Single-Variant and Multi-Variant Trend Tests for Genetic Association with Next-Generation Sequencing That Are Robust to Sequencing Error

Wonkuk Kim^a Douglas Londono^b Lisheng Zhou^b Jinchuan Xing^b
Alejandro Q. Nato^b Anthony Musolf^b Tara C. Matise^b Stephen J. Finch^c
Derek Gordon^b

^aDepartment of Mathematics and Statistics, University of South Florida, Tampa, Fla., ^bDepartment of Genetics and Human Genetics Institute, Rutgers, The State University of New Jersey, Piscataway, N.J., and ^cDepartment of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, N.Y., USA

Key Words

Next-generation sequencing · Rare variant · Trend test · Genetic association · GWAS · Allele · Locus

Abstract

As with any new technology, next-generation sequencing (NGS) has potential advantages and potential challenges. One advantage is the identification of multiple causal variants for disease that might otherwise be missed by SNP-chip technology. One potential challenge is misclassification error (as with any emerging technology) and the issue of power loss due to multiple testing. Here, we develop an extension of the linear trend test for association that incorporates differential misclassification error and may be applied to any number of SNPs. We call the statistic the linear trend test allowing for error, applied to NGS, or $LTT_{ae,NGS}$. This statistic allows for differential misclassification. The observed data are phenotypes for unrelated cases and controls, coverage, and the number of putative causal variants for every individual at all SNPs. We simulate data considering multiple factors (disease mode of inheritance, genotype relative risk, causal variant frequency, sequence error rate in cases, sequence error rate in controls, number of loci, and others) and evaluate

type I error rate and power for each vector of factor settings. We compare our results with two recently published NGS statistics. Also, we create a fictitious disease model based on downloaded 1000 Genomes data for 5 SNPs and 388 individuals, and apply our statistic to those data. We find that the $LTT_{ae,NGS}$ maintains the correct type I error rate in all simulations (differential and non-differential error), while the other statistics show large inflation in type I error for lower coverage. Power for all three methods is approximately the same for all three statistics in the presence of non-differential error. Application of our statistic to the 1000 Genomes data suggests that, for the data downloaded, there is a 1.5% sequence misclassification rate over all SNPs. Finally, application of the multi-variant form of $LTT_{ae,NGS}$ shows high power for a number of simulation settings, although it can have lower power than the corresponding single-variant simulation results, most probably due to our specification of multi-variant SNP correlation values. In conclusion, our $LTT_{ae,NGS}$ addresses two key challenges with NGS disease studies; first, it allows for differential misclassification when computing the statistic; and second, it addresses the multiple-testing issue in that there is a multi-variant form of the statistic that has only one degree of freedom, and provides a single p value, no matter how many loci.

Copyright © 2013 S. Karger AG, Basel

Introduction

By now, it is well established that causal variants (CVs) will play an important role in disease gene mapping for complex traits. By CV, we mean an allele at a polymorphic locus that increases risk of acquiring a disease. Over the course of the last several decades, CVs have unquestionably played an important role in Mendelian diseases, i.e. diseases for which the probability of acquiring the disease conditional on having at least one copy of CV is large (near 100%) [1].

One of the critical issues when applying statistical association methods with CVs is that of CV frequency (CVF). Mathematically, if a SNP locus has two variants, a (non-causal), and A (causal), the CVF equals $\Pr(A)$ in the population. As computed using methods implemented in the programs Genetic Power Calculator [2], CaTS [3], and others, statistical power to detect association with a single CV when its frequency is very small (<0.01) and the genotype relative risks are moderate (≤ 2.0) is very low, for any fixed significance level and any 'reasonable' sample (say 1,000 cases, 1,000 controls). Equivalently, to achieve a statistical power of 80% for any fixed significance level requires an extremely large number of cases and controls ($>10,000$ each). For example, using the PAWE-PH website [4], if the genotype relative risk of the heterozygote is 1.5, the disease mode of inheritance is multiplicative, the disease allele frequency is 0.01, and the SNP marker locus is the disease locus, then we require 13,926 cases and 13,926 controls to achieve 80% power at the $10E-4$ significance level, even without phenotype misclassification.

Furthermore, the issue of misclassification of the CV call becomes extremely important. Several researchers have investigated the effects of genotype misclassification on the power and robustness of statistical association methods [5–23]. As documented by Kang et al. [24, 25] and Ahn et al. [26, 27], and as can be illustrated using the method implemented in the PAWE software [28] among others, the limit of the 'cost' of genotype misclassification approaches infinity as the CVF approaches 0. Here, cost is defined as the ratio:

(Minimum sample size necessary to detect association in the presence of misclassification):(minimum sample size necessary to detect association when no misclassification is present).

In other words, when the CVF is small, even very small misclassification rates (say 1%) require prohibitively large sample sizes to detect association.

A number of statistical approaches have been proposed to deal with this situation [10, 11, 16, 29–36]. A brief sum-

mary of findings is that there are two types of misclassification errors with which to be concerned; non-differential misclassification, in which error penetrances at a locus are the same in the affected and unaffected samples; and differential misclassification, in which penetrances differ among affected and unaffected samples. The former affects only statistical power in case-control association studies, but affects both power and type I error rate in family-based association. The latter affects type I error rates and statistical power even in case-control association studies.

A recent publication has looked at non-differential misclassification in NGS studies [37]. These authors found that even at very low error rates, misclassifying a common homozygote as a heterozygote causes (sometimes substantial) loss of statistic power to detect association, and this power loss grows as the minor allele frequency decreases. These results are consistent with the work by Kang et al. [24, 25], who observed the same consequences for genotype misclassification.

From our perspective, differential misclassification error is more troublesome statistically, since one no longer knows the true null distribution and, therefore, true power cannot be determined.

One of the primary approaches for studying CVs is direct sequencing of the individuals of interest. In the past several years, the development of the next-generation sequencing (NGS) technology has made whole-exome (the protein-coding regions of a genome) or whole-genome sequencing an increasingly important component of medical genetic studies [38–49]. In addition to identifying disease-causing mutations (variants) in classic Mendelian inheritance disorders, the whole-exome/genome sequencing data has been used to identify CVs in complex diseases and 'orphan' diseases that occur in a very small number of individuals (e.g. Ng et al. [50]). The whole-genome sequencing approach is particularly powerful when heterogeneity is present in the disease. For example, in a study of Kabuki syndrome, 32 different mutations were found in the *MLL2* gene among the 53 families, including 12 de novo mutations [51]. Under this setting, the exceedingly low allele frequency of individual causal mutations makes misclassification a major problem. This documents the clinical importance of identifying very rare variants. The focus of our current work is an assessment of a new test statistic applied to more common variants (e.g. where the CVF is at least 1%). We provide more information in the Discussion.

Another issue affecting statistical power is that of multiple testing. Because the number of CVs observed will most certainly be larger than the number of common

Table 1. Notation for the $LTT_{\text{ae,NGS}}$ statistic

Throughout this work, the superscript 't' indicates the true value of a variable.

The following notation is in force:

M = The number of loci/sites in the study.

i_k^t = The phenotype code of individual k .

The value $i_k^t = 1$ means the individual is a case (affected), and the value $i_k^t = 0$ means the individual is a control (unaffected).

$\mathbf{x}_k = (x_{1,k}, \dots, x_{M,k})$ = The CV count vector for individual k .

$\mathbf{v}_k^t = v_{1,k}^t, \dots, v_{M,k}^t$ = The coverage vector for individual k .

$\mathbf{G}^t$ = The vector of true genotypes at the M loci; this vector can take on 3^M values, denoted by $(j_1^t, \dots, j_M^t)$.

$\mathbf{g}_{(j_1^t, \dots, j_M^t)} = \Pr(\mathbf{G}^t = (j_1^t, \dots, j_M^t))$ = The genotype frequency of the multi-locus genotype $\mathbf{G}^t$ in the population.

$f_{i_k^t, \mathbf{G}^t} = f_{i_k^t(j_1^t, \dots, j_M^t)} = \Pr(i_k^t | \mathbf{G}^t = (j_1^t, \dots, j_M^t))$ = Penetrance of phenotype code i_k^t for individuals with genotype vector $\mathbf{G}^t = (j_1^t, \dots, j_M^t)$.

α = Baseline odds ratio.

β = log odds ratio.

$w_{\mathbf{G}^t}$ = Weight corresponding to the multi-locus genotype $\mathbf{G}^t$. As above, we use dominant weight parameterization. Specifically:

$$w_{\mathbf{G}^t} = \begin{cases} 0, & \mathbf{G}^t = (0, \dots, 0) \\ 1, & \text{otherwise} \end{cases}.$$

$\alpha + \beta w_{\mathbf{G}^t} = \ln(f_{1, \mathbf{G}^t} / f_{0, \mathbf{G}^t})$. Note: as with the single-locus association situation, the mixing proportion in the case group is given by:

$$f_{1, \mathbf{G}^t} = f_{0, \mathbf{G}^t} \times e^{\alpha + \beta w_{\mathbf{G}^t}}.$$

Also, using the notation $\mathbf{G}^t = (j_1^t, \dots, j_M^t)$, we have:

$$f_{0, \mathbf{G}^t} = \frac{1}{1 + e^{\alpha + \beta w_{\mathbf{G}^t}}},$$

so that:

$$f_{1, \mathbf{G}^t} = \frac{e^{\alpha + \beta w_{\mathbf{G}^t}}}{1 + e^{\alpha + \beta w_{\mathbf{G}^t}}}.$$

H_0 for our test statistic is: $\beta = 0$.

Under H_0 :

$$f_{i_k^t, \mathbf{G}^t} = f_{i_k^t, *} = \frac{(e^\alpha)^{i_k^t}}{1 + e^\alpha}.$$

$\tau_{\mathbf{G}^t, k} = \tau_{(j_1^t, \dots, j_M^t), k}$ = Conditional probability that individual k has true genotype vector $\mathbf{G}^t = (j_1^t, \dots, j_M^t)$, where conditioning is on the observed data $\mathbf{x}_k$.

SNPs observed in any data set, any statistical procedure that tests all CVs must pay a larger 'penalty' in terms of correcting for multiple tests, thereby reducing power further. Several authors have considered the issue of multiple testing and have sought ways to address it in genetic studies [52–57]. One solution is to determine the 'true' number of independent SNPs, so that the multiple testing correction is not as egregious. Ott and colleagues also worked on this problem [58–62]. Specifically, they considered the sums of single-locus statistics, and corrected for multiple testing through permutation. An advantage of this approach is that it incorporates the linkage-disequilibrium (correlation) structure among individual markers. Also, the sum statistic provides a single p value even though multiple markers are used, so no correction for multiple testing is necessary. This approach has also been incorporated in work by Purcell et al. [63] and Zhou and Pan [64], among others.

In this work, we present a statistic based on the trend test developed by Cochran [65] and Armitage [66]. It is a likelihood ratio test in which the observed data are phenotype and NGS data, specifically the coverage and number of observed CV counts at one or more polymorphic non-synonymous coding SNPs. We note that work using the trend test for rare variants has already been published [67]. These authors found that in disease models with only rare risk variants, an statistical method based on the Cochran-Armitage trend test had power comparable to or greater than tests that pool (i.e. bin) rare variants. The authors concluded that efficient locus-wide inference using single-variant test statistics should be reconsidered as a useful framework for devising powerful association with rare-variant sequence data. As noted above, in this work we focus on slightly more common CVs.

Our statistic extends the work by Kinnamon et al. [67] in that it probabilistically estimates the true (unobserved) genotype based on the NGS data, and also estimates the sequence misclassification error probabilities in cases and controls (either separately or jointly). We evaluate the performance of our statistic on null and alternative data using simulations. We also use empirical 1000 Genomes SNP data to generate a fictitious disease phenotype, and apply our test statistic to such data to determine the p value and estimate of various parameters of interest.

Methods

We begin by providing some notation in table 1. We follow with a description of the test statistic, simulations, and empirical data.

log-Likelihoods and Computation of Test Statistic

Here, we present the log-likelihoods of the observed data, and we state the definition of the linear trend test allowing for errors applied to NGS ($LTT_{ae,NGS}$) in terms of the log-likelihoods.

$$\ln(L_{H_d}) = \sum_{k=1}^N \ln \Pr(\mathbf{x}_k, \mathbf{v}_k^t, \mathbf{i}_k^t) \\ = \text{log-likelihood equation of the observed data.}$$

Note:

$$d = 0 \text{ for the null hypothesis: } (H_0: \beta^t = 0),$$

$$d = 1 \text{ for the alternative hypothesis: } (H_1: \beta^t \neq 0).$$

$\ln(\hat{L}_{H_d})$ is the maximum log-likelihood of the data for each hypothesis. This maximum is achieved by applying the EM algorithm in the following way:

- (1) Specify a certain number of starting points (i.e. randomly generated vectors $\vec{\psi}$ of parameter settings for α, β , etc.).
- (2) For each vector $\vec{\psi}$ in item 1, update the log-likelihoods under each hypothesis until some stopping condition is satisfied, such as:

$$|(r+1)\text{-st step of } \ln(L_{H_d}) - (r)\text{-th step of } \ln(L_{H_d})| < \delta, \quad (1)$$

for some tolerance δ . In this work, we use $\delta = 0.00001$. The maximum log-likelihood is then the (r) -th step of $\ln(L_{H_d})$. We denote this value by

$$\ln(L_{H_d})_{r(\vec{\psi})}.$$

Note that for an arbitrary vector $\vec{\psi}$ in item 1, if the stopping condition (eq. 1) is not met after the maximum number of steps, we define the log-likelihood as:

$$\ln(L_{H_d})_{r_{\max}(\vec{\psi})},$$

where $r_{\max}$ is the total number of steps specified for the EM algorithm. For example, in the Simulations section below, $r_{\max} = 1,000$ (item xi).

- (3) We define the log-likelihood of the observed data, denoted $\ln(\hat{L}_{H_d})$, as:

$$\ln(\hat{L}_{H_d}) = \max_{\vec{\psi}} \left(\ln(L_{H_d})_{r(\vec{\psi})} \right).$$

$LTT_{ae,NGS} = 2[\ln(\hat{L}_{H_1}) - \ln(\hat{L}_{H_0})]$. As noted above in item 3, the caret symbol $\wedge$ indicates that we have obtained the maximum log-likelihood of the data under the particular hypothesis. Note that $LTT_{ae,NGS}$ is asymptotically a χ^2 distribution with 1 degree of freedom. We design the $LTT_{ae,NGS}$ statistic to be robust to differential misclassification error among cases and controls, i.e. we specify $\epsilon_0^t \neq \epsilon_1^t$.

Simulations

To evaluate the type I error rate and power of the test statistic under different scenarios, we perform simulations. In this section, we describe simulations where we consider type I error rate and power for a single CV. First, we simulate observed data for a single locus containing a CV for each individual based on their phenotype status. Next, we compute observed data for other loci based on a correlation coefficient.

The parameter settings that we consider for the first locus are:

(i) Disease MOI:	Dominant, recessive, multiplicative;
(ii) Prevalence (ϕ):	0.05, 0.15;
(iii) Genotype relative risk (R_2):	1.0 (H_0), 2.5, 5.0;
(iv) CVF:	0.01, 0.05;
(v) Coverage v_k^t (for all k):	3, 8, 25;
(vi) Control error probability ϵ_0^t :	0.001, 0.04;
(vii) Case error probability ϵ_1^t :	0.001, 0.04;
(viii) Number of cases:	500;
(ix) Number of controls:	500;
(x) Number of starting points:	200;
(xi) Number of EM steps per starting point:	1,000;
(xii) Tolerance ϵ :	10^{-5} ;
(xiii) Number of replicates per vector (i–iv):	500 (H_0), 250 (H_1);
(xiv) Number of loci:	5, 8;
(xv) Correlation ρ :	0.90.

Under item iii, we specify that the disease locus is in Hardy-Weinberg equilibrium. Also, the genotype relative risk (R_2) is defined in terms of the penetrances f_i , $0 \leq i \leq 2$ and the disease MOI. More specifically, $f_i = \Pr(\text{affected} | i \text{ copies of the CV at a locus})$,

$$R_1 = \frac{f_1}{f_0}, R_2 = \frac{f_2}{f_0}$$

[68].

To determine genotypes at the remaining loci the $LTT_{ae,NGS}$ statistic for multiple loci, we use the correlation coefficient specified under item xv.

For any consecutive pair of loci, genotype frequencies (and thus CVFs) conditional on affection status are computed using the well-known correlation coefficient method for determining linkage disequilibrium. A full description is given in the online supplementary Appendix (for all online suppl. material, see www.karger.com/doi/10.1159/000346824).

For the null simulations, using parameter settings i–vii, and noting that disease MOI reduces to 1 MOI when $R_2 = 1$, we have a total of 96 simulation vectors. Also, for the alternative simulations, using the same parameter settings and noting that R_2 has only two values, we compute a total of 576 simulation vectors.

We compute the empirical power (see below for the H_0 term) at different significance levels (5%, 1%) for each of these settings by computing the $LTT_{ae,NGS}$ for each of the replicates listed under item xiii, determining the corresponding p value assuming a central χ^2 distribution with 1 degree of freedom under H_0 , and evaluating the proportion of replicates for which the p value is less than the corresponding significance level. When H_0 is true, we refer to the proportion as the empirical type I error rate for a given vector of parameter settings. When H_1 is true, we refer to the proportion as the empirical power for a given vector of parameter settings.

Comparison of the $LTT_{ae,NGS}$ Statistic with CMAT and SKAT Statistics

We compare our statistic with two other commonly used statistics that test for association with NGS data. They are: (a) the cumulative minor-allele test (CMAT) [69], and (b) the sequence

Table 2. Non-synonymous chromosome 20 SNPs selected for multi-locus power calculation

SNP	Position, bp	Reference	Alternative
rs2274934	60897487	C	T
rs3737137	60897721	C	T
rs6143021	60897772	T	C
rs3810548	60905878	A	G
rs13042941	60908969	T	C

Position = Position of SNP on chromosome 20; Reference = reference allele/variant; Alternative = alternative allele/CV.

kernel association test (SKAT) [70]. We choose these statistics because: first, they test different components of a compound H_0 , as recently documented by Liu et al. (N. Tintle, pers. commun.; see Acknowledgements); second, computation is straightforward, so that we can easily program each statistic; and third, each has been well-cited (according to ISI Web of Science).

For the three statistics, we simulate CV count data for multiple loci in cases and controls. We use our statistic to determine (through the EM algorithm) the probabilities of each genotype at each locus. We use this information to determine the genotype probability calculations for SKAT and CMAT (see online suppl. Appendix). We compute empirical type I error and empirical power values as described above.

1000 Genomes Data Example

Using information from the most recent 1000 Genomes SNP annotation [71], we select 5 non-synonymous SNPs within a maximum range of 20 kilobases of each other for multiple persons. The SNPs are listed in table 2. The individuals chosen come from the following populations:

- (i) ASW (African ancestry in Southwest US);
- (ii) CEU (Utah residents (CEPH) with Northern and Western European ancestry);
- (iii) FIN (Finnish natives);
- (iv) GBR (British in England and Scotland);
- (v) TSI (Toscani in Italy);
- (vi) YRI (Yoruba in Ibadan, Nigeria).

We remove all individuals related to the probands of our data set, so that all individuals in our remaining data set are unrelated. Next, we download these unrelated individuals' sequenced raw file of exome alignment from the 1000 Genomes Project FTP site [72]. The region of interest we specify (chromosome 20: base-pair positions 60,897,487–60,908,969) is extracted from each individual by applying SAMtools [73]. This region is chosen because chromosome 20 is a relatively small chromosome and the corresponding data sets are likewise much smaller and easier to process computationally.

With the 'mpileup' function in SAMtools, we align the extracted region to the reference human genome Build 37 [74]. Then we

perform SNP calling using BCFtools. Following this protocol, we obtain coverage and the number of CV reads for each of the 5 SNP positions. Finally, we filter individuals one last time by removing each individual whose coverage at any of the SNPs is ≤ 1 . We obtain a final set of 388 unrelated individuals.

Next, we create a fictitious disease phenotype as follows:

- (1) For each individual's vectors of coverage $\mathbf{v}_k^t = v_{1k}^t, \dots, v_{5k}^t$ and CV reads $\mathbf{x}_k = x_{1k}, \dots, x_{5k}$, we compute a new vector, the proportion of each SNP's coverage that consists of CV reads. In mathematical terms, we create

$$\mathbf{U}_k = (u_{1k}, \dots, u_{5k}) = \left(\frac{x_{1k}}{v_{1k}^t}, \dots, \frac{x_{5k}}{v_{5k}^t} \right).$$

- (2) For each individual, we compute the linear combination

$$L_k = \sum_{m=1}^5 b_m u_{mk}.$$

The weights we specify are: $b_1 = -1.0$, $b_2 = 0.5$, $b_3 = 0.0$, $b_4 = 8.0$, $b_5 = 5.0$. These weights are selected to provide a range of effects. For example, weight b_1 is -1.0 , indicating that the CV is protective, which is contrary to our stated specification. Weights b_2 and weight b_3 are 0.5 and 0.0, respectively, which means they have little or no effect on disease risk. Weights b_4 and b_5 are rather large, suggesting that, if we find association, it is due to effects of SNPs 4 and 5.

- (3) We compute the median L_k value over all 388 individuals. We define as 'affected' any individual whose L_k value is above the median, and define as 'unaffected' any individual whose L_k value is below the median. In this way, we obtain exactly 50% cases and 50% controls.

Finally, we apply the $LTT_{ae,NGS}$ to this data set and compute significance, as well as parameter estimates.

Results

Comparison of the $LTT_{ae,NGS}$ Statistic with CMAT and SKAT Statistics

Simulations under H_0

Here, we present results of the null simulations. Specifically, we report summary statistics for the empirical type I error rates with different settings of coverage (v_k^t). These results may be found in figure 1 (results presented in box and whisker plots [75]). In this figure, the empirical type I error rates are computed only for the differential misclassification simulation settings ($\epsilon_0^t \neq \epsilon_1^t$) at the 5% significance level (results for the 1% level are very similar and therefore not presented). Each box and whiskers plot is computed from 16 empirical type I error rates. The number 16 comes from 2 settings each for: (1) Φ ; (2) CVF; (3) number of loci, and (4) differential error rates ($\epsilon_0^t \neq \epsilon_1^t$).

One finding is immediate. The $LTT_{ae,NGS}$ statistic appears to maintain correct type I error rate for all coverage settings in the presence of differential misclassification.

Fig. 1. Box and whiskers plots of empirical type I error rates for $LTT_{ae,NGS}$, CMAT, and SKAT statistics with different coverage settings. Each box and whiskers plot is determined using empirical type I error rates for 16 simulations (2 settings each for: (1) number of loci; (2) Φ ; (3) CVF; (4) $\varepsilon_0^i \neq \varepsilon_1^i$; and 3 settings for ν^t (see notation in Methods – Simulations)).

The vector values (1st quartile, median, 3rd quartile) for coverages $\nu_k^t = 3, 8, 25$ are: (0.050, 0.058, 0.066), (0.055, 0.060, 0.067), (0.051, 0.060, 0.072), respectively, for the $LTT_{ae,NGS}$ statistic. By contrast, for the same coverage ordering, the CMAT vector values are: (0.913, 0.921, 0.975), (0.341, 0.353, 0.367), (0.048, 0.053, 0.059). Likewise, the SKAT vector values are: (0.918, 0.935, 0.964), (0.340, 0.364, 0.374), (0.051, 0.057, 0.063). One may see all these quantities in figure 1.

Note that, when the coverage is 3, the non-differential statistics have empirical type I error rates that are approximately 18 times the desired rate (0.05). More worrisome, from our perspective, is the fact that the median empirical type I error rates for these statistics are near 1.00, i.e. one would almost always reject the true H_0 if simulation settings are correct. When the coverage is 8, the non-differential statistics have empirical type I error rates that are approximately 7 times the desired rate (0.05). While the empirical type I error rates for this coverage are much closer to the 5% significance level, they still indicate substantial inflation in type I error. Only when the coverage rises to 25 do the empirical type I error rates for all three

statistics appear to approximately equal to the 5% significance level. These results strongly suggest that, at lower coverage rates, not allowing for differential misclassification may result in a (sometimes substantial) increase in type I error for an NGS statistic. We note that a non-differential form of our $LTT_{ae,NGS}$ statistic produced similar inflation in empirical type I error to the CMAT and SKAT statistics (results not shown).

As previous research has documented [11, 26, 30, 36, 76–78], when applying a statistic test that does not account for the potential differential genotype misclassification error, we observe an inflated empirical type I error rate. This makes sense because, unless we account for the differential errors in the analysis, the differential errors manifest themselves as a true phenotype-genotype association. Another question is raised by studying figure 1: why is there substantial inflation in empirical type I error for lower coverage, but none observed for higher coverage? This question may be answered by observing that we specify the probability of the observed CV counts conditional on the underlying true genotype (and other parameters) as a binomial distribution (see online suppl. Ap-

Table 3. Summary statistics for empirical power differences (CMAT – $LTT_{ae,NGS}$) and (SKAT – $LTT_{ae,NGS}$)

Statistic	Disease MOI	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
CMAT	Dominant	–0.096	0.000	0.000	0.004	0.004	0.100
	Multiplicative	–0.080	–0.004	0.010	0.012	0.032	0.076
	Recessive	–0.040	–0.012	0.004	0.014	0.030	0.104
SKAT	Dominant	–0.148	0.000	0.000	0.009	0.004	0.144
	Multiplicative	–0.116	0.000	0.016	0.022	0.048	0.104
	Recessive	–0.047	–0.008	0.008	0.016	0.028	0.112

All empirical powers were computed at the 5% significance level only for non-differential error simulations.

pendix, section entitled ‘Component probability mass function for Causal Variant (CV) count data for single locus’). Consider the values: $v_k^t = 3$, $j^t = 0$ genotype (homozygous for the non-CV allele), and misclassification rate $\varepsilon_j^t = 0.04$. The probability of an observed count $x_k = 1$ is 11% (we drop the m subscript from x to indicate we are considering data for a single variant). If we consider the heterozygote: $j^t = 1$, the probability of an observed count $x_k = 1$ is 37.5%. From these values, we see that it is somewhat difficult to determine with certainty the true underlying genotype.

Now consider the values: $v_k^t = 25$, $j^t = 0$ genotype, and misclassification rate $\varepsilon_j^t = 0.04$. The probability of an observed count $x_k = 1$ is 37.5%. For the heterozygote: $j^t = 1$, the probability of an observed count $x_k = 1$ is 7.45E-05% (virtually 0). Also, for the $j^t = 2$ genotype, the probability is 6.76E-31% (we would call this value 0). From these results, we would conclude with near certainty that the underlying genotype is $j^t = 0$. Hence, even in the presence of misclassification error, higher coverage enables us to determine (much more precisely) the true underlying genotypes. It is for this reason that all three statistics maintain a correct type I error rate when $v_k^t = 25$.

Finally, we note that, for the non-differential error simulations ($\varepsilon_0^t = \varepsilon_1^t$), all three statistics demonstrate correct empirical type I error rates at all coverages (results not shown). Therefore, we present empirical power results only for the non-differential settings.

Simulations under H_1

In table 3, we present summary statistics of the empirical power differences (CMAT – $LTT_{ae,NGS}$) and (SKAT – $LTT_{ae,NGS}$) at the 5% significance level. Results are stratified by the disease MOI. Studying this table, we observe that the median and mean empirical power dif-

ferences are close to 0 for nearly all disease MOI and statistics. Similarly, the mean differences are also close to 0.

Another observation is that the power differences are slightly skewed in favor of the CMAT and SKAT statistics. Specifically, the 3rd quartile values are all larger than the corresponding 1st quartile values for all rows. Having said this, we note that the minimum difference (in absolute value) is close to, if not greater than, the corresponding maximum value for each row, with the exception of the recessive disease MOI. We conjecture that power differences are ‘more positive’ for the recessive disease MOI because our statistic is really a ‘dominant weight’ statistic [79], i.e. *any* multi-locus genotype containing at least one CV is given a weight of 1, and the only multi-locus genotype with a 0 weight is the one containing no CVs.

Because the empirical powers of the three methods are so similar, we focus on power results for our statistic. We run an analysis of variance (ANOVA) as implemented in the R software package [80] to determine what factors most significantly affect empirical power at the 5% level (results for the 1% significance level are nearly identical and therefore not shown). The factors we consider as input variables are items i–viii and xiv listed in the Methods section (under Simulations) and all two-way interactions. The response variable is the empirical power corresponding to the specific simulation vector. Note that, because the $LTT_{ae,NGS}$ statistic is robust to differential misclassification, we include empirical power results for differential error as well.

Here, we present just the terms of the ANOVA whose F statistic have a p value <2.00E-16 (full results not shown). In order of their F-statistic values (largest to smallest), the most significant factors are: (1) disease MOI

Fig. 2. Box and whiskers plots of empirical powers for $LTT_{ae,NGS}$ for different settings of disease MOI and CVF. Each box and whiskers plot is determined using empirical powers for 96 simulations (2 settings each for: (1) number of loci; (2) Φ ; (3) R_2 ; (4) ε_0^t ; (5) ε_1^t ; and 3 settings for v^t (see notation in Methods – Simulations)). Dom = Dominant; Mult = multiplicative; Rec = recessive disease MOI, respectively.

($F = 13,119$); (2) CVF ($F = 2,372$); (3) R_2 ($F = 1,607$); (4) disease MOI $\times$ CVF ($F = 928$); (5) disease MOI $\times$ R_2 ($F = 488$), and (6) CVF $\times$ R_2 ($F = 164$).

To illustrate the importance of the first two factors (disease MOI and CVF) on empirical powers, we present a box and whiskers plot of the powers, stratified by settings for these two factors (fig. 2). Examining this figure, we observe that minimum, median, and maximum power can vary substantially, depending upon the disease MOI. For example, the minimum power for the dominant MOI is 0.5, when the CVF is 0.01. By contrast, the maximum power for the recessive MOI is 0.14, when the CVF is 0.05.

It is also reasonably clear that the CVF makes a difference when considering the disease MOI. For example, when the CVF is 0.01, the dominant MOI's most empirical powers range from 0.70 (1st quartile) to 1.00 (3rd quartile). By contrast, when the CVF is 0.05, power is virtually 1.00 for all other settings. Also, for the multiplicative MOI, when the CVF is 0.01, empirical powers range primarily from 0.20 (1st quartile) to 0.55 (3rd quartile).

However, when the CVF is 0.05, empirical powers increase so that the range is mostly 0.67 (1st quartile) to 0.99 (3rd quartile).

1000 Genomes Data Example

When running the $LTT_{ae,NGS}$ on the 1000 Genomes data set, we obtain a test statistic value of 42.643, with a corresponding p value of $6.57E-11$. The final parameter estimates are: $\hat{\alpha} = -13.144$, $\hat{\beta} = 13.334$, $\hat{\varepsilon}_0 = 0.017$, $\hat{\varepsilon}_1 = 0.014$, under H_1 . The genotype frequencies are not listed, since there are 243 such frequencies. The caret above each parameter indicates that this is a maximum likelihood estimate. It is interesting to note that the error values are independent of the fictitious disease model, i.e. it appears that there is an approximately 1.5% probability of misclassifying sequence data in the 1000 Genomes Project for these individuals and SNPs. Finally, we note that, for our disease model, the increase in risk of getting the disease is $e^{(-13.144 + 13.334)} = 1.21$ for any individual harboring a multi-locus genotype with at least one CV.

Summary and Discussion

The purpose of this work is to develop a linear trend test that can utilize phenotype and NGS data to perform a case-control genetic association analysis. As mentioned in the Introduction and documented in the online supplementary Appendix, our statistic probabilistically estimates the true (unobserved) genotype based on the NGS data, and also estimates the sequence misclassification error probabilities in cases and controls separately. Based on the results of simulations, our statistic appears to maintain the correct type I error rate in the presence of either differential or non-differential misclassification. Also, empirical power appears to be comparable to statistics such as the CMAT and SKAT statistics, depending upon the R_2 and CVF. One interesting result from the 1000 Genomes SNP data set is that our statistic estimates sequence misclassification errors of approximately 1.5% for the SNPs sequenced.

One of the key findings in this work is that the development of such a statistic is indeed possible; in the online supplementary Appendix, we present closed-form solutions of all necessary parameters. From this information, we may be able to extend the statistic to other models that are binary, in the sense that weights are either 0 or 1. We performed such computations for our linear trend test using double sampling [32].

In previous research, we found that phenotype misclassification error has much larger effect, in terms of power loss, than genotype misclassification error. The one kind of genotype misclassification error that can be vexing is differential genotype misclassification. As noted in the Introduction, and as documented in our Results (fig. 1), differential genotype misclassification, even at low rates, can produce substantially inflated type I error rates [11, 26, 30, 36, 78], so that the null distribution of the test statistic is no longer known. For that reason, we recommend that statistics like $LTT_{ae,NGS}$ be used when testing for association, or that already developed NGS statistics be modified to allow for differential misclassification. This recommendation is particularly important for designs with low coverage. In results not shown, even when error is non-differential, there is practically no power loss for the $LTT_{ae,NGS}$ statistic as compared with a non-differential form of the same statistic, so we suggest always using the statistic we report in this work. Such a suggestion is analogous to recommendations that one should use the H-LOD statistic (as compared with the LOD statistic) when performing linkage analysis, or that one should use the t test allowing for unequal variance as

compared with the one that assumes equal variance in two groups; that is, these statistics are robust to some deviation from a common assumption (e.g. linkage homogeneity, equal variance) and there is very little power loss when using these statistics.

We are aware that the format of NGS data is evolving. The latest set of data from the 1000 Genomes Project has actual genotype calls along with quality scores. One of the nice features of likelihood-based methodology is that, as long as probabilities can be determined, statistics can be computed for multiple formats. We look forward to developing such statistics.

As mentioned in the Introduction, we are keenly aware of the clinical importance regarding identification of very rare variants. While the focus this work has been an evaluation of the $LTT_{ae,NGS}$ statistic applied to more common variants ($1\% \leq CVF \leq 5\%$ for each CVF), we note that our statistic can handle any CVFs and determine accurate results. The reason is that the asymptotic distribution of the test statistic is more accurate as the number of loci considered is increased (irrespective of the CVFs), needing less reliance on permutation methods to compute significance. Also, our statistic was designed to handle multiple CVs, since increase in the number does not affect the degrees of freedom.

Is it possible to determine differential misclassification without the application of our statistic? We think so. In fact, we are aware of at least one paper [78] that looked at Q-Q plots to determine differential misclassification. If researchers have a sufficient number of polymorphic sites for which there is NGS data, we recommend using Q-Q plots. Another diagnostic test involves checking whether the missing rates among genomic sites differ between cases and controls [81]. While neither of these methods is definitive in determining differential misclassification (only double-sampling [82] can provide near-perfect evidence), results that deviate from expected results can be strongly suggestive. We note that not all researchers will perform either whole-exome or whole-genome NGS studies; some may perform candidate region studies with a relatively small number of sites tested. In those instances, Q-Q plots may not be as helpful, as they may not have a sufficient number of data points. In either situation, we comment that results from our simulations suggest that the $LTT_{ae,NGS}$ is robust to differential misclassification. Therefore, regardless of the nature of the misclassification, we anticipate that our statistic will provide valid results.

Along with the comments above, our statistic may be used in a confirmatory analysis for researchers who have found significant association with other methods. In this

way, researchers can protect themselves from following up false positives.

Finally, we have developed an R version of the analysis software. In addition, we are developing a C version of the software for results using permutation testing. We plan to make these items available on the web by May 2013 to any interested researchers.

Acknowledgements

J.X. is supported by the National Institutes of Health, National Human Genome Research Institute (R00HG005846). T.C.M. is supported by Population Architecture Using Genetics and Epide-

miology (PAGE), which is funded by the National Human Genome Research Institute (NHGRI) and is supported by the National Institutes of Health grant U01HG004801 (Coordinating Center). The authors gratefully acknowledge E. Génin, who provided the authors with a much-needed deadline extension that enabled the authors to complete this work. Also, the authors express sincere gratitude to Dr. Nathan Tintle, who provided the authors with a draft of a manuscript, now in revision, in which Dr. Tintle and colleagues provide an elegant parameterization of NGS statistics. Results from this article were critical in assisting the authors to choose NGS statistics for comparison with the $LTT_{ae,NGS}$. Finally, the authors thank two anonymous reviewers, whose comments helped the authors draft a much more complete revision.

References

- Ott J: Analysis of Human Genetic Linkage, 3 ed. Baltimore, Johns Hopkins University Press, 1999.
- Purcell S, Cherny SS, Sham PC: Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003;19:149–150.
- Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–213.
- Edwards BJ, Haynes C, Levenstien MA, Finch SJ, Gordon D: Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet* 2005;6:18.
- Lamina C, Kuchenhoff H, Chang-Claude J, Paulweber B, Wichmann HE, Illig T, Hoehe MR, Kronenberg F, Heid IM: Haplotype misclassification resulting from statistical reconstruction and genotype error, and its impact on association estimates. *Ann Hum Genet* 2010;74:452–462.
- Kennedy J, Mandouli I, Pasaniuc B: Genotype error detection using Hidden Markov Models of haplotype diversity. *J Comput Biol* 2008;15:1155–1171.
- Browning BL, Browning SR: Haplotypic analysis of Wellcome Trust Case Control Consortium data. *Hum Genet* 2008;123:273–280.
- Cheng KF, Lin WJ: Simultaneously correcting for population stratification and for genotyping error in case-control association studies. *Am J Hum Genet* 2007;81:726–743.
- Cheng KF, Chen JH: A simple and robust TDT-type test against genotyping error with error rates varying across families. *Hum Hered* 2007;64:114–122.
- Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V: 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* 2004;3:Article26.
- Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC: Effects of differential genotyping error rate on the type I error probability of case-control studies. *Hum Hered* 2006;61:55–64.
- Hao K, Wang X: Incorporating individual error rate into association test of unmatched case-control design. *Hum Hered* 2004;58:154–163.
- Kang SJ, Gordon D, Brown AM, Ott J, Finch SJ: Tradeoff between no-call reduction in genotyping error rate and loss of sample size for genetic case/control association studies. *Pac Symp Biocomput* 2004;116–127.
- Morris RW, Kaplan NL: Testing for association with a case-parents design in the presence of genotyping errors. *Genet Epidemiol* 2004;26:142–154.
- Bernardinelli L, Berzuini C, Seaman S, Holmans P: Bayesian trio models for association in the presence of genotyping errors. *Genet Epidemiol* 2004;26:70–80.
- Barral S, Haynes C, Stone M, Gordon D: LR-Tae: improving statistical power for genetic association with case/control data when phenotype and/or genotype misclassification errors are present. *BMC Genet* 2006;7:24.
- Dahabreh IJ, Linardou H, Bouzika P, Varvarigou V, Murray S: TP53 Arg72Pro polymorphism and colorectal cancer risk: a systematic review and meta-analysis. *Cancer Epidemiol Biomarkers Prev* 2010;19:1840–1847.
- Heid IM, Lamina C, Kuchenhoff H, Fischer G, Klopp N, Kolz M, Grallert H, Vollmert C, Wagner S, Huth C, Muller J, Muller M, Hunt SC, Peters A, Paulweber B, Wichmann HE, Kronenberg F, Illig T: Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *Am J Epidemiol* 2008;168:878–889.
- Holmes MV, Perel P, Shah T, Hingorani AD, Casas JP: CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA* 2011;306:2704–2714.
- Hossain S, Le ND, Brooks-Wilson AR, Spinelli JJ: Impact of genotype misclassification on genetic association estimates and the bayesian adjustment. *Am J Epidemiol* 2009;170:994–1004.
- Ji F, Yang Y, Haynes C, Finch SJ, Gordon D: 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* 2005;4:Article37.
- Lai R, Zhang H, Yang Y: Repeated measurement sampling in genetic association analysis with genotyping errors. *Genet Epidemiol* 2007;31:143–153.
- Li J, Coates RJ, Gwinn M, Khoury MJ: Steroid 5- α -reductase Type 2 (SRD5A2) gene polymorphisms and risk of prostate cancer: a HuGE review. *Am J Epidemiol* 2010;171:1–13.
- Kang SJ, Finch SJ, Haynes C, Gordon D: Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. *Hum Hered* 2004;58:139–144.
- Kang SJ, Gordon D, Finch SJ: What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 2004;26:132–141.
- Ahn K, Gordon D, Finch SJ: Increase of rejection rate in case-control studies with the differential genotyping error rates. *Stat Appl Genet Mol Biol* 2009;8:Article25.

- 27 Ahn K, Haynes C, Kim W, Fleur RS, Gordon D, Finch SJ: The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann Hum Genet* 2007;71:249–261.
- 28 Gordon D, Haynes C, Blumenfeld J, Finch SJ: PAWE-3D: visualizing power for association with error in case-control genetic studies of complex traits. *Bioinformatics* 2005;21:3935–3937.
- 29 Marquard V, Beckmann L, Heid IM, Lamina C, Chang-Claude J: Impact of genotyping errors on the type I error rate and the power of haplotype-based association methods. *BMC Genet* 2009;10:3.
- 30 Londono D, Haynes C, De La Vega FM, Finch SJ, Gordon D: A cost-effective statistical method to correct for differential genotype misclassification when performing case-control genetic association. *Hum Hered* 2010;70:102–108.
- 31 Yang Y, Wise CA, Gordon D, Finch SJ: A family-based likelihood ratio test for general pedigree structures that allows for genotyping error and missing data. *Hum Hered* 2008;66:99–110.
- 32 Gordon D, Haynes C, Yang Y, Kramer PL, Finch SJ: Linear trend tests for case-control genetic association that incorporate random phenotype and genotype misclassification error. *Genet Epidemiol* 2007;31:853–870.
- 33 Maraganore DM, de Andrade M, Elbaz A, Farrer MJ, Ioannidis JP, Kruger R, Rocca WA, Schneider NK, Lesnick TG, Lincoln SJ, Hulihan MM, Aasly JO, Ashizawa T, Chartier-Harlin MC, Checkoway H, Ferrarese C, Hadjigeorgiou G, Hattori N, Kawakami H, Lambert JC, Lynch T, Mellick GD, Papapetropoulos S, Parsian A, Quattrone A, Riess O, Tan EK, Van Broeckhoven C: Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease. *JAMA* 2006;296:661–670.
- 34 Govindarajulu US, Spiegelman D, Miller KL, Kraft P: Quantifying bias due to allele misclassification in case-control studies of haplotypes. *Genet Epidemiol* 2006;30:590–601.
- 35 Mitchell AA, Cutler DJ, Chakravarti A: Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* 2003;72:598–610.
- 36 Plagnol V, Cooper JD, Todd JA, Clayton DG: A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet* 2007;3:e74.
- 37 Powers S, Gopalakrishnan S, Tintle N: Assessing the impact of non-differential genotyping errors on rare variant tests of association. *Hum Hered* 2011;72:153–160.
- 38 Sastre L: New DNA sequencing technologies open a promising era for cancer research and treatment. *Clin Transl Oncol* 2011;13:301–306.
- 39 Lindblom A, Robinson PN: Bioinformatics for human genetics: promises and challenges. *Hum Mutat* 2011;32:495–500.
- 40 Patrinos GP, Innocenti F, Cox N, Fortina P: Genetic Analysis in Translational Medicine: The 2010 GOLDEN HELIX Symposium. *Hum Mutat* 2011;32:698–703.
- 41 Schweiger MR, Kerick M, Timmermann B, Isau M: The power of NGS technologies to delineate the genome organization in cancer: from mutations to structural variations and epigenetic alterations. *Cancer Metastasis Rev* 2011;30:199–210.
- 42 Cloonan N, Waddell N, Grimmond SM: The clinical potential and challenges of sequencing cancer genomes for personalized medical genomics. *IDrugs* 2010;13:778–781.
- 43 Weksberg R: Imprinted genes and human disease. *Am J Med Genet (Part C)* 2010;154C:317–320.
- 44 Kabesch M: Next generation genetics in allergy. *Curr Opin Allergy Clin Immunol* 2010;10:407.
- 45 Roukos DH: Next-generation, genome sequencing-based biomarkers: concerns and challenges for medical practice. *Biomark Med* 2010;4:583–586.
- 46 Chen JM, Ferec C, Cooper DN: Revealing the human mutome. *Clinical Genet* 2010;78:310–320.
- 47 Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS: The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* 2010;55:403–415.
- 48 Dias-Neto E, Nunes DN, Giordano RJ, Sun J, Botz GH, Yang K, Setubal JC, Pasqualini R, Arap W: Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PLoS One* 2009;4:e8338.
- 49 ten Bosch JR, Grody WW: Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol Diagn* 2008;10:484–492.
- 50 Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010;42:30–35.
- 51 Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010;42:790–793.
- 52 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004;74:765–769.
- 53 de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005;37:1217–1223.
- 54 Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003;100:9440–9445.
- 55 Ott J, Liu Z, Shen Y: Challenging false discovery rate: a partition test based on p values in human case-control association studies. *Hum Hered* 2012;74:45–50.
- 56 Chen Z, Liu Q, McGee M, Kong M, Huang X, Deng Y, Scheuermann RH: A gene selection method for GeneChip array data with small sample sizes. *BMC Genomics* 2011;12(suppl 5):S7.
- 57 Muller BU, Stich B, Piepho HP: A general method for controlling the genome-wide type I error rate in linkage and association mapping experiments in plants. *Heredity (Edinburgh)* 2011;106:825–831.
- 58 Hoh J, Ott J: Scan statistics to scan markers for susceptibility genes. *Proc Natl Acad Sci USA* 2000;97:9615–9617.
- 59 Hoh J, Ott J: A train of thoughts on gene mapping. *Theor Popul Biol* 2001;60:149–153.
- 60 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003;4:701–709.
- 61 Hoh J, Ott J: Genetic dissection of diseases: design and methods. *Curr Opin Genet Dev* 2004;14:229–232.
- 62 Hoh J, Wille A, Ott J: Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* 2001;11:2115–2119.
- 63 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–575.
- 64 Zhou H, Pan W: Binomial mixture model-based association tests under genetic heterogeneity. *Ann Hum Genet* 2009;73:614–630.
- 65 Cochran WG: Some methods for strengthening the common χ^2 tests. *Biometrics* 1954;10:417–451.
- 66 Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11:375–386.
- 67 Kinnamoon DD, Hershberger RE, Martin ER: Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. *PLoS One* 2012;7:e30238.
- 68 Schaid DJ, Sommer SS: Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 1993;53:1114–1126.
- 69 Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S: Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 2010;87:604–617.
- 70 Wu Michael C, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;89:82–93.
- 71 1000 Genomes Project, Phase 1, Version 3, 2012.
- 72 1000 Genome Project, Exome Alignment File, 2012.

- 73 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25: 2078–2079.
- 74 Ensemble Human Genome Build 37 – Chromosome 20, 2012.
- 75 Tukey JW: *Exploratory Data Analysis*. Upper Saddle River, Pearson Education – Addison Wesley, 1977.
- 76 Pluzhnikov A, Below JE, Konkashbaev A, Tikhomirov A, Kistner-Griffin E, Roe CA, Nicolae DL, Cox NJ: Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am J Hum Genet* 2010;87:123–128.
- 77 Cheng KF, Lin WJ: The effects of misclassification in studies of gene-environment interactions. *Hum Hered* 2009;67:77–87.
- 78 Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA: Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 2005;37:1243–1246.
- 79 Slager SL, Schaid DJ: Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum Hered* 2001;52:149–153.
- 80 R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, R Foundation for Statistical Computing. ISBN 3-900051-07-0. <http://www.R-project.org/>.
- 81 Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: Data quality control in genetic case-control association studies. *Nat Protoc* 2010;5:1564–1573.
- 82 Tenenbein A: A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics* 1972;14:187–202.