BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 15 2014, pages 2179–2188
doi:10.1093/bioinformatics/btu196

Genetics and population analysis
Advance Access publication April 14, 2014

Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic

Andriy Derkach[1], Theodore Chiang[2], Jiafen Gong[2], Laura Addis[3], Sara Dobbins[4], Ian Tomlinson[5], Richard Houlston[4], Deb K. Pal[3] and Lisa J. Strug[2,6,*]

[1]Department of Statistical Science, University of Toronto, Toronto, ON, Canada, [2]Program in Child Health Evaluative Sciences, the Hospital for Sick Children Research Institute, Toronto, ON, Canada, [3]Department of Clinical Neuroscience, Institute of Psychiatry, King's College London, London, [4]Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, [5]Molecular and Population Genetics and NIHR Comprehensive Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK, [6]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Sufficiently powered case–control studies with next-generation sequence (NGS) data remain prohibitively expensive for many investigators. If feasible, a more efficient strategy would be to include publicly available sequenced controls. However, these studies can be confounded by differences in sequencing platform; alignment, single nucleotide polymorphism and variant calling algorithms; read depth; and selection thresholds. Assuming one can match cases and controls on the basis of ethnicity and other potential confounding factors, and one has access to the aligned reads in both groups, we investigate the effect of systematic differences in read depth and selection threshold when comparing allele frequencies between cases and controls. We propose a novel likelihood-based method, the robust variance score (RVS), that substitutes genotype calls by their expected values given observed sequence data.

Results: We show theoretically that the RVS eliminates read depth bias in the estimation of minor allele frequency. We also demonstrate that, using simulated and real NGS data, the RVS method controls Type I error and has comparable power to the 'gold standard' analysis with the true underlying genotypes for both common and rare variants.

Availability and implementation: An RVS R script and instructions can be found at strug.research.sickkids.ca, and at https://github.com/strug-lab/RVS.

Contact: lisa.strug@utoronto.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 17, 2013; revised on March 3, 2014; accepted on April 9, 2014

1 INTRODUCTION

Genome-wide case–control association studies with single nucleotide polymorphisms (SNPs) can incorporate publicly available genome-wide control data for comparison to disease group allele frequencies (The Wellcome Trust Case Control Consortium, 2007). This convenient strategy not only increases statistical power by using larger numbers of controls but also allows precious resources to be focused on data collection from diseased individuals. Publicly available genome-wide next-generation sequence (NGS) data exist [e.g. 1000 Genomes Project (Abecasis *et al*., 2012)]; however, it has generally been used as a tool to identify novel or rare variants in individual studies, rather than as a control group for association analysis. One explanation for this underutilization of NGS genotypes for association may be the bias that results from sequencing cases and controls with different genomic platforms and protocols; these biases tend to be less of a concern when using microarray data if properly accounted for (Sebastiani *et al*., 2011).

A more commonly implemented design for association with NGS data is to use sequenced cases for variant discovery and then genotype the identified variants in a larger sample of cases and controls (Liu and Leal, 2012; Longmate *et al*., 2010; Sanna *et al*., 2011). Such two-step sampling designs can be cost effective and can ensure there is no Type I error inflation. However, this approach cannot detect protective variants that are present only in the discovery sample and in general are overly conservative. Here we develop statistical methodology for a design in which publicly available sequenced controls are used for association studies with 'in-study' NGS sequenced cases to prioritize variants for further investigation. Public controls could augment 'in-study' sequenced controls or, in the case that we consider here, public controls could be the only control group used for analysis.

Possible confounders that could influence findings when using an external NGS control group in genetic association studies can be divided into two general categories: (i) those that can be controlled by design considerations such as appropriately matched control groups on ethnicity (i.e. basic epidemiologic principles unrelated to the type and production of genetic data); and (ii) factors directly related to the sequencing and variant calling technology (Nielsen *et al*., 2011): base calling procedures (e.g. various sequencing platforms), alignment (e.g. algorithm and reference genome), read depth, SNP detection and genotype calling algorithms (DePristo *et al*., 2011; Li *et al*., 2009; McKenna

*To whom correspondence should be addressed.

et al., 2010). For example, large publicly available datasets that were sequenced at low read depth (LRD) can result in biased estimation of allele frequencies. This bias diminishes with increasing read depth (Kim *et al.*, 2011). However, if allele frequencies are compared from variants sequenced with different read depth in case and control cohorts, false associations may be generated and true ones masked. Even when average read depth is similar coverage could vary in individual regions across platforms, samples and experiments. This would likewise bias results in regions with low coverage in one group by chance, and preclude comparison between cases and controls in regions completely lacking sequence in one group or the other. Lastly, differences in SNP discovery and variant calling algorithms can also lead to spurious association findings. As a consequence of these shortcomings, statistical methodology designed to assess association in sequence data has generally required both cases and controls to be sequenced together using a common platform, depth and design.

Here, we focus on the technical aspects of comparing allele frequencies between cases and controls that were sequenced as part of different projects with different experimental designs. When the matched case and control groups with their aligned NGS data (e.g. binary version of Sequence Alignment/MAP (BAM) files) are available, we can apply the variant calling algorithm to the combined data so that the resulting case and control data would be well-matched with the exception of design parameters such as enrichment strategy, sequencing platform, read depth and resulting coverage (Fig. 1).

It is well-documented that differences in read depth between cases and controls have large effects on estimation of minor allele frequency (MAF) and can lead to inflated Type I error in association studies (Kim *et al.*, 2011). Less attention has been paid to the selection threshold used in genotype calling algorithms (genotypers) such as those implemented in Samtools and Genome Analysis Toolkit (GATK) (DePristo *et al.*, 2011; Li *et al.*, 2009; McKenna, *et al.*, 2010). These genotypers provide confidence/quality scores for a genotype call (e.g. GQ scores) and based on these scores and a predetermined threshold, low confidence/quality calls are filtered out.

To address variant call differences that can occur even within a study design that sequences cases and controls, one could incorporate read depth or quality score differences into the association analysis by using a logistic regression analysis with read depth as a covariate or by weighting each variant call by quality score (Daye *et al.*, 2012; Garner, 2011). However, in the setting where cases and controls are distinguishable by read depth, these approaches are not applicable because they would be confounded by case–control status, and the corresponding parameters would not be estimable. Another approach to account for differential read depth is implemented in the GATK toolkit (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_PrintReads.html), which randomly down-samples BAM files for the higher read depth group. This approach is a less powerful strategy in comparison with methodology that incorporates all observed data, as we show in Section 3.2.

When cases and controls are sequenced as part of the same experimental design, Skotte *et al* (2012) suggest substituting genotype calls by their expected values given the observed sequence data. This can result in higher power and better control of Type I error than methods based on called genotypes, while taking into account uncertainty in the calls without requiring filtering by arbitrary quality score thresholds. This approach incorporates read depth by constructing the joint likelihood of observed phenotypes and observed sequence data, and significance testing is conducted using a score statistic. This approach, however, would not control Type I error when there are case–control differences in read depth because these differences may produce inflated estimates of the score statistic variance, especially for rare variants.

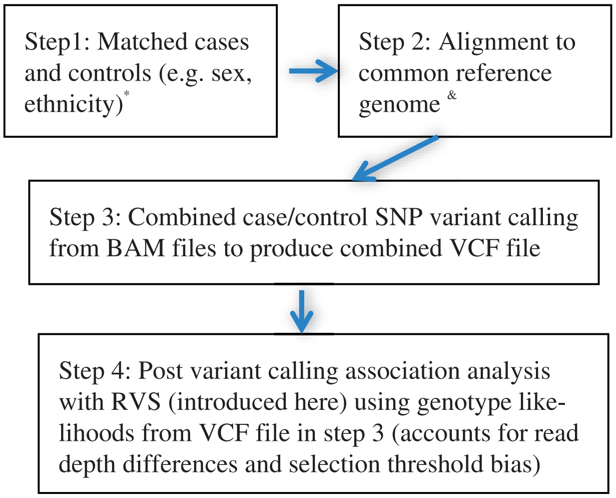

Step1: Matched cases and controls (e.g. sex, ethnicity)*
Step 2: Alignment to common reference genome &
Step 3: Combined case/control SNP variant calling from BAM files to produce combined VCF file
Step 4: Post variant calling association analysis with RVS (introduced here) using genotype likelihoods from VCF file in step 3 (accounts for read depth differences and selection threshold bias)

Fig. 1. Workflow proposed for the NGS association analysis when external NGS control data are used. We assume both case and control NGS data have passed standard quality control metrics. Asterisk indicates that ideally cases and controls would also be matched on sequencing platform and enrichment strategy. However, our results in Section 3 indicates that this is not necessary. [&]Different alignment algorithms are implicitly accounted for by the RVS because the unit of analysis is genotype probability rather than the genotype calls in the association analysis

In Section 2, we propose the robust variance score (RVS), which repurposes and extends the approach by Skotte *et al.* (2012). In Section 3.1, we illustrate analytically and by simulation how differences in read depth and variant screening parameters affect Type I error in association studies using called genotypes. We then present simulation results under the null and alternative models for association with single and multiple variants using the RVS method. In Section 3.2, we show that analysis using the RVS has comparable power to an analysis with the true genotypes. In Section 3.3, we apply the RVS to analysis of several studies using NGS technology, and we compare our findings with those from association studies using genotype calls with quality score thresholds.

2 METHODS

2.1 Defining the RVS method

We use a score statistic derived from the joint likelihood of observed phenotypes and observed sequence data. We assume that for individual i, the phenotype Y_i depends on the observed sequencing data D_{ij}, through unobserved genotype G_{ij} (G_{ij} is coded 0, 1 and 2) at locus j. The corresponding joint likelihood can be written as

$$P(\mathrm{Y} = (Y_{1,\dots,}\ Y_n), \mathrm{D} = (D_{1j,\dots,}\ D_{nj})) \\ = \prod_{i=1}^{n}\left(\sum_{g=0}^{2} P(Y_i|G_{ij} = g)P(G_{ij} = g, D_{ij})\right).$$

With $\mathrm{logit}(P(Y_i|G_{ij} = g)) = \beta_o + \beta_1 g$, the score statistic for β_1 is $S_j = \sum_{i=1}^{n}(Y_i - \overline{Y})E(G_{ij}|D_{ij})$ and the corresponding score test statistic $T_j = \frac{S_j^2}{var(S_j)}$ is constructed. Calculation of the expected value of genotype G_{ij} given the sequence data D_{ij} is given by $E(G_{ij}|D_{ij}) = \sum_{g=0}^{2} gP(G_{ij} = g|D_{ij})$, which requires estimation of the conditional probability $P(G_{ij} = g|D_{ij}) = \frac{P(D_{ij}|G_{ij}=g)P(G_{ij}=g)}{P(D_{ij})}$. This probability is calculated from genotype likelihood probabilities $P(D_{ij}|G_{ij}= g)$ and genotype frequencies. The conditional probabilities $P(D_{ij}|G_{ij}= g)$ are provided in the output of standard genotype calling packages (DePristo *et al.*, 2011; Li *et al.*, 2009), such as the variant calling format (VCF) files. They can also be calculated from the aligned reads by applying the simple Bayesian genotyper (McKenna *et al.*, 2010). Genotype frequencies $P(G_{ij} = g)$ are calculated from the full sample by the EM algorithm (McKenna *et al.*, 2010; Skotte *et al.*, 2012).

To calculate the test statistics, we also need to calculate the variance of S_j and, therefore, the variance of $E(G_{ij}|D_{ij})$. The expected value of the score statistic is 0 under the null hypothesis because the mean of $E(G_{ij}|D_{ij})$ is equal for cases and controls when trait Y_i and genotype G_{ij} are independent. The law of total variance is defined as $Var(G_{ij}) = Var(E(G_{ij}|D_{ij})) + E(Var(G_{ij}|D_{ij}))$. The conditional expected value $E(G_{ij}|D_{ij}) = \sum_{g=0}^{2} gP(G_{ij} = g|D_{ij})$ converges to the true value of genotype G_{ij} with high read depth (HRD) because $P(G_{ij} = g|D_{ij})$ goes to 1 for true G_{ij}. Therefore, $Var(E(G_{ij}|D_{ij}))$ is converging to the $Var(G_{ij})$. This can also be seen from $E(Var(G_{ij}|D_{ij}))$ where, in HRD data, it converges to 0 by the mathematical properties of consistency and conditional expectation (see Supplementary Fig. S1). When read depth is not sufficiently high, the second term $E(Var(G_{ij}|D_{ij}))$ is >0 and the first term $Var(E(G_{ij}|D_{ij}))$ is smaller than $Var(G_{ij})$. Therefore, $Var(E(G_{ij}|D_{ij}))$ is read depth-dependent and in low read depth data, variance of $E(G_{ij}|D_{ij})$ is smaller than the variance of the true genotype.

If cases and controls do not have systematic differences in the read depth at a given locus, the usual estimate of the variance for S_j, derived from logistic regression, can be used. However, if there is a difference in the variances of $E(G_{ij}|D_{ij})$ between two groups because of differences in read depth, the variance estimate of S_j is biased. The bias depends on the number of samples in the LRD and HRD groups and the difference in variances between the two groups. For example, for $N_{control} >> N_{case}$, the variance of the score statistic is underestimated, whereas if $N_{case} >> N_{control}$, the variance of the score statistic is overestimated. As a consequence, variance estimation of the score statistics must distinguish between the two groups.

We propose to estimate variance of the conditional expectation for cases and controls separately, as we derive in Appendix A. Briefly, to achieve the variance robustness when the number of cases is smaller than the number of controls, we propose to estimate $Var_{case}(E(G_{ij}|D_{ij}))$ by $\widehat{V}ar(G_{ij})$ with estimated genotype frequencies $P(G_{ij} = g)$, and we estimate the variance of the conditional expectation for controls by the sample variance of $E(G_{ij}|D_{ij})$ in the controls (see details in Appendix A).

Similarly, we use the score statistic $S_j = \sum_{i=1}^{n}(Y_i - \overline{Y})E(G_{ij}|D_{ij})$ to construct the test statistics for jointly analyzing several rare variants using standard published approaches (Basu and Pan, 2011; Lee *et al.*, 2012; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Neale *et al.*, 2011; Wu *et al.*, 2011), which are in general comparable (Derkach *et al.*, 2012). Similar to the single variant analysis, we can estimate the variance of $S = (S_{1,\dots,}\ S_J)$, defined as the score statistic for J rare variants. In this case, we combine the covariance matrices estimated separately for cases and controls by the same principle as in single variant analysis.

For common variants, P-values can be computed using the asymptotic distribution of the score test statistic, which is chi-square with 1 degree of freedom. For rare variants, the asymptotic chi-square approximation to the distribution of the score statistic $T_j = \frac{S_j^2}{var(S_j)}$ is often inadequate, and a permutation procedure is preferred (Basu and Pan, 2011; Derkach *et al.*, 2012). However, permutation cannot be used when the observed data consist of an external control group because the distribution of $E(G_{ij}|D_{ij})$ will then depend on read depth (Appendix A). Instead, we could calculate P-values using the bootstrap, where we sample centered values $E(G_{ij}|D_{ij}) - \overline{E(G_{ij}|D_{ij})}$ with replacement, separately for cases and controls. For joint rare variant analysis, instead of sampling $E(G_{ij}|D_{ij})$ for a single variant, we sample with replacement a centered vector of values $(E(G_{i1}|D_{i1}) - \overline{E(G_{i1}|D_{i1})}, \dots, E(G_{iJ}|D_{iJ}) - \overline{E(G_{ij}|D_{iJ})})$ for the case and control groups separately. P-value computation is based on 10 000 replicates. This non-parametric approach is also used to test the equality of two sample means without assuming equality of the distributions (Hall and Hart, 1990).

2.2 Simulation methodology

We simulated sequence reads based on the simple Bayesian genotyper as described in Appendix B. We assume all cases are sequenced at HRD and all controls are sequenced at LRD. In cases, read depth is simulated using a normal distribution with a mean of 100 and an SD of 10. In the control group, read depth is normally distributed with a mean of 4 and an SD of 1 (minimum read depth is set to 1). For each individual's locus, we generate a 'true' genotype, and then for that genotype, we generate sequence reads. The distribution of the reads given the genotype is described in Appendix B. Measurement error is present in the reads with the k$^{\mathrm{th}}$ error $\mathrm{e_k}$ following a normal distribution with a mean of 0.01 and an SD of 0.025. Then, from the generated sequenced reads, the simple Bayesian genotyper is used to call the genotype. The likelihood equations are also constructed from these sequence reads, and then we implement the RVS to evaluate Type I error and power.

For the simulations under the null model, we generate genotypes for each individual's locus based on the same MAF regardless of case or control status. For the simulations under the alternative model, we calculate MAF at the specific locus for the case and control groups separately. Specifically, we assume that $P(Y_1 = 1|X_{ij}) = exp(\beta_0 + \beta_1 X_{ij})/(1 + exp(\beta_0 + \beta_1 X_{ij}))$, where, for example, $\beta_0 = log(0.1/(1 - 0.1))$ and $\beta_1 = log(1.5)$, and these values correspond to $P(Y_i = 1|X_{ij} = 0)$ equal to 0.1 and the odds ratio (OR) for a causal variant equal to 1.5.

2.2.1 Simulation parameters and association analysis Simulations vary as a function of sample size, case control ratios and MAF. Specific scenarios are provided in Table 1. To understand the effect of differential read depth for fixed MAF using genotype calls, we simulate 1000 variants to have the same MAF, and genotype calling is done with various selection thresholds (e.g. R = 0, 0.5 or 1, see Appendix B). Association analysis using the genotype calls is conducted using the conventional score statistic (Armitage, 1955).

Single variant association analysis of the simulated data is conducted using (i) a score statistic with the *true genotypes*, (ii) the RVS method and (iii) a genotype likelihood approach that does not implement the robust variance estimate (Skotte *et al.*, 2012). We calculate P-values from the RVS using bootstrap. The corresponding P-values from analysis with true genotypes and the genotype likelihood approach are calculated via permutation. P-value computations are based on 10 000 replicates. We assess the performance of the RVS as a function of sample size (Table 1).

Table 1. Summary of simulation studies

Study design Number of cases: Number of controls	Data generated	Genetic effect	Analysis methods applied	Purpose
50:150 500:1500	1000 replicates for each MAF = 0.01, 0.1, 0.2, 0.3, 0.4 and selection threshold R = 0, 0.5, 1 combination	Under the null hypothesis	Score test with genotype calls	Assess Type I error inflation using genotype calls
500:500 500:1500	10 000 replicates for MAF = 0.01 and ranging from 0.1 to 0.5	Under the null hypothesis	(1) Score test with true genotypes, (2) RVS, (3) genotype likelihood without RVS	Comparing Type I error between the three methods in single variant analysis
50:50 50:100 50:200 500:500 500:1500	10 000 (100 000 for number of case = 50) replicates of 5 rare variants with MAF ranging from 0.001 to 0.05	Under the null hypothesis	(1) CAST and C-alpha with true genotypes, (2) RVS	Comparing Type I error between the two methods in joint rare variant analysis
500:500 500:1500	1000 replicates for MAF = 0.01 and 0.1	Under the alternative hypothesis; OR = 1.5.	(1) Score test with true genotypes , (2) RVS	Comparing empirical power of the two methods in single variant analysis
50:50 50:100 50:200 500:500 500:1500	1000 replicates of five rare variants with MAF ranging from 0.001 to 0.05	Under the alternative hypothesis; OR = 1.5	(1) CAST and C-alpha with true genotypes, (2) RVS	Comparing empirical power of the two methods in joint rare variant analysis

Note: All cases are simulated to be sequenced at HRD and all controls at LRD.

In joint rare variant analysis, we collapse five rare variants with MAF ranging from 0.001 to 0.05 and performance is evaluated across 10 000 replicates. We use the cohort allelic sums test (CAST) (Morgenthaler and Thilly, 2007) as an example of a linear statistic and C-alpha (Neale *et al.*, 2011) as an example of a quadratic statistic. We compare results from an analysis that uses the true genotypes with the RVS test that use the genotype likelihood-derived score statistic with the robust variance, as described in Section 2.1.

To assess power and Type I error, we use 1000 and 10 000 simulated replicates, respectively. For simulations under the null model, we assess the deviation of *P*-values from a uniform distribution in quantile–quantile (QQ) plots. In the analysis with genotype calls, we also assess the effect on Type I error when variants with low confidence/quality scores are filtered based on a selection threshold (see detailed description of confidence scores and selection thresholds in Appendix B). We evaluate power to detect association under various scenarios using the RVS method and compare our findings with the power we obtain using the true genotypes as the gold standard (Table 1).

2.3 NGS study data

2.3.1 Data from the 1000 Genomes Project To make comparisons between two independently sequenced samples with different enrichment strategies under the null hypothesis, we consider data from the 1000 Genomes Project using samples of European descent (CEU + GBR). Aligned reads from chromosome 11 are downloaded from the Phase 3 release [20130502]. One sample consists of exome data from 56 individuals (average read depth ~50), and the other sample includes 113 individuals who were sequenced at LRD (~6.5). A multisample VCF file was generated using the Genome Analysis Toolkit (GATK, version 2.4-9) (DePristo *et al.*, 2011; McKenna *et al.*, 2010) on the combined set of the aligned reads to identify SNPs and Indels in the samples. We excluded variants satisfying any of the following criteria: variants with phred scaled probability <30 (Qual < 30), with phred-scale strand bias *P*-value by Fisher's exact test >60 (FS > 60) and low quality depth QD < 2. Additional filtering parameters are presented in Supplementary Table S8. From the VCF files, we extract the individual genotype calls and their corresponding genotype likelihoods. To study the effect of filtering on Type I error inflation, we also analyzed datasets removing genotype calls with a quality score <5 and 10 (GQ < 5, 10). The GQ value is the phred quality score $-10log_{10}P(genetype\ call\ is\ wrong|variant)$. Therefore, GQ = 5 and GQ = 10 correspond to R = 0.5 and R = 1, as defined in Appendix B. Here we restrict analysis to biallelic loci and variants that have ≤20% missing calls, and compare association results using the genotype calls and the RVS constructed from the likelihoods supplied in the VCF file. In Supplementary Table S8, we provide the number of variants analyzed in each dataset.

2.3.2 NGS Sequencing in a Rolandic epilepsy-associated region Rolandic epilepsy (RE) is a childhood-onset epilepsy, and its electroencephalography (EEG) endophenotype is linked and associated with a 600 kb chromosomal region of 11p13 from 31 243 672 to 31 893 146 using NCBI Human Reference Assembly build 37 (Strug *et al.*, 2009). To evaluate the reliability of using the RVS with an external control group in an associated region, we compare 27 HRD RE cases with 113 LRD NGS controls from the 1000 Genomes Project.

We obtained targeted resequencing data on 27 RE patients of European descent, ascertained in the northeastern USA. The 600 kb region of chromosome 11p13 was enriched using long-range PCR. The multiplex samples were then resequenced on the Illumina GAIIX platform with an average read depth of 197× by deCODE Scientific Services (Iceland). Thirty-six base-pair end reads were aligned to a 700 kb region of 11p13 using the Novoalign algorithm within the GATK analysis pipeline (DePristo *et al.*, 2011; McKenna *et al.*, 2010). Here we compare the RE cases with the 113 controls of European descent from the 1000 Genomes Project, Phase 3 release [20130502]. A multisample VCF file was produced using GATK's Unified Genotyper module on the combined set of aligned reads to identify variant sites in the samples. Only biallelic variants are analyzed here. Additional filtering parameters and the number of variants categorized by MAF threshold 0.05 are in Supplementary Table S9. Given there are only 27 RE cases sequenced, we focus on variants with MAF >0.05. We then selected the 491 SNPs that have ≤20% missing calls and have estimated MAF >0.05. Likelihood information from the VCF file was used to implement the RVS method to compare the two groups.

For comparison, we assess the association evidence using genotype calls from the 27 epilepsy cases and an independent sample of 200 colorectal cancer cases from the UK, whole-genome sequenced by Complete Genomics. Sequencing was performed at high coverage (70% of the genome with an average 35×) using unchained combinatorial probe anchor ligation chemistry on arrays of self-assembling SNA nanoballs (Drmanac *et al.*, 2010). In this case we did not have access to locus-specific coverage information. We investigate the variant call set in a sample spanning the 600 kb region of interest at 11p13. In this dataset we have 453 variant sites with ≤20% missing calls and estimated MAF >0.05 available for analysis.

3 RESULTS

3.1 Theoretical and empirical investigations of the effect of read depth and selection threshold differences

To guide our theoretical investigations, without loss of generality, we assign A and C to be major and minor alleles, respectively. For simplicity assume that all reads are sequenced without errors ($e_k = 0$ for $k = 1, ..., r_{ij}$). Under this setting, if r_{ij} reads consist of minor and major allele calls, then the genotype call is always set to AC. We assume that the true genotype at the locus is CC and all r_{ij} reads consist of the minor allele C. From Appendix B, the likelihoods for the genotypes AA, AC and CC are $L(D_{ij}|AA) = 0$, $L(D_{ij}|AC) = (1/2)^{r_{ij}}$ and $L(D_{ij}|CC) = 1$, where D_{ij} consists of r_{ij} reads all equal to C. The posterior probabilities are equal to $P(AA|D_{ij}) = 0$, $P(AC|D_{ij}) = (1/2)^{r_{ij}} \cdot P(AC)/P(D_{ij})$ and $P(CC|D_{ij}) = P(CC)/P(D_{ij})$. By ignoring $P(D_{ij})$ and assuming Hardy-Weinberg equilibrium (HWE) with $P(C) = p_j$ and $P(A) = q_j$, we rewrite the posterior probabilities as $P(AA|D_{ij}) = 0$, $P(AC|D_{ij}) = (1/2)^{r_{ij}} 2p_j q_j$ and $P(CC|D_{ij}) = p_j^2$. The genotype call is chosen to be CC if $P(CC|D_{ij}) = p_j^2 > P(AC|D_{ij}) = 2p_j q_j (1/2)^{r_{ij}}$, and it is miscalled as AC otherwise. This implies that for a variant with read depth r_{ij} and a specific MAF $p_j < \frac{1}{1+2^{r_{ij}-1}}$, the rare homozygous genotype can be misclassified as a heterozygote. Naturally, this kind of misclassification leads to underestimated MAFs, and the degree of bias decreases with MAF as the number of rare homozygous genotypes decreases. For a given read depth, Table 2 provides the MAF threshold below which rare homozygous genotypes are misclassified.

The degree of bias in the estimation of MAF is also related to the selection threshold R, which screens out calls with low confidence scores. With a selection threshold R = 0, estimation of MAF is affected mainly by misclassified rare homozygous calls. However, with R > 0, those misclassified calls are often filtered out because of low confidence/quality scores. This suggests the estimated MAF would be further underestimated because of screening out some misclassified calls and some weak true calls. Similarly, sequencing error, which is not modeled in this theoretical investigation, would further affect estimation of MAF because posterior probabilities for genotypes with rare alleles are affected. Our empirical investigations do incorporate sequencing error (Section 3.2).

Table 2. Relationship between read depth and critical MAF threshold

Read depth	2	3	4	5	30	100
Minimum MAF	0.33	0.20	0.11	0.059	$1.8{\cdot}10^{-9}$	$1.6{\cdot}10^{-30}$

Note: For MAF below the listed value, rare homozygous genotypes are miscalled as heterozygotes. Selection threshold R = 0.

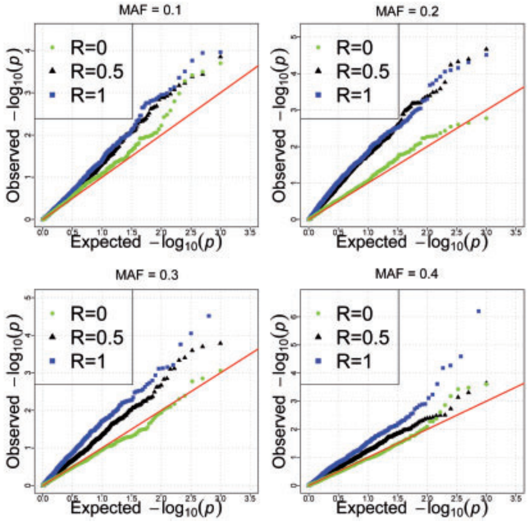


Fig. 2. QQ plot for *P*-values from an association study with 50 HRD cases and 150 LRD controls, as a function of MAF. *P*-values are calculated by the score statistic on 1000 variants using called genotypes. *R* is the selection threshold

These theoretical findings have particular implications when case and control samples differ systematically: (i) If cases and controls are sequenced with the same read depth but two different selection thresholds are applied, spurious results can occur for some MAFs. (ii) If both cases and controls are sequenced with different read depths but the same selection threshold, R > 0, is applied, spurious results can also occur for some MAFs. (With R = 0, for variants with MAF larger than the critical MAF provided in Table 2, the analysis is unlikely to produce spurious results.) (iii) For a given read depth, at variants with MAF below the threshold (Table 2), spurious association results are more likely for all R because of bias in MAF estimation.

Our simulation study confirms our theoretical conclusions and further investigates the relationship between read depth r_{ij}, selection threshold R and Type I error. As the theoretical derivations

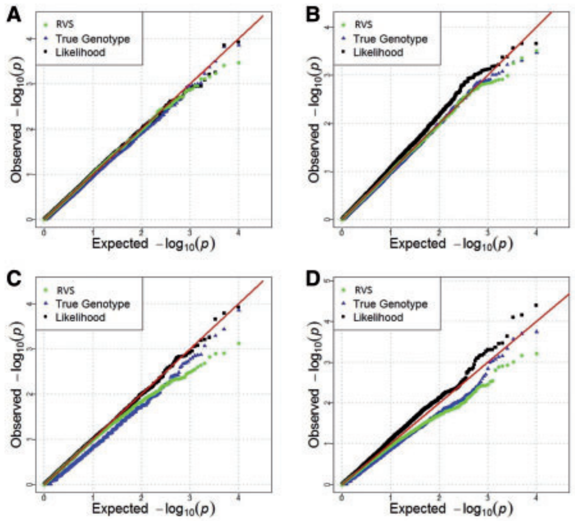


Fig. 3. QQ plots for *P*-values from RVS for 10000 variants with MAF equal to 0.1 (**A** and **B**) and 0.01 (**C** and **D**). True genotype analysis uses the score statistic; Likelihood uses genotype likelihoods without the robust variance. Plots (A and C), 500 cases and 500 controls; Plots (B and D) 500 cases and 1500 controls. Sequencing error is set to 0.01

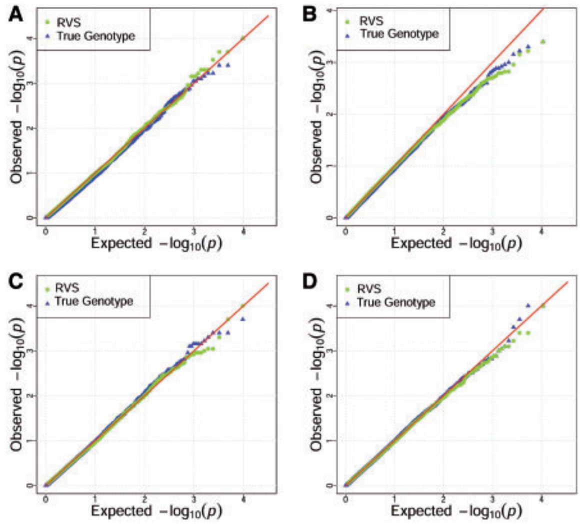


Fig. 4. QQ plots for *P*-values from rare variant analysis using the CAST linear statistic (plots **A** and **B**) and C-alpha quadratic statistic (plots **C** and **D**). Five rare variants are grouped together with MAF ranging from 0.001 to 0.05. Analysis with true genotypes uses the score statistic. Plots (A and C), 500 cases, 500 controls; Plots (B and D) 500 cases, 1500 controls. Sequencing error is set to 0.01

predicted, deviation in QQ plots or the extent of bias in the estimation of a given variant's MAF depends largely on the MAF (Fig. 2). The main contributor to the deviation in the QQ plots is that the rare homozygotes are in most cases screened out (R = 1 and 0.5) or miscalled (R = 0) in the LRD control sample. A deviation of *P*-values from the expected uniform distribution is less apparent when MAF is large (e.g. Fig. 2; MAF = 0.4); this is because of less misspecification/filtering for large MAF (as seen in Table 2).

To confirm that a deviation in *P*-values from uniform exists in the absence of filtering (R = 0), we considered analysis in larger samples. Analysis of simulated replicates with 500 HRD cases and 1500 LRD controls confirms our theoretical prediction (Supplementary Fig. S2). Larger deviations from expected are observed in larger sample sizes where the differences in estimated MAF between cases and controls are amplified and the greatest bias occurs when the MAF is ~0.2.

3.2 Empirical investigation of RVS

We begin our empirical investigation of the properties of the RVS approach by showing that it controls Type I error under a variety of settings. As predicted, the *P*-values from the genotype likelihood approach (Skotte *et al.*, 2012) are inflated when the number of LRD controls is larger than the number of cases (Fig. 3B and D). This inflation increases as the case control ratio decreases. In contrast, *P*-values from the RVS using the bootstrap are not affected because it uses the robust variance estimate (Fig. 3, Supplementary Fig. S3 and Supplementary Table S1). For example, for a test size of 0.01, the empirical Type 1 error for RVS is 0.0094. The Type I error is slightly conservative for the analysis of single rare variants using RVS, but this is as expected and remains the case even when the true genotypes are analyzed (Supplementary Fig. S3A and Supplementary Table S2). This indicates that both permutation and bootstrap approaches are conservative when there is sparsity in the data. A similar result is observed when asymptotic distributions are used to compute *P*-values (Supplementary Fig. S4). In contrast to single rare variant analysis, when five rare variants are grouped for analysis, Type I errors resulting from the bootstrap and permutation approaches are well controlled (Fig. 4 and Supplementary Tables S3 and S4). Investigations with smaller sample sizes magnify these observations (Supplementary Figs. S5 and S6 and Supplementary Table S5).

Table 3. Empirical power of the RVS for single common variants

Type of analysis	Sample size (case:control)	Level of the test			
		0.05	10^{-2}	10^{-3}	10^{-4}
RVS	500:500	0.81	0.62	0.32	0.13
True genotypes		0.83	0.62	0.34	0.15
RVS	500:1500	0.91	0.80	0.53	0.29
True genotypes		0.95	0.84	0.62	0.41

Note: All cases are simulated to be sequenced at HRD and all controls at LRD. Results are based on 1000 replicates. The variant has MAF equal to 0.1 and OR equal to 1.5. Empirical power for analysis with the true genotypes is provided for comparison. Sequencing error is set to 0.01.

For common variants, the RVS method is comparable in power with the score test applied to the true genotypes (Table 3). When a single rare variant is considered, both the score test applied to the true genotypes and the RVS have substantially less power than joint analysis (Supplementary Table S6). When we jointly analyze five rare variants by CAST

Table 4. Empirical power of the RVS for joint rare variant analysis using a linear statistic (CAST) and quadratic statistic (C-alpha)

Method	Type of analysis	Sample size (case:control)	Level of the test			
			0.05	10^{-2}	10^{-3}	10^{-4}
CAST	RVS	500:500	0.86	0.69	0.47	0.26
	True genotypes		0.89	0.74	0.50	0.29
C-alpha	RVS	500:500	0.71	0.49	0.26	0.12
	True genotypes		0.74	0.53	0.28	0.12
CAST	RVS	500:1500	0.96	0.89	0.72	0.51
	True genotypes		0.97	0.92	0.79	0.61
C-alpha	RVS	500:1500	0.89	0.76	0.55	0.35
	True genotypes		0.92	0.80	0.61	0.40

Note: All cases are simulated to be sequenced at HRD and all controls at low read depth. Results based on 1000 replicates. For each replicate, 5 variants with MAF raging form 0.001 to 0.05 and OR equal to 1.5 are grouped and analyzed by linear and quadratic statistics with RVS. Empirical power for analysis with true genotypes is also provided. Sequencing error is set to 0.01.

Table 5. Empirical power comparison between RVS with HRD cases and LRD controls, and logistic regression when both case and control groups have equal average LRD (4×)

Type of analysis	Sample size (case:control)	Level of the test			
		0.05	10^{-2}	10^{-3}	10^{-4}
Logistic regression	500:500	0.73	0.51	0.25	0.1
RVS		0.81	0.62	0.32	0.13
Logistic regression	500:1500	0.89	0.75	0.51	0.29
RVS		0.91	0.80	0.53	0.29

Note: Results are based on 1000 replicates. The variant has MAF = 0.1 and OR = 1.5. Empirical power for association using RVS is provided for comparison.

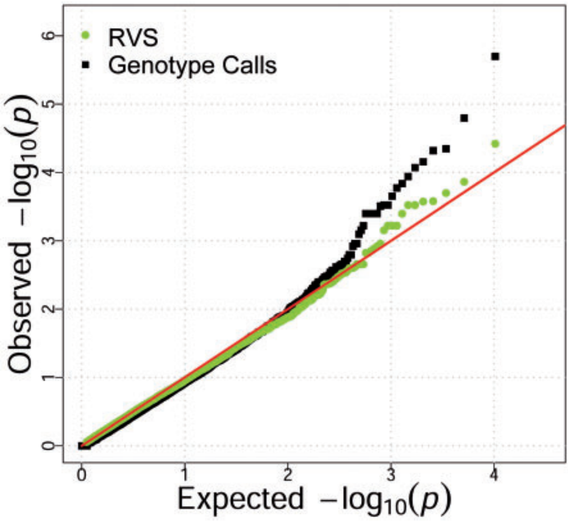


Fig. 5. QQ plots for *P*-values obtained from the RVS and the score statistic. Analysis based on 56 HRD cases and 113 LRD controls from the 1000 Genomes Phase 3 release [20130502]. *P*-value based on 10^5 replicates. We use 10 269 SNPs that have MAF >0.05, missing rate smaller than 20% and are present in both datasets

and the quadratic test C-alpha, a noticeable improvement in power is apparent (Table 4). Not surprisingly, results from single and joint rare variant analysis indicate that analysis using the RVS is not as powerful as the score statistic with the true genotypes, when the number of controls is larger than the number of cases. This is because genotype frequencies used in the posterior are estimated from the observed case control data, and the accuracy of these estimates is affected when read depth is low. We also note that the RVS method, in most cases, has similar power to an analysis with the true genotypes with test sizes between 0.05 and 0.0001.

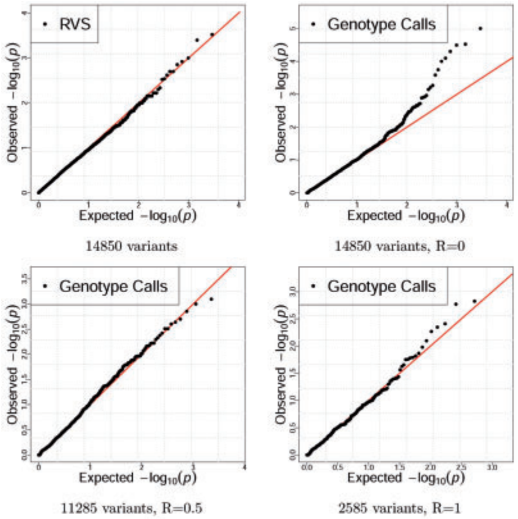
We also consider an empirical power comparison between the RVS method with HRD cases and LRD controls, and conventional logistic regression where both groups have the same average LRD (4×). We consider this analysis to illustrate that downsampling HRD cases to the same average read depth as controls reduces power. Results in Table 5 (500 cases and 500/1500 controls at MAF = 0.1) and Supplementary Table S7 (500 cases and 500/1500 controls at MAF = 0.01) indicate that using RVS in studies with HRD cases is a more powerful strategy than using logistic regression in studies with LRD in both case and control groups.

3.3 Application of the RVS method

3.3.1 1000 Genomes Project data We used two publicly available datasets from the 1000 Genomes Project: 56 HRD exome sequencing and 113 LRD whole genome sequencing at 11p13 to assess the Type I error inflation using RVS under the null hypothesis. We compare results for single SNP analysis of common variants (MAF > 0.05) and rare variant analysis using groups of five rare variants (MAF < 0.05).

A comparison of *P*-values from common variants using the RVS method with those from the score statistic using genotype calls that are not filtered (R = 0) indicates that the RVS controls Type I error well (Fig. 5). Results from these two analyses are similar, although analysis with unfiltered genotype calls results in several false-positive variants. These spuriously associated variants would have been filtered out had genotype quality filters been applied. However, as we observed in our simulations and our theoretical findings, filtering also results in increasing Type I error inflation, as a function of more stringent quality thresholds. Filtering can also significantly reduce the number of variants for analysis, which would reduce power. For example, with GQ = 10 (R = 1), we analyze only 1760 common SNPs (see Supplementary Fig. S7). The RVS does not require any filtering and analyzes the observed data as is.

It should be noted that the inflation in Type I error is only marginal in Figure 5, consistent with our theoretical predictions. There is a reasonably high average read depth for the 10 269 variants analyzed in the 'LRD' control group (7.5× versus 71× for the HRD control group; Supplementary Fig. S8 and Supplementary Table S8). With a read depth of 8, the theoretical



14850 variants

14850 variants, R=0

11285 variants, R=0.5

2585 variants, R=1

Fig. 6. QQ plots for joint rare variant analysis under the null model: with the CAST statistic using the RVS, and a score test using genotype calls with selection threshold R = 0, 0.5 and 1, respectively. Analysis based on 56 HRD cases and 113 LRD controls from the 1000 Genomes Project Phase 3 release [20130502]. *P*-value results are based on 10^5 replicates. Analysis combines five rare variants (MAF < 0.05, missing rate smaller than 20%)

Table 6. A comparison of the variants' rankings from a score test using genotype calls with 27 RE cases and HRD Complete Genomics controls (n = 200), versus the rankings from RE cases and 1000 Genomes LRD controls (n = 113)

Name	*P*-value (rank) based on HRD cases and HRD controls; genotype calls	*P*-value (rank) based on HRD cases and LRD controls; RVS
rs6484504	0.00008 (1)	0.0010 (3)
rs578666	0.00012 (2)	0.0001 (1)
rs674035	0.0007 (3)	0.0004 (2)
rs11031375	0.003 (4)	0.009 (7)
rs662702	0.012 (5)	0.0523 (191)
rs11031330	0.011 (7)	0.0018 (4)
rs603202	NA	0.0027 (5)

calculations indicate that it is unlikely to have a biased estimate of MAF for variants with MAF >0.007. Individual variants within these 10 269 that are sequenced with lower read depth than 7× contribute to the slightly inflated Type I error, highlighting the important point that average read depth is not sufficient to assume unbiased estimation of MAF at all loci when using genotype calls for analysis.

Figure 6 compares the results for joint rare variant analysis using the linear statistic (CAST) between (i) the RVS method and (ii) the score statistic using genotype calls with filtering thresholds as R = 0, R = 0.5 and R = 1. The RVS approach controls the Type I error well, while *P*-values are inflated in all three scenarios with genotype calls. The filtering reduces the overall inflation in Type I error, at the expense, however, of significantly reducing the number of rare variants that can be analyzed, from 14 850 for RVS and R = 0, to 11 285 and 2585 for R = 0.5, 1, respectively, suggesting filtering can negatively impact power. Results from analysis with quadratic statistics are similar (Supplementary Fig. S9).

3.3.2 Association in targeted resequencing of individuals with Rolandic epilepsy Last, we are interested in how the RVS performs in a region of association with the RE EEG endophenotype. We identify variants from the aligned reads of our 27 RE cases and 113 LRD whole genome-sequenced individuals from the 1000 Genomes Project. In addition, we have access to genotype calls in 200 HRD controls sequenced by Complete Genomics. A summary of the number of variants analyzed with our cases and each control group indicates similar identification of variants with MAF >0.05 (Supplementary Table S9). We compare the top-ranked variants by the RVS method using the 1000 Genomes Project LRD control group, with the top rankings based on an analysis that implements the conventional score statistic using genotype calls with the 200 HRD controls sequenced by Complete Genomics. Table 6 indicates that the top-ranked variants are similar across the two analyses; that is, using the RVS with LRD controls indicates similar prioritization of variants for follow-up to an analysis with genotype calls from two HRD sequenced samples in an associated region. The LRD group is smaller than the HRD control group, which may explain why the *P*-values from the RVS are slightly larger.

4 DISCUSSION

Publicly available genome-wide microarray datasets have been widely used as controls in genome-wide association studies (The Wellcome Trust Case Control Consortium, 2007). Here we provide a new method, the RVS, to test for association in studies that use NGS from external control groups. Confounding factors associated with NGS data processing, such as SNP and genotype calling algorithms, read depth and selection parameters can all contribute to spurious or masked findings. Here we focus on statistical adjustment for the bias in MAF estimation (and consequently Type I error) introduced by differential read depth between cases and controls, and the selection threshold. In the absence of a unified study design for sequencing cases and controls, using theoretical and empirical investigations, we show that the RVS is a useful tool to incorporate external control groups in genetic association studies with NGS data, in an effort to prioritize sequence variants for follow-up.

The RVS can be used for single variant or joint rare variant analysis, and does not require arbitrary selection of parameter values for filtering but rather analyzes all observed data. The Type I error associated with the RVS is well controlled by the use of robust variance estimates, and the power is comparable to analyses using genotypes called without error.

Our theoretical and simulation results indicate that systematic differences between cases and controls lead to spurious results in association analysis using genotype calls from NGS technology. The degree of deviation in *P*-value distribution from expected under the null hypothesis depends on MAF, difference in read depth and the applied selection threshold. Particularly, when $R = 0$, genotype call-based association analyses can be applied for some variants even when there is a significant difference in read depth between cases and controls (Fig. 2 and Table 2). Table 2 provides lower bounds on MAF at which point MAF estimates from low read depth data remain close to the true value. However, using these theoretical predictions to justify analysis at certain MAFs would perforce preclude analysis at the remaining variants because the biased allele frequency estimation for the remaining data would remain unaccounted for. A HRD control group with ample coverage to allow a selection threshold (i.e. $R > 0$) can be used to avoid this bias; however, this does not ensure that there is ample read depth in cases and controls at every locus.

To address systematic difference in read depth, the GATK toolkit (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_PrintReads.html) proposes to randomly downsample BAM files for the higher read depth group. This approach is a less powerful strategy in comparison to methodology that incorporates all observed data as we show in Section 3.2. Other methods that use logistic regression analysis with read depth as a covariate, or by weighting each variant call by quality score (Daye *et al*., 2012; Garner, 2011), as well as methods that substitute genotype calls by their expected values (Skotte *et al*., 2012) are only applicable if both groups are not distinguishable by read depth and sequencing error.

We implement the RVS in a case–control study with 27 HRD cases and 113 LRD controls. The top-ranked variants are in agreement with an association study based on genotype calls from the 27 HRD cases and 200 HRD controls. However, as was the case in the present study, if one has access only to genotype calls, as opposed to the aligned BAM files or the raw data, only variants with genotype calls present in both datasets can be analyzed; this could potentially be restrictive. For example, rs603202 in Table 6 was sequenced in the 1000 Genomes Project controls and epilepsy cases, but calls at this SNP in the Complete Genomics control group are absent. Without additional information about locus-specific coverage in the Complete Genomics control set, we are not able to determine whether the missing variant is monomorphic in the sample or there was simply no sequencing coverage. Coverage information is integrated into the RVS analysis of the aligned reads.

Currently the RVS cannot accommodate covariate adjustment, and this will be an area of future development. The RVS is easily extendable to accommodate a design in which a subset of controls is sequenced alongside the cases in addition to incorporating an external control group. This type of study design may prove preferable to assess the comparability of the external control group; this more costly approach, however, requires a formal evaluation. Ensuring the comparability of the case and control groups on the basis of epidemiologic principles is paramount. This assumption requires careful consideration before moving forward with any statistical analysis.

Ideally, cases and controls would also be matched on sequencing platform and enrichment strategy; however, our results indicate that this is not necessary because the RVS adjusts for differences in variability and missing rates because of platform/enrichment differences. We suggest that when cases and controls are sequenced using different technologies, conducting association analysis with a second control group can provide confidence that systematic bias due to platform/enrichment differences is not driving the observed signals. Different alignment algorithms are implicitly accounted for by the RVS because the unit of analysis is the genotype probability rather than the genotype calls in the association analysis.

Whole genome sequencing of large samples remains cost prohibitive for many investigators. Using external control groups in NGS association studies, to augment a smaller set of sequenced controls or as the only control set for comparison, can reserve precious resources for the sequencing of cases. Therefore, if NGS service providers (or other public initiatives) make control samples available to customers (or the public), then the RVS makes it feasible to use external control groups in association studies with NGS data.

ACKNOWLEDGEMENT

The authors thank Drs Lei Sun and Jerry Lawless for helpful discussions.

Funding: This work has been supported by an Ontario Graduate Scholarship (OGS) and the CIHR Strategic Training for Advanced Genetic Epidemiology (STAGE) to A.D; Natural Sciences and Engineering Research Council, the Ontario Ministry of Research and Innovation early researcher award program and the Canadian Institutes of Health Research (MOP-258916) grants to L.J.S; European Union Marie Curie International Reintegration Award of the Seventh Framework Programme (DKP); Waterloo Foundation (DKP); Epilepsy Research UK (DKP); Charles Sykes Epilepsy Research Trust (DKP); NIHR Specialist Biomedical Research Centre for Mental Health of South London and Maudsley NHS Foundation Trust (DKP). Cancer Research UK, Bobby Moore Fund for Cancer Research UK (RSH, IPT), Research Network and the NHS via the Biological Research Center of the NIHR at the Royal Marsden Hospital NHS Trust (RSH); Core infrastructure support to the Welcome Trust Center for Human Genetics (IPT).

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. *et al*. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

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

Basu,S. and Pan,W. (2011) Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol*., **35**, 606–619.

Daye,Z.J. *et al*. (2012) A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res*., **40**, e60.

DePristo,M.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*., **43**, 491–498.

Derkach,A. *et al*. (2012) Pooled association tests for rare genetic variants: a review and some new results. Statistical Science. To appear.

Drmanac,R. *et al*. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.

Garner,C. (2011) Confounded by sequencing depth in association studies of rare alleles. *Genet. Epidemiol*., **35**, 261–268.

Hall,P. and Hart,J.D. (1990) Bootstrap test for difference between means in nonparametric regression. *J. Am. Stat. Assoc*., **85**, 1039–1049.

Kim,S.Y. *et al*. (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, **12**, 231.

Lee,S. *et al*. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu,D.J. and Leal,S.M. (2012) SEQCHIP: a powerful method to integrate sequence and genotype data for the detection of rare variant associations. *Bioinformatics*, **28**, 1745–1751.

Longmate,J.A. *et al*. (2010) Three ways of combining genotyping and resequencing in case-control association studies. *PLoS One*, **5**, e14318.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*., **5**, e1000384.

McKenna,A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*. *Res*., **20**, 1297–1303.

Morgenthaler,S. and Thilly,W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat*. *Res*., **615**, 28–56.

Neale,B. *et al*. (2011) Testing for an unusual distribution of rare variants. *PLoS*. *Genet*., **7**, e1001322.

Nielsen,R. *et al*. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat*. *Rev*. *Genet*., **12**, 443–451.

Sanna,S. *et al*. (2011) Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet*., **7**, e1002198.

Sebastiani,P. *et al*. (2011) Retraction. *Science*, **333**, 404.

Skotte,L. *et al*. (2012) Association testing for next-generation sequencing data using score statistics. *Genet*. *Epidemiol*., **36**, 430–437.

Strug,L.J. *et al*. (2009) Centrotemporal sharp wave EEG trait in rolandic epilepsy maps to Elongator Protein Complex 4 (ELP4). *Eur*. *J*. *Hum*. *Genet*., **17**, 1171–1181.

The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Wu,M.C. *et al*. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am*. *J*. *Hum*. *Genet*., **89**, 82–93.