity of a signal at a probe (or probe sets). In contrast, the two alleles of a SNP have two distinct nucleotides that can be represented by two distinct probes. Moreover, for multi-allelic CNPs, only the total number of copies (or categories) and not the alleles are observed for each individual. As an example, for a CNP locus of three alleles, with 1 copy, 2 copies and 3 copies respectively and probe intensity proportional to the number of copies, an intensity observation of 4 can be a genotype of 2/2 copies or a genotype of 1/3 copies.

Consider the pictorial examples in Figures 1a and 1b, which are created to represent a hypothetical CNP with a total of four different copy number categories (labeled "1" through "4"). For each category, different subjects will have a quantitative measure following a fixed continuous distribution, whether or not the subject is a case or control. The case category frequencies are then the mixing proportions of the component distributions [25]; similarly for the controls. In Figure 1a, the quantitative measures for CNP category i, $1\leq i\leq 4$ comes from a univariate normal distribution with mean i and variance $^{1}/_{36}$ each. Studying the figure, we see there is clear separation among the component normal distributions, so that classification of individuals into categories 1,…, 4 is highly accurate [26]. An example where classification is more problematic is presented in Figure 1b. In this figure, the CNP quantitative measures for subjects have normal distributions with the same means as in Figure 1a, but with variance ¼ each. That is, for each univariate distribution in Figure 1b, the variance is nine times that of the variance in Figure 1a, resulting in greater overlap among component distributions. As suggested by Figure 1b, when the component distributions have more overlap, the rate of misclassifying an individual having true CNP category i as having CNP category $j\neq i$ is much higher. It has been reported that the chi-square test of independence loses power as the misclassification rate increases [26,27,28].

An additional concern is that the CNP category that increases risk may occur with low frequency, as is often the situation with Mendelian diseases [29]. In case-control association studies using SNPs with low at-risk allele frequency, an increase in the genotype misclassification error rates requires indefinitely large increases in sample size to maintain constant power [30,31]. We hypothesize that CNP classification errors may lead to underpowered studies when the at-risk CNP category has low frequency. Challenges of performing association studies using CNP data were recently documented by McCarroll and Altshuler [32], who note, "To the extent that the precise allelic state of any DNA is not well measured, power declines." We raise the question: *Is there a more (statistically) powerful method of using CNP data when testing for association with a complex trait than the usual chi-square test of independence?*

To answer this question, we propose use of the likelihood ratio test statistic (*LRTS*) comparing the mixing proportions of cases and controls estimated from the underlying quantitative measures for CNPs. Rather than assign classifications to each individual's CNP, we perform a test of association on the CNP quantitative measure. We present an analytic solution to computation of power and sample size calculations for genetic association with CNP quantitative measures. We then calculate the efficiency of the chi-square test of independence using Bayesian classification compared to the *LRTS* to examine which test statistic has greater power for a wide range of trait specifications. By efficiency, we mean the ratio of sample size requirements for the chi-square test of independence and *LRTS*, respectively, for a fixed power and type I error rate. Finally, we demonstrate the use of the *LRTS* for differences in the mixing proportions of the CNP categories between two ethnic groups for a CNP with relevance to a genetic disease.

Methods

Notation

The following notation is used throughout this work:

X_α = A continuous random variable representing the CNP quantitative measure; α is an index indicating control ($\alpha=1$) or case ($\alpha=2$) status.

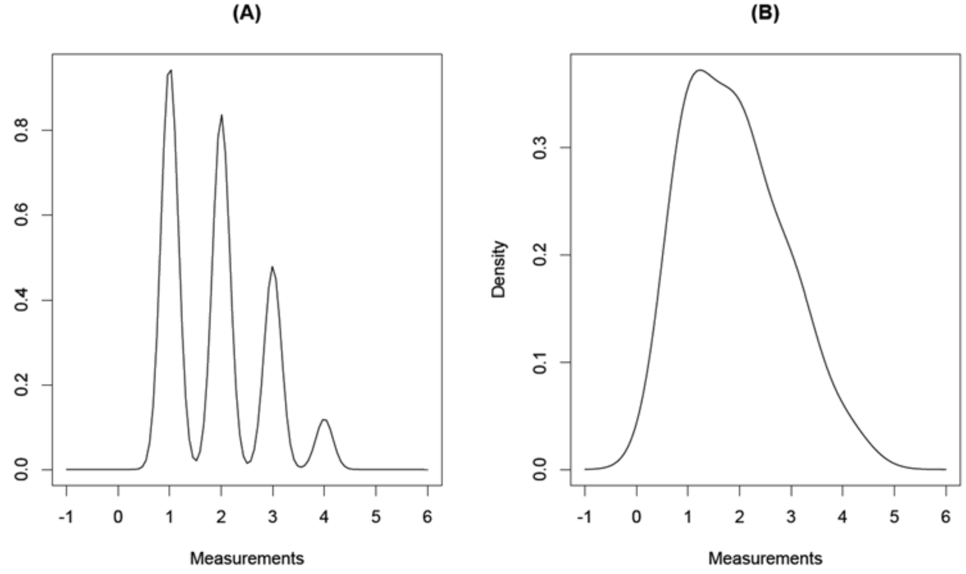

Figure 1. 1a and 1b. In this figures, we present probability density plots for statistical distributions that are mixtures of four univariate normal distributions with equally spaced means 1, 2, 3, and 4, and a common variance. In Figure 1a, the variance of each component distribution is $^{1}/_{36}$. In Figure 1b, the variance of each component distribution is ¼.
doi:10.1371/journal.pone.0003475.g001

The number of controls is n_1, and cases n_2, with $N = n_1 + n_2$ and $Q_\alpha = \frac{n_\alpha}{n_1 + n_2}$, $\alpha = 1,2$, which is the proportion of controls or cases in the total sample.

$d =$ The number of CNP categories; the subscript i indexes the category, $1 \le i \le d$.

$f(x \mid \theta_i, \eta) =$ The probability density function (pdf) of the continuous random variable $X = x$, conditional on the CNP category. This pdf is a function of the parameters θ_i and η, where η is a parameter that is constant for all component distributions. For example, if f is a normal pdf, then θ_i is the mean and η is the variance.

$\overrightarrow{p}_\alpha = (p_{\alpha 1}, \cdots, p_{\alpha d}) =$ A vector of mixing proportions; here, the values $p_{\alpha i}$, $1 \le i \le d$, are the proportions of CNP category i in the α affection status class ($\alpha = 1$ for controls, $\alpha = 2$ for cases). Under the null hypothesis, $p_{1i} = p_{2i}$.

$p_{0i} = Q_1 p_{1i} + Q_2 p_{2i}$. Since, under the null hypothesis, $p_{1i} = p_{2i}$, then $p_{1i} = p_{2i} = p_{0i}$, under the null.

$q_i =$ The CNP category frequencies in the population from which cases and controls are drawn.

$c_{\alpha i} = \sqrt{N}(p_{\alpha i} - p_{0i}) =$ A parameter needed for specification of the alternative hypothesis and hence power and sample size calculations.

$\overrightarrow{\theta} = (\theta_1, \cdots, \theta_d) =$ A vector of parameters for the probability density functions $f(x \mid \theta_i, \eta)$. In the examples used here, θ_i is the mean of the CNP category i distribution. Also, in the efficiency calculations reported later, $\theta_{i+1} - \theta_i = 1$, $i = 1, \ldots, d-1$. The separation $S = \frac{\theta_{i+1} - \theta_i}{\sqrt{\eta}}$ is the number of standard deviations between adjacent CNP category means.

Probability density function of the CNP quantitative measure

The probability density function of the random variable X_α is given by $h_\alpha\left(x \middle| \overrightarrow{p}_\alpha, \overrightarrow{\theta}, \eta\right) = \sum_{i=1}^{d} p_{\alpha i} f(x \mid \theta_i, \eta)$, where we assume the number of categories d is known and equal in both cases and controls. Given a CNP category i, the underlying pdf $f(\cdot)$ is the same for cases and controls. When $f(x \mid \theta_i, \eta)$ is a normal distribution, we specify that the variance (η) is equal across all CNP categories i and affection statuses α. While these specifications are not critical for performing power and sample size calculations, they may be advantageous when performing mixture analyses of real data. For instance, there may be convergence problems for the computed maximum likelihood of a univariate normal mixture if one allows the category variances to be unequal. Methods such as those proposed by Hathaway [33,34] may be used when the equal variance assumption does not hold.

Likelihood function

The likelihood function under the null hypothesis is given by:

$$L_0 = \left(\prod_{j=1}^{n_1+n_2} \left(\sum_{i=1}^{d} p_{0i} f\left(x_j \middle| \theta_i, \eta\right) \right) \right). \tag{1}$$

The likelihood function under the alternative hypothesis is given by:

$$L_1 = \left(\left(\prod_{j=1}^{n_1} \left(\sum_{i=1}^{d} p_{1i} f\left(x_j \middle| \theta_i, \eta\right) \right) \right) \left(\prod_{k=1}^{n_2} \left(\sum_{i=1}^{d} p_{2i} f\left(x_k \middle| \theta_i, \eta\right) \right) \right) \right). \tag{2}$$

Computationally, L_0 and L_1 are calculated by using the maximum likelihood estimates (MLEs) of the parameters.

LRTS

In this work, we consider two test statistics: (1) the *LRTS* applied directly to the CNP quantitative measures for cases and controls; and (2) the chi-square test of independence applied to $2 \times d$ tables after the CNP quantitative measures have been classified into one of d categories for cases and controls using a Bayesian classification rule (see section immediately following). The *LRTS* (1) is defined as

$$LRTS = 2\left(\max_{p_{1i}, p_{2i}, \theta_i, \eta} \ln(L_1) - \max_{p_{0i}, \theta_i, \eta} \ln(L_0) \right), \tag{3}$$

where the likelihoods are defined in equations (1) and (2).

Bayesian classification rule for univariate CNP quantitative measures

To categorize CNP quantitative measures into a CNP category, we consider a classification formula based on Bayes rule [35]. Since this approach minimizes the expected cost of misclassification, as proven in Anderson [36], it is a well-accepted approach. An observation x is assigned to CNP category i if and only if $p_{0i} f(x \mid \theta_i, \eta) \ge \max_{1 \le j \le d} p_{0j} f\left(x \middle| \theta_j, \eta\right)$, where $p_{0i} = Q_1 p_{1i} + Q_2 p_{2i}$, (defined above – Notation). For an example with $d = 3$ copy number categories and a normal CNP category distribution, application of the Bayes rule yields:

x is placed in the left-most component if $x < \min(\gamma_{12}, \gamma_{13})$,

x is placed in the middle component if $\gamma_{12} < x < \gamma_{23}$,

x is placed in the right-most component if $\max(\gamma_{13}, \gamma_{23}) < x$,

where $\gamma_{ij} = \frac{\theta_j + \theta_i}{2} - \frac{\eta}{\theta_j - \theta_i} \log\left(\frac{p_{0j}}{p_{0i}}\right)$, $1 \le i < j \le 3$. In applications, $(\gamma_{12}, \gamma_{13}, \gamma_{23})$ are estimated using the MLEs of the parameters.

Simulation studies to verify asymptotic null distribution of chi-square test with Bayesian classification

We perform simulation studies to verify the accuracy of the asymptotic null distribution of the chi-square test of independence applied to CNP counts after classification using the Bayesian classification rule (described above). We consider two settings each of sample size and mixing proportion vectors (a total of four settings). Our mixture model is a mixture of four univariate normal distributions with consecutive mean distances $\theta_{i+1} - \theta_i = 1$ unit apart. Separations are fixed to be $\frac{\theta_{i+1} - \theta_i}{\sqrt{\eta}} = 3$. We specify sample sizes $n_1 (= n_2) = 200$ *or* 500, and mixing proportions $\overrightarrow{p}_1 \left(= \overrightarrow{p}_2\right) = (0.25, \ 0.25, \ 0.25, \ 0.25)$ *or* $(0.1, \ 0.2, \ 0.3, \ 0.4)$.

Computing asymptotic power for the LRTS of the CNP quantitative measure

The asymptotic distribution of the *LRTS* under the null hypothesis follows a χ^2_{d-1} distribution under certain conditions [37] (referred to as "classic regularity conditions"); and the asymptotic power under the alternative specified hypothesis $H_N : p_{\alpha i} = p_{0i} + \frac{c_{\alpha i}}{\sqrt{N}}$ can be calculated using the non-central chi-square distribution $\chi^2_{d-1}(\lambda_{LRTS})$ with the non-centrality parameter (NCP) λ_{LRTS} given in Appendix S1.

Computing asymptotic power for chi-square test of independence

The $2 \times d$ test under an alternative hypothesis H_N asymptotically follows a non-central chi-square distribution [38]. When the component distributions have more overlap, the misclassification rates are much higher. If the misclassification error mechanism is random and non-differential, the observed classification probabilities p^* can be written in terms of a matrix of classification probabilities $\varepsilon = (\varepsilon_{ij})$, where $\varepsilon_{ij} = \Pr$ (subject's observed genotype $= i|$ subject's true genotype $= j$). The power for the chi-square test of independence with misclassification errors can be calculated from the NCP λ_{CS} [27,28,38], where

$$\lambda_{CS} = NQ_1Q_2\sum_{i=1}^{d}\frac{\left(p^*_{1i}-p^*_{2i}\right)^2}{p^*_{0i}}, \text{ and } p^*_{\alpha i} = \sum_{j=1}^{d}\varepsilon_{ij}p_{\alpha j}.$$

Genetic model parameters for efficiency analysis

We calculate the efficiency of the chi-square test on $2\times d$ contingency tables with respect to the *LRTS* on CNP quantitative measures for two genetic models of inheritance (MOI) associated with CNPs that have been documented as a possible MOI for CNPs [2,39]. We first specify the disease prevalence ϕ, the population frequencies q_i for CNP category $i, 1\le i \le d$, and the relative risks R_i of becoming affected, given that an individual has CNP category i. We then compute the penetrances $g_i = \Pr(affected | CNP\_category = i)$, where we specify $R_i = g_i/g_1$, so that the reference *CNP category* relative risk is 1. The reference CNP category may be chosen arbitrarily without loss of generality. The penetrances are given by $g_1 = \frac{\phi}{\sum_{i=1}^{d} R_i q_i}$, and $g_i = R_i g_1$. Using Bayes Theorem, the CNP category mixing proportions conditional on affection status are:

$$\begin{aligned} p_{1i} &= \Pr(CNP\_category = i|unaffected) \\ &= \frac{\Pr(unaffected, CNP\_category = i)}{\Pr(unaffected)} \qquad (4) \\ &= \frac{(1-g_i)q_i}{1-\phi}, \end{aligned}$$

$$\begin{aligned} p_{2i} &= \Pr(CNP\_category = i|affected) \\ &= \frac{\Pr(affected, CNP\_category = i)}{\Pr(affected)} \\ &= \frac{g_i q_i}{\phi}. \end{aligned}$$

For our comparative analyses, we set $d = 4$, $q_1 = 0.4$, $q_2 = 0.35$, $q3 = 0.2$, $q_4 = 0.05$, and $\phi = 0.05$. In the first (Dosage) model, the risk of becoming affected increases geometrically with increase in CNP category. We specify $R_2 = 1.8$, $R_3 = 1.8^2 = 3.24$, and $R_4 = 1.8^3 = 5.83$, so that risk increases by a factor of 1.8 for each increase in CNP category.

In the second (Extremes) model, risk of becoming affected increases for CNP categories 1 and d and decreases for all other categories. For this work, we specify $R_2 = 0.3$, $R_3 = 0.3$, and $R_1 = R_4 = 1$. Finally, we set the means to be equally spaced for all components. Specifically, $\theta_i = i$ for comparative analyses so that separation is given by $\frac{\theta_{i+1}-\theta_i}{\sqrt{\eta}} = \frac{1}{\sqrt{\eta}}$.

Simulation studies to verify asymptotic null and alternative distributions of LRTS

We perform simulation studies to verify the accuracy of the asymptotic null and alternative distributions of the *LRTS*. For the null distribution simulations, we consider two settings each of sample size and mixing proportion vectors (a total of four settings). For the alternative distribution simulations, we consider one setting of sample size and two different MOIs (a total of two settings). Also, for both sets of simulations, our mixture model is a mixture of four univariate normal distributions with consecutive mean distances $\theta_{i+1} - \theta_i = 1$ unit apart. Separations are fixed to be $\frac{\theta_{i+1}-\theta_i}{\sqrt{\eta}} = 3$.

For the null distribution simulations, we specify sample sizes $n_1(=n_2) = 200$ *or* 500, and mixing proportions $\overrightarrow{p}_1\left(=\overrightarrow{p}_2\right) = (0.25,\ 0.25,\ 0.25,\ 0.25)$ *or* $(0.1,\ 0.2,\ 0.3,\ 0.4)$. For the alternative distribution simulations, sample sizes are n_1 $(=n_2) = 200$, and mixing proportions are determined using equations (4) with the specified parameters (including CNP population frequencies) for the Dosage and Extremes MOIs, given above (Methods - Genetic model parameters for efficiency analysis).

To find the global maximum (equations (1) and (2)), we use Expectation-Maximization algorithms (EM). A small pilot study found that there were typically three relative maxima under the null specification and two under the alternative. Consequently, we use 100 random starting points (RSPs) for parameter estimation under the null distribution simulations and 50 RSPs for the estimation under the alternative distribution simulations. EM algorithm computations are performed using MCLUST in the R programming environment [40]. For each RSP, the convergence tolerance is set at 10^{-5} and the maximum iteration number is set at 300.

Efficiency of the chi-square test relative to the LRTS

The efficiency of the $2\times d$ test relative to the *LRTS* is denoted *Eff* and is the ratio $Eff = \frac{\lambda_{CS}}{\lambda_{LRTS}}$ of the NCP of the chi-square test to the NCP of the *LRTS*. When the relative efficiency is less than 1, the chi-square test requires a larger sample size to achieve the same power as the *LRTS*, given that both tests have the same level of significance. For example, if the relative efficiency of the $2\times d$ test is 0.8, the $2\times d$ test requires 100 observations to have the same power as the *LRTS* using 80 observations.

Example CNP data for two ancestral populations

Since recent work documents different CNP distributions in different ethnic populations [41,42] we apply our *LRTS* to test for differences in mixing proportions of CNP categories between two groups of individuals (Caucasian and Taiwanese) using probe ratio data for a multi-allelic CNP probe in the *FCGR3* gene on Chromosome 1. We also apply the chi-square test of independence to the probe ratio data after the individuals are classified into categories using the Bayesian classification rule described above. Oligonucleotide probes are designed as described previously [43].

To be consistent with notation used throughout this work, from this point forward we label the Taiwanese samples as "controls" and the Caucasian samples as "cases", although individuals in this study were not ascertained for any particular disease phenotype.

Results

Simulation studies to verify asymptotic null distribution of chi-square test with Bayesian classification

In Table 1, we report the *empirical type I error rates* at the 0.975, 0.10, 0.05, 0.025, and 0.01 significance levels for each set of parameter settings. For each simulation, these type I error rates are the proportion of replicates for which the computed *LRTS* exceeds 0.2157, 6.25, 7.81, 9.348 or 11.34, which correspond to the 0.975, 0.10, 0.05, 0.025 and 0.01 significance level cutoffs for a central chi-square distribution with 3 degrees of freedom (the asymptotic null distribution for each simulation). For each empirical type I error rate, we report the 95% confidence interval,

Table 1. Simulation results of the null distribution of chi-squared test.

Sample size	Proportions	*Empirical type I error rate**					KS-Test P-value
		0.975 Level	**0.10 Level**	**0.05 Level**	**0.025 Level**	**0.01 Level**	
200	(0.25, 0.25, 0.25, 0.25)	0.976	0.107	0.042	0.018	0.005	0.72
500	(0.25, 0.25, 0.25, 0.25)	0.971	0.092	0.047	0.025	0.007	0.78
200	(0.1, 0.2, 0.3, 0.4)	0.979	0.094	0.046	0.018	0.006	0.54
500	(0.1, 0.2, 0.3, 0.4)	0.983	0.106	0.056	0.036	0.010	0.36

Based on 1000 replications for each settings.

doi:10.1371/journal.pone.0003475.t001

based on 1000 replicates. As an additional confirmation, we apply the Kolmogorov-Smirnoff (KS) goodness of fit test [44,45] to each simulations' set of 1000 *LRTS* values (i.e., sample size for KS test is 1000), and report the p-values in Table 1.

In each simulation, the target type I error rate is contained in the 95% confidence interval for the corresponding empirical type I error rate. In addition, the smallest KS test p-value is 0.36, indicating that we do not reject the null hypothesis that the data are drawn from a central chi-square distribution with 3 degrees of freedom.

Computing asymptotic power for LRTS of copy number measurement

When the alternative hypothesis $H_N : p_{\alpha i} = p_{0i} + \frac{c_{\alpha i}}{\sqrt{N}}$ is true, the NCP of the *LRTS* may be written in a quadratic form as:

$$\lambda_{LRTS} = NQ_1Q_2\left(p_{11} - p_{21}, \quad \cdots, \quad p_{1(d-1)} - p_{2(d-1)}\right) J_0 \begin{pmatrix} p_{11} - p_{21} \\ \vdots \\ p_{1(d-1)} - p_{2(d-1)} \end{pmatrix}$$

where $\mathcal{J}_0$ is the $(d-1) \times (d-1)$ symmetric matrix specified in Appendix S1.

Simulation studies to verify asymptotic null and alternative distributions of LRTS

As in Table 1, in Table 2, we report the empirical type I error rates at the 0.10, 0.05, and 0.01 significance levels for each set of parameter settings. For each simulation, these type I error rates are the proportion of replicates for which the computed *LRTS* exceeds 6.25, 7.81, or 11.34, which correspond to the 0.10, 0.05, and 0.01 significance level cutoffs for a central chi-square distribution with 3 degrees of freedom (the asymptotic null distribution for each simulation). For each empirical type I error rate, we report the 95% confidence interval, based on 1000 replicates. As an additional confirmation, we apply the Kolmogorov-Smirnoff (KS) goodness of fit test [44,45] to each simulations' set of 1000 *LRTS* values (i.e., sample size for KS test is 1000), and report the p-values in Table 2.

In each simulation, the target type I error rate is contained in the 95% confidence interval for the corresponding empirical type I error rate. In addition, the smallest KS test p-value is 0.34, indicating that we do not reject the null hypothesis that the data are drawn from a central chi-square distribution with 3 degrees of freedom.

In Table 3, we report the *simulation power* at the 10^{-3}, 10^{-4}, and 10^{-5} significance levels for each set of parameter settings. For each simulation, these powers are the proportion of replicates for which the computed *LRTS* exceeds 16.27, 21.11, or 25.90, which correspond to the 10^{-3}, 10^{-4}, and 10^{-5} cutoffs for a central chi-square distribution with 3 degrees of freedom (the asymptotic null distribution for each simulation). More stringent significance level cutoffs are chosen for the power analyses since power at the 0.10, 0.05, and 0.01 levels is close to or equal to 100% for these parameter specifications. As with the empirical type I error rates in Table 1, we report the 95% confidence intervals, based on 1000 replicates each. We also report the asymptotic power at each of the significance levels, determined by computing the non-centrality parameter (equation (A1)) for each set of parameter settings. As an additional confirmation, we apply the Kolmogorov-Smirnoff (KS) goodness of fit test [44,45] to each simulations' set of 200 *LRTS* values (i.e., sample size for KS test is 200), and report the p-values in Table 3.

While the KS p-values are much smaller, we see that, for the 10^{-3} and 10^{-4} significance levels, the simulation power is contained in the 95% confidence interval for each simulation. The results of this table suggest that our simulation results are consistent with asymptotic results for at least the 10^{-3} and 10^{-4} significance levels.

Table 2. Simulation results of the null distribution of *LRTS*.

Sample size	Proportions	*Empirical type I error rate**					KS-Test P-value
		0.975 Level	**0.10 Level**	**0.05 Level**	**0.025 Level**	**0.01 Level**	
200	(0.25, 0.25, 0.25, 0.25)	0.979	0.103	0.045	0.015	0.007	0.81
500	(0.25, 0.25, 0.25, 0.25)	0.971	0.097	0.052	0.021	0.013	0.79
200	(0.1, 0.2, 0.3, 0.4)	0.977	0.106	0.046	0.020	0.005	0.34
500	(0.1, 0.2, 0.3, 0.4)	0.982	0.109	0.060	0.028	0.011	0.41

Based on 1000 replications for each settings.

doi:10.1371/journal.pone.0003475.t002

Table 3. Simulation results for *LRTS* under alternative distributions.

MOI	Method to calculate power	*Simulation Power**			KS-Test P-value
		10^{-3} Level	10^{-4} Level	10^{-5} Level	
Dosage	Simulation	0.958 (0.946, 0.970)	0.866 (0.845, 0887)	0.735 (0.708, 0.762)	0.01
	Asymptotic	0.949	0.856	0.712	
Extremes	Simulation	0.950 (0.936, 0.964)	0.857 (0.835, 0.879)	0.738 (0.711, 0.765)	0.07
	Asymptotic	0.946	0.848	0.700	

Legend for Table 2. Based on 1000 replications and 200 sample size per case/control group.

*95% approximate confidence intervals for simulated power are given in parentheses.

Here, we present simulated and asymptotic power for the *LRTS* when the alternative hypothesis that mixing proportions are different in each of two groups is true. The mixing proportions are computed using equations (4) for the Dosage and Extremes models, where CNP population frequencies are as specified above (Methods - Genetic model parameters for efficiency analysis). For the Dosage model, the relative risks are: $R_2 = 1.8$, $R_3 = 1.8^2 = 3.64$, $R_4 = 1.8^3 = 5.83$. For the Extremes model, the relative risks are: $R_1 = 1$, $R_2 = 0.3$, $R_3 = 0.3$, $R_4 = 1$. Asymptotic power is computed using the non-centrality parameter documented in equation (A1). The column "KS-Test P-value" refers to the p-value computed using the Kolmogoroff-Smirnoff goodness of fit test, as implemented in R programming environment.

doi:10.1371/journal.pone.0003475.t003

Relative efficiency of the 2×*d* chi-square test relative to the LRTS

Using the result for the NCP of the *LRTS*,

$$Eff = \frac{\sum_{i=1}^{d} \frac{\left(p_{1i}^{*} - p_{2i}^{*}\right)^2}{Q_1 p_{1i}^{*} + Q_2 p_{2i}^{*}}}{\sum_{i=1}^{d-1} \sum_{j=1}^{d-1} J_{ij}(p_{1i} - p_{2i})\left(p_{1j} - p_{2j}\right)},$$

where

$$J_{ij} = E_0 \left(\frac{(f(x|\theta_i,\eta) - f(x|\theta_d,\eta))\left(f\left(x\middle|\theta_j,\eta\right) - f(x|\theta_d,\eta)\right)}{\left[\sum_{k=1}^{d} (Q_1 p_{1k} + Q_2 p_{2k}) f(x|\theta_k,\eta)\right]^2} \right).$$

Figure 2 contains the relative efficiency of the 2×4 chi-square test with Bayesian rule classification with respect to the *LRTS* for the Extremes and Dosage models against the separation between successive category means. In all models, the relative efficiency is less than 1; that is, the *LRTS* is more powerful. When the separation is 5 standard deviations or greater, both tests have essentially the same power. The relative efficiency steadily declines as the separation between category means decreases, with less efficiency for the Extremes model.

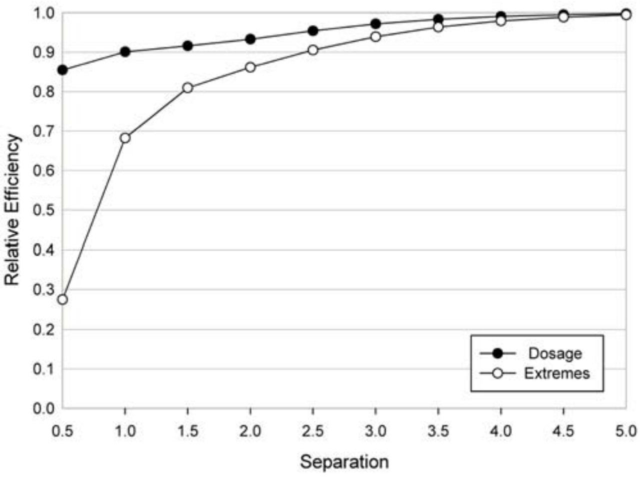

Figure 2. Here we present the relative efficiency *Eff* (defined in Methods) of the chi-square test of independence in relation to the *LRTS* as a function of separation ($\frac{1}{\sqrt{\eta}}$) between the four component distributions that comprise the mixture distribution. All information regarding parameter specification for the Dosage and Extremes models for which relative efficiencies are calculated is presented in the Methods section (Genetic model parameters for efficiency analysis).

doi:10.1371/journal.pone.0003475.g002

Example CNP data for two populations

Results for the *LRTS* applied to P4077 probe ratio data for the Caucasian and Taiwanese samples are presented in Table 4. Figure 3 contains the histograms of each group's probe ratio data, as well as of the combined groups (Caucasians and Taiwanese). There are an estimated three CNP categories, and the *LRTS* p-value for the P4077 probe is 0.014. In comparison, the chi-square test of independence p-value based on the asymptotic null distribution for the P4077 probe data with classification by the Bayesian rule is 0.03. The p-value based on Fisher's Exact Test is 0.0175. The numbers of Caucasian and Taiwanese individuals in CNP categories 1, 2, and 3 are: 229, 31, and 1; and 67, 20, and 1, respectively, as determined by the Bayesian classification rule. Additionally, we report the estimated classification rates as follows:

$$\varepsilon = \begin{pmatrix} 0.963 & 0.037 & 0.000 \\ 0.335 & 0.663 & 0.002 \\ 0.000 & 0.087 & 0.913 \end{pmatrix},$$

where $\varepsilon_{ij} = \Pr(\textit{reported CNP classification} = j \,|\, \textit{true CNP classification} = i)$. The *LRTS* method provides a slightly more significant p-value.

When we use the estimated misclassification parameters in the matrix ε along with the estimated mixing proportions under the alternative hypothesis (Table 4) in the *P*ower for *A*ssociation *W*ith *E*rror (PAWE) webtool, the power at the 5% significance level for the sample sizes specified in our example is 98% with error-free data, and is 76% with error rates given in ε, a power loss of 22%. From the perspective of power loss, Kang et al. [30,31] showed that misclassification of the most common category to any other category is the most costly; here, the estimated error rate of 3.7% in classification CNP category "1" as category "2" results in the greatest power loss. Other investigators have previously documented that the chi-square test of independence and the linear trend test lose power under such misclassification when data are genotypes or multi-locus haplotypes [30,31,46,47].

Additionally, if we compute the separation values $\frac{\theta_{i+1} - \theta_i}{\sqrt{\eta}}$, $i = 1$, 2, using the estimated parameters from Table 3, we see that separation between categories 1 and 2 is $\frac{1.42 - 1.056}{\sqrt{0.03}} = 2.09$, and separation between categories 2 and 3 is $\frac{2.18 - 1.42}{\sqrt{0.03}} = 4.38$. That is,

Table 4. Parameter estimation with 3 component normal mixtures for probe P4077 ratio data.

Hypothesis	Estimated parameters	CNP Category		
		$i=1$	$i=2$	$i=3$
Null (H_0)	Mixing proportions	0.815	0.179	0.006
	Means (θ_i)	1.062	1.446	2.191
Alternative (H_N)	Mixing proportions for Taiwanese (p_{1i})	0.626	0.362	0.011
	Mixing proportions for Caucasians (p_{2i})	0.843	0.152	0.005
	Means (θ_i)	1.056	1.420	2.180

Legend for Table 4. Data are determined for 261 individuals of Caucasian ethnicity and 88 individuals of Taiwanese ethnicity. The estimated variance (η) under both the null and alternative hypotheses is 0.03.

doi:10.1371/journal.pone.0003475.t004

for the majority of samples (categories 1 and 2) the separation is only 2.09. Our results of the relative efficiency studies in Figure 2 also suggest that, for such separation, the chi-square test with Bayesian classification is a less powerful procedure than the *LRTS*.

While one cannot use parameters estimated from data collected to calculate actual power, we present these calculations as indications of the source of the greater power of the *LRTS* due to the relatively high misclassification rates that are consistent with the estimated parameters.

Discussion

We have derived the non-centrality parameter for the *LRTS* of the mixture proportions applied to the CNP quantitative measurements. The relative efficiency of the 2×4 chi-square test is less than 1 for the example disease MOIs considered here, with greater decreases as the separation between category-means decreases. That is, for the models considered, the *LRTS* is more powerful than the chi-square test. In the example, power may have been lost for the chi-square test because of relatively high estimated misclassification rate from the most common category to the second most common category. The chi-square test of independence can lose substantial power under such misclassification [30,31,47].

A key advantage of the *LRTS* is that it can be computed on *any* CNP data, whether or not that data can be categorized. While the example presented (Table 4 and Figure 3) used only a single CNP

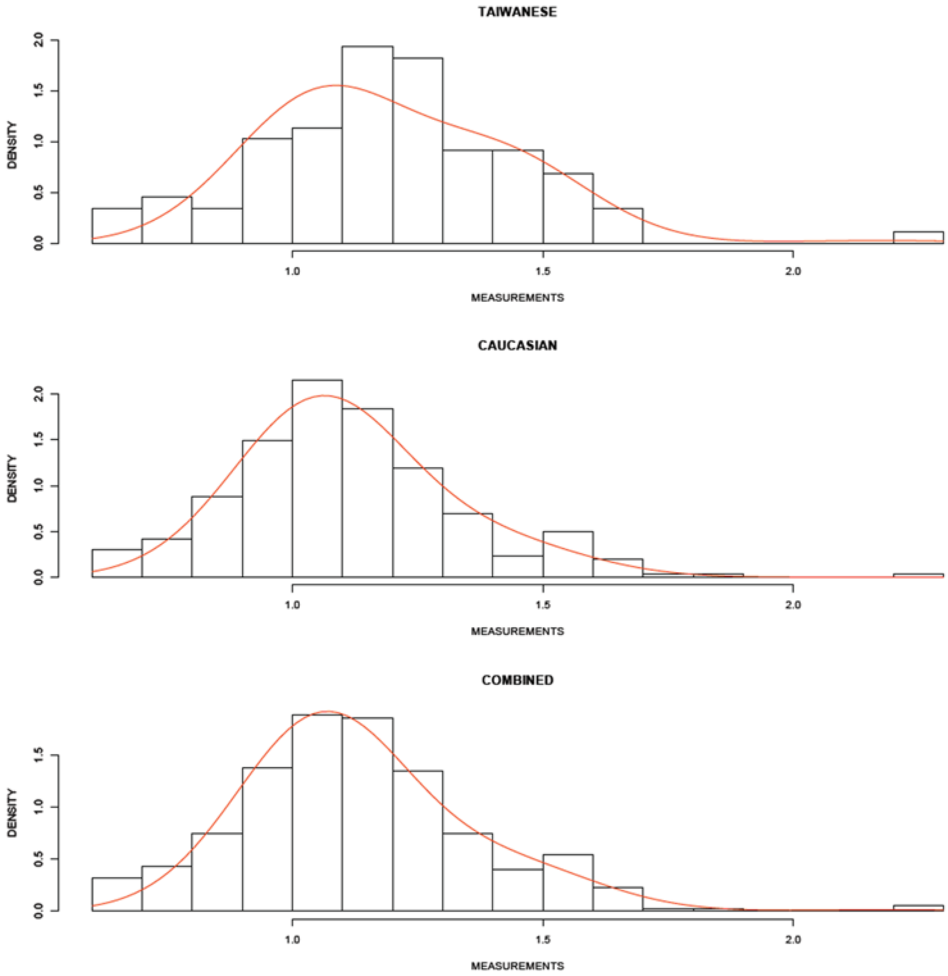

Figure 3. In these figures, we provide histograms of P4077 probe ratio data for Taiwanese, Caucasian and Combined (Taiwanese and Caucasian) samples. We also provide a fitted probability density function line for each data set. These graphs were created using the R programming environment. The horizontal axis labeled "MEASUREMENT" refers to each individual's probe ratio data value (after log transform) for the P4077 probe.

doi:10.1371/journal.pone.0003475.g003

as an illustration, the *LRTS* can be calculated for multiple SNPs analyzed simultaneously through specification of a multivariate pdf. The formal statistical analysis is the same, in that the *LRTS* is calculated as shown in Equation (3). Additionally, extensions of a multivariate procedure can incorporate more complex modeling of the mixture mechanism, for example, including a Hidden Markov Model approach.

The results indicated in Table 1 and Figure 2, namely that non-differential misclassification errors do not result in a change in the type I error rate and that there is power loss for the chi-square test of association, are consistent with numerous publications on the subject of non-differential genotyping error. Pompanon et al. [48] and Gordon and Finch [49,50] provide reviews of the literature.

As an alternative analysis, one might consider a logistic regression model with case/control status as the dependent variable and CNP quantitative measure as the independent variable. One potential advantage of this method is that determination of optimal estimates is less computationally intensive than the LRTS procedure documented in this work. Another potential advantage of logistic regression is that it allows for the possible inclusion of covariates. In this work we focus on the *LRTS* to avoid specification of a mathematical model of association. That is, the *LRTS* presented here only tests whether mixing proportions are different in two groups. There are mixture models that examine whether covariates are associated with CNP category membership [51,52]. A natural next step to extend our work is to allow the inclusion of covariates. The *LRTS* is similar in spirit to the commonly used chi-square test of independence for genotype data on cases and controls. That statistic similarly tests for differences in allele or genotype frequencies among different categories (e.g., cases and controls). We further note that there is literature on power and sample size for logistic regression [53,54]. While robustness of logistic regression procedures when the independent variable is drawn from a single univariate normal distribution is well documented (e.g., see [55]), the extension to logistic regression procedures when the independent variable is drawn from a mixture of distributions, as is the situation with CNPs, needs further investigation.

The recent work documenting differences in CNP distributions for different ethnic populations is consistent with the frequently replicated results that there are different allele and genotype frequency distributions in different ethnic populations [13,56]. Yu et al. [57] confirmed CNP values with "gold-standard" sequencing data. It is a limitation of our example that our estimated CNP classifications are not confirmed with sequencing data. Recent methodological research has documented several benefits of having standard and gold-standard measurements simultaneously on a subset of individuals [58,59,60]. Such sampling has been referred to as double-sampling [61,62].

An additional limitation in the data analysis of our example is our assumption of equal variances among the component distributions. While this assumption appeared to be true for this example, it will not hold in general. In that event, methods such as those proposed by Hathaway [33,34] may be used.

The power and sample size calculations presented here are based on asymptotic theory; that is, our results should hold when sample sizes are sufficiently large. When sample sizes are smaller, one can use simulation methods to estimate power. Of course, p-values should be based on permutation tests in such instances.

Web Resources

Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/Omim)

Power for Association With Error (http://linkage.rockefeller.edu/pawe/)

Supporting Information

Appendix S1

Found at: doi:10.1371/journal.pone.0003475.s001 (0.02 MB PDF)

Author Contributions

Conceived and designed the experiments: DG JS KY SJF. Performed the experiments: WK. Analyzed the data: WK DG SJF. Contributed reagents/materials/analysis tools: WK DG JS KY SJF. Wrote the paper: WK DG JS KY SJF.

References

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–528.
2. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307: 1434–1440.
3. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, et al. (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439: 851–855.
4. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, et al. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39: 721–723.
5. Pollex RL, Hegele RA (2007) Copy number variation in the human genome and its implications for cardiovascular disease. Circulation 115: 3130–3138.
6. Pollex RL, Hegele RA (2007) Genomic copy number variation and its potential role in lipoprotein and metabolic phenotypes. Curr Opin Lipidol 18: 174–180.
7. Lee JA, Lupski JR (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 52: 103–121.
8. Goverdhan SV, Hannan S, Newsom RB, Luff AJ, Griffiths H, et al. (2007) An analysis of the CFH Y402H genotype in AMD patients and controls from the UK, and response to PDT treatment. Eye.
9. Wegscheider BJ, Weger M, Renner W, Steinbrugger I, Marz W, et al. (2007) Association of complement factor H Y402H gene polymorphism with different subtypes of exudative age-related macular degeneration. Ophthalmology 114: 738–742.
10. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445–449.
11. Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, et al. (2007) Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. Carcinogenesis.
12. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539–543.
13. Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, et al. (2000) ALFRED: a Web-accessible allele frequency database. Pac Symp Biocomput. pp 639–650.
14. Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, et al. (2001) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms–an update. Nucleic Acids Res 29: 317–319.
15. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, et al. (2006) A common variant associated with prostate cancer in European and African populations. Nat Genet 38: 652–658.
16. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, et al. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39: 631–637.
17. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
18. Ozaki K, Tanaka T (2005) Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses. Cell Mol Life Sci 62: 1804–1813.
19. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881–885.
20. Cochran WG (1954) Some methods for strengthening the common chi-squared tests. Biometrics 10: 417–451.
21. Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386.
22. (2007) Illumina GenCall Data Analysis Software Download. http://www.illumina.com/downloads/GenCallTechSpotlight.pdf.

23. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, et al. (2006) Copy number variation: new insights in genome diversity. Genome Res 16: 949–961.
24. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, et al. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 13: 2291–2305.
25. Titterington D, Smith A, Makov U (1985) Statistical Analysis of Finite Mixture Distributions. New York: J. Wiley and Sons. pp 254.
26. Kang SJ, Gordon D, Brown AM, Ott J, Finch SJ (2004) Tradeoff between no-call reduction in genotyping error rate and loss of sample size for genetic case/control association studies. Pac Symp Biocomput. pp 116–127.
27. Gordon D, Finch SJ, Nothnagel M, Ott J (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 54: 22–33.
28. Mote VL, Anderson RL (1965) An investigation of the effect of misclassification on the properties of chisquare-tests in the analysis of categorical data. Biometrika 52: 95–109.
29. Ott J (1999) Analysis of Human Genetic Linkage. Baltimore: Johns Hopkins.
30. Kang SJ, Finch SJ, Haynes C, Gordon D (2004) Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. Hum Hered 58: 139–144.
31. Kang SJ, Gordon D, Finch SJ (2004) What SNP genotyping errors are most costly for genetic association studies? Genet Epidemiol 26: 132–141.
32. McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37–42.
33. Hathaway RJ (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Stat 13: 795–800.
34. Hathaway RJ (1986) A constrained EM-algorithm for univariate normal mixtures. J Stat Comp Simulation 23: 211–230.
35. McLachlan GJ, Peel D (2000) Finite mixture models. New York: J. Wiley and Sons. 456 p.
36. Anderson TW (2003) An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons. 675 p.
37. van der Vaart AW (1998) Asymptotic statistics; and CSiS, Mathematics P, ed. Cambridge: Cambridge University Press.
38. Mitra SK (1958) On the limiting power function of the frequency chi-square test. Ann Math Stat 29: 1221–1233.
39. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, et al. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80: 1037–1054.
40. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97: 611–631.
41. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
42. White SJ, Vissers LE, Geurts van Kessel A, de Menezes RX, Kalay E, et al. (2007) Variation of CNV distribution in five different ethnic populations. Cytogenet Genome Res 118: 19–30.
43. Healy J, Thomas EE, Schwartz JT, Wigler M (2003) Annotating large genomes with exact word matches. Genome Res 13: 2306–2315.
44. Smirnoff N (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin de l Universite de Moscou, Serie internationale (Mathematiques) 2: 3–14.
45. Kolmogoroff A (1941) Confidence limits for an unknown distribution function. Ann Math Stat 12: 461–463.
46. Ahn K, Haynes C, Kim W, Fleur RS, Gordon D, et al. (2007) The effects of SNP genotyping errors on the power of the cochran-armitage linear trend test for case/control association studies. Ann Hum Genet 71: 249–261.
47. Levenstien MA, Ott J, Gordon D (2006) Are Molecular Haplotypes Worth the Time and Expense? A Cost-Effective Method for Applying Molecular Haplotypes. PLoS Genet 2: e127.
48. Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6: 847–859.
49. Gordon D, Finch SJ (2005) Factors affecting statistical power in the detection of genetic association. J Clin Invest 115: 1408–1418.
50. Gordon D, Finch SJ (2006) Consequences of error. In: Dunn MJ, Jorde LB, Little PFR, Subramaniam S, eds. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Hoboken: J. Wiley and Sons.
51. Jones B, Nagin D, Roeder K (2001) A SAS procedure based on mixture models for estimating developmental trajectories. Sociol Method Res 29: 374–393.
52. Corbiere F, Joly P (2007) A SAS macro for parametric and semiparametric mixture cure models. Comput Methods Programs Biomed 85: 173–180.
53. Hsieh FY (1989) Sample size tables for logistic regression. Stat Med 8: 795–802.
54. Hsieh FY, Bloch DA, Larsen MD (1998) A simple method of sample size calculation for linear and logistic regression. Stat Med 17: 1623–1634.
55. Agresti A (2002) Categorical Data Analysis. Hoboken: John Wiley and Sons. 710 p.
56. Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, et al. (2000) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 28: 361–363.
57. Yu CE, Dawson G, Munson J, D'Souza I, Osterling J, et al. (2002) Presence of large deletions in kindreds with autism. Am J Hum Genet 71: 100–115.
58. Gordon D, Haynes C, Yang Y, Kramer PL, Finch SJ (2007) Linear trend tests for case-control genetic association that incorporate random phenotype and genotype misclassification error. Genet Epidemiol 31: 853–870.
59. Ji F, Yang Y, Haynes C, Finch SJ, Gordon D (2005) Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. Stat Appl Genet Mol Biol 4: Article 37.
60. Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, et al. (2004) Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Stat Appl Genet Mol Biol 3: Article 26.
61. Tenenbein A (1970) A double sampling scheme for estimating from binomial data with misclassifications. J Am Stat Assoc 65: 1350–1361.
62. Tenenbein A (1972) A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. Technometrics 14: 187–202.