ARIEL and AMELIA: Testing for an Accumulation of Rare Variants Using Next-Generation Sequencing Data

Jennifer L. Asimit^a Aaron G. Day-Williams^a Andrew P. Morris^b
Eleftheria Zeggini^a

^aWellcome Trust Sanger Institute, Hinxton, and ^bWellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Key Words

Whole-genome sequencing • Exome sequencing • Association analysis • Accounting for uncertainty • Complex trait

Abstract

Objectives: There is increasing evidence that rare variants play a role in some complex traits, but their analysis is not straightforward. Locus-based tests become necessary due to low power in rare variant single-point association analyses. In addition, variant quality scores are available for sequencing data, but are rarely taken into account. Here, we propose two locus-based methods that incorporate variant quality scores: a regression-based collapsing approach and an allele-matching method. **Methods:** Using simulated sequencing data we compare 4 locus-based tests of trait association under different scenarios of data quality. We test two collapsing-based approaches and two allele-matching-based approaches, taking into account variant quality scores and ignoring variant quality scores. We implement the collapsing and allele-matching approaches accounting for variant quality in the freely available ARIEL and AMELIA software. **Results:** The incorporation of variant quality scores in locus-based association tests has power advantages over

weighting each variant equally. The allele-matching methods are robust to the presence of both protective and risk variants in a locus, while collapsing methods exhibit a dramatic loss of power in this scenario. **Conclusions:** The incorporation of variant quality scores should be a standard protocol when performing locus-based association analysis on sequencing data. The ARIEL and AMELIA software implement collapsing and allele-matching locus association analysis methods, respectively, that allow the incorporation of variant quality scores.

Copyright © 2012 S. Karger AG, Basel

Introduction

Rare variants have been proven to play a role in some complex traits [1–3], but their contribution has not been studied as extensively as for common SNPs. This is partly due to the current cost limitations in next-generation sequencing technologies, the lack of a rare variant catalog with reference genotypes (poised to change with the 1,000 Genomes and UK10K Projects), and the lack of an appropriate analytical toolbox to enable powerful rare variant association analysis given currently available sample sizes. Following the usual convention, we define

low-frequency variants to be those with a minor allele frequency (MAF) of 1–5%, and those with MAF <1% as rare variants.

The need for locus-based methods in low-frequency/rare variant analyses has become clear as single-variant tests have been demonstrated to be underpowered to detect associations. It has been shown that the power of single-variant tests is greatly affected by allele frequency and is very sensitive to the effect size [4, 5]. The regions for analysis via locus-based methods may be genes or other functional units, e.g. regulatory regions.

Collapsing (or aggregate) methods have become a popular locus-based approach to low-frequency/rare variant analyses and numerous strategies have been proposed [4, 6–12]. However, all of these methods, with the exception of RARECOVER by Bhatia et al. [11] either assume that all variants within the functional unit analysed have the same direction of effect (i.e. protective or risk-causing), or require a test set/testing stage to predict the direction of effect. A replication-based strategy has been recently introduced as a method that is less sensitive to the presence of variants with opposite directions of effect [13], while the C-alpha test is a novel method that tests for the presence of a mixture of effect directions among rare variants in a locus [14].

An allele-matching approach proposed within a kernel-based association test (KBAT; [15]) is an alternative method that may be applied to variants of all frequencies and does not rely on any information regarding the direction of effect.

Both the collapsing method and allele-matching tests combine information across multiple variant sites within the region for analysis, but using different approaches. In the collapsing method information across multiple variants is collapsed into a single quantity which is used to test for an association with an accumulation of low-frequency/rare minor alleles. An implicit assumption of this collapsing approach is that the causal variants have the same direction of effect, for example in a case-control scenario all are either protective or risk-increasing. In contrast, the allele-matching test is a multi-marker test that combines single-variant test statistics that do not rely on knowledge of the risk alleles.

We focus on a collapsing method under a regression framework [7] which is applicable to both case-control data and quantitative traits. This approach applies a single univariate test to the collapsed data within a group, resulting in enriched signals and fewer degrees of freedom, rather than facing multiple correction factors for many single-marker tests, or high degrees of freedom in

a multiple-marker test [4]. There are various versions of the collapsing method according to the approach used to collapse the variants in a locus for each individual (presence/absence indicator of at least one minor allele at any low-frequency variant or proportion of low-frequency variants with at least one minor allele) and the chosen statistical test [4, 7, 16]. Our focus is on a regression framework incorporating proportions of variants that carry low-frequency minor alleles. Although this formulation is not robust to linkage disequilibrium (i.e. correlation between the collapsed variants can lead to inflation of the test statistic), it has been demonstrated to be more robust to the presence of minor alleles at noncausal variants than the model that uses the presence/absence of at least one variant with a minor allele [see 7]. In addition, the regression framework is flexible to the inclusion of covariates.

In the analysis of the association of rare variants and disease, there is a loss of power due to genotype mis-specification. Quality scores are available for genotype and sequence-derived data, but this information is not routinely put to use. A common measure of variant quality is the phred-scaled quality score, which is related to the variant-calling error probabilities. In practice, variants with phred-scaled scores below 10 are filtered out and not included in analyses; this corresponds to a correct variant call probability of 0.90. Histograms of the probabilities of correct variant calls were generated for 16 random samples of 50-kb regions from chromosome 1 data of the pilot study of 1,000 Genomes (August 2009 Release, 68 CEU samples) and are provided in the online supplementary figure 1 (www.karger.com/doi/10.1159/000336982) [17]. In general, the majority of variants have a probability of a correct variant call (converted quality score) at one of the two extremes of the probability distribution: probability greater than 0.90 or probability below 0.10.

We propose methods for rare variant analyses that take advantage of the additional information contained in variant quality scores derived from sequencing. More specifically, rather than filtering out variants that have a quality score below 10 (correct variant call probability below 90%), we opt to include all variants with a positive probability of a correct variant call and weight variants according to these probabilities. These methods are extensions of the collapsing method and allele-matching test detailed above: Accumulation of Rare variants Integrated and Extended Locus-specific test (ARIEL) and Allele Matching Empirical Locus-specific Integrated Association test (AMELIA). We investigate power comparisons among the original methods and extensions to

account for variant quality scores under various scenarios of disease allelic architecture, focusing on low-frequency/rare variants. We find that ARIEL and AMELIA are more powerful than the collapsing method and KBAT, particularly when the causal variants are of high quality. In addition, the allele-matching tests display a large power increase over the collapsing approaches when the locus contains both protective and risk variants.

Methods

A logistic regression-modelling framework is employed for the collapsing methods applied to case-control data. Let p_i denote the probability that individual i has the trait, while k_i denotes the number of successfully genotyped low-frequency/rare variants for individual i , and r_i denotes the number of low-frequency/rare variants that carry at least one copy of the minor allele. Incorporating a vector of covariates for individual i , x_i , the logistic regression model is given by

$$\text{logit}(p_i) = \alpha + \lambda \frac{r_i}{k_i} + \beta x_i + \varepsilon_i \quad (1)$$

where λ is the expected increase when an individual carries at least one minor allele at any low-frequency/rare variant compared to one that has a complete absence of low-frequency/rare minor alleles and the odds ratio is given by $\exp(\lambda)$.

The collapsing method extension to account for uncertainty in the form of variant quality scores is introduced as ARIEL (Accumulation of Rare variants Integrated and Extended Locus-specific test). Noting that all of the variants in a locus do not have the same level of quality, we consider a quality-weighted proportion, where higher quality variants contribute a larger weight to the proportion of rare variants at which subject i carries a minor allele. The phred-scaled quality score Q ($0 < Q < 100$) is logarithmically related to the variant-calling error probabilities P by the relation $Q = -10 \log_{10} P$, so that the probability that a variant is correctly called is $1 - 10^{-Q/10}$. This means that a phred-scaled score above 10 is assigned to variants that have a correct variant call probability of at least 90%. For this reason we convert the phred-scaled scores to probabilities of correct variant calls, since in our methods the probabilities will more fairly reflect the quality of the variant. As a more suggestive notation we let q_{ij} be the transformed quality score for variant j of subject i , and let G_{ij} equal the number of rare alleles at variant j for subject i . For the i -th subject we then define the weighted proportion as:

$$w_i = \frac{\sum_{j=1}^K I(G_{ij} > 0) q_{ij}}{\sum_{j=1}^K q_{ij}}, \quad (2)$$

which simplifies to the original proportion when perfect quality (all $q_{ij} = 1$, all phred-scaled scores are 100) is obtained. The model becomes:

$$\text{logit}(p_i) = \alpha + \lambda w_i + \beta x_i + \varepsilon_i \quad (3)$$

and as before, estimation is done by least-squares, and a parametric p value for testing the effect $\lambda \neq 0$ is obtained. Although per-

mutation testing is not a requirement for the collapsing approaches, permutation-based p values may also be calculated.

Sequencing studies are currently carried out in two study designs: (1) high-depth sequencing of individuals and individual variant calling/genotyping and (2) population-based low-depth resequencing with population variant calling followed by individual genotyping. The first study design will result in each individual having a variant quality score for each variant calculated by the reads for just that individual, i.e. variant scores vary across both variants ($j = 1, \dots, K$) and individuals ($i = 1, \dots, n$). This scenario results in what we term non-consensus variant scores. In the case of low-read depth population resequencing data, variant quality scores are calculated from the reads of all individuals, i.e. quality scores only vary among variants ($q_{ij} = q_j$; $i = 1, \dots, n$). This second scenario results in what we term consensus variant scores. The derivation of ARIEL above is based on the analysis of non-consensus variant quality scores. In the case of consensus quality scores, the weighted proportions (2) are calculated in the same manner, with q_{ij} replaced by q_j .

For both the collapsing method and ARIEL, the odds ratio (OR) is adjusted by the number of variants in the region that are used in the analysis. More explicitly in (1) and (3), the adjusted OR is $\exp(\lambda/K)$, where K is the number of variants with MAF below the specified threshold.

The allele-matching test is an example of a kernel-based association test (KBAT) designed for case-control data, and jointly tests multiple variants (correlated or independent) without making any assumptions on the direction of individual variant effects [15]. KBAT is based on genotype similarity scores, measured by a kernel function, between individuals within the same group (e.g. cases or controls). We focus on the Allele Match (AM) kernel, which is the count of common alleles between the genotypes of two individuals. At each variant similarity scores are calculated between each pair of individuals in cases and between each pair of individuals in controls. A one-way ANOVA model is then used to test whether or not the two groups of scores have the same mean at the specific variant, which is equivalent to testing for an association between the variant and disease. That is, if the average number of shared alleles at a variant between cases is significantly different from that for controls then there is disease association with the variant. The multiple single variant tests are then combined into a single multi-marker test for an association between disease and multiple rare/low-frequency variants within a specified region.

Details of KBAT are as follows: first, similarity scores $y_{l(ij)}$ between individuals i and j in group l ($l = 1$ = cases, 2 = controls) are determined by using a kernel, such as the Allele Match (AM) kernel, which is the count of common alleles between the genotypes of two individuals. A genotype score at a variant for individual i in group l ($g_{l(i)}$) is defined as the number of minor alleles (arbitrary which allele is counted) it carries, and therefore takes values 0, 1 or 2. At a given variant, for individuals $i \neq j$ in group l with respective genotype scores $g_{l(i)}$ and $g_{l(j)}$, the AM similarity score is defined by

$$y_{l(ij)} = \begin{cases} 4, & \text{if } g_{l(i)} = g_{l(j)} \\ 2, & \text{if } g_{l(i)} = 1, g_{l(j)} \in \{0, 2\} \text{ or } g_{l(i)} = 1, g_{l(j)} \in \{0, 2\}, \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

By defining the kernel in this way, there is no need to have knowledge of the risk allele at each variant. Similarity scores that de-

pend on knowledge of the risk allele are also explored in Mukhopadhyay et al. [15]. This method is general to any number of $L \geq 2$ groups, where group l consists of n_l individuals.

The similarity scores $y_{l(ij)}^k$ between individuals i and j in group l at variant k are modelled using a one-way ANOVA model at each variant:

$$y_{l(ij)}^k = \mu^k + \alpha_l^k + \varepsilon_{l(ij)k}, \quad i < j = 1, \dots, n_l; \quad l = 1, 2,$$

where at variant k , μ^k is the general effect for pairs of individuals, α_l^k is the group-specific treatment effect, and to test for disease association the null hypothesis is $H_0: \alpha_1^k = \alpha_2^k$. Let n_l be the number of individuals in group l , and let m_l be the number of possible distinct pairs of individuals (i.e. $m_l = n_l(n_l - 1)/2$), which corresponds to the number of distinct similarity scores. Let $\bar{y}_l^k = \sum_{i < j} y_{l(ij)}^k / m_l$ be the mean for group l , and denote the grand mean by $\bar{y}^k = \sum_{l=1}^L \sum_{i < j} y_{l(ij)}^k / \sum_{l=1}^L m_l$. The between-group sum of squares at marker k is defined by $SSB_k = \sum_{l=1}^L m_l (\bar{y}_l^k - \bar{y}^k)^2$, while the corresponding within-group sum of squares is given by $SSW_k = \sum_{l=1}^L \sum_{i < j} (y_{l(ij)}^k - \bar{y}_l^k)^2$. The single variant test statistic at marker k is then given by

$$\frac{SSB_k}{SSW_k},$$

and the K -marker KBAT test statistic to test for disease association with the set of K markers is:

$$\frac{\sum_{k=1}^K SSB_k}{\sum_{k=1}^K SSW_k}. \quad (5)$$

Rather than summing over the K single variant test statistics (ratios), the K -marker test statistic takes the form of (5), which was found to have a higher power when the variants are correlated [see 15]. Clearly the similarity scores $y_{l(ij)}$ are not independent normal random variables, so that neither the single variant test statistics nor the KBAT test statistic (5) may be approximated by an F-distribution. Thus, permutation is required to obtain the p value for each locus.

In extending the allele-matching test to account for genotype uncertainty we refer to the extension as AMELIA (Allele Matching Empirical Locus-specific Integrated Association test). AMELIA is developed as two approaches depending on whether there are consensus or non-consensus variant quality scores. When there are non-consensus quality scores, the (transformed) variant quality scores are incorporated into the analysis by fitting a weighted ANOVA model at each variant k . The weight for the pair of individuals (i, j) in group l is a function of the quality scores $q_{l(i)}^k$ and $q_{l(j)}^k$, with the simplest weight function being $w_{l(ij)}^k = (q_{l(i)}^k * q_{l(j)}^k)$. The weighting scheme that takes the average of the two quality scores was also explored, but tended to have reduced power compared to the product weight function (results not shown). In KBAT, each of the similarity scores contributes a unit weight to the variant-level test statistic. However, with the simple weighting scheme that we consider, similarity scores for which both genotype calls have a high probability of being correct are assigned a weight near 1, while others are appropriately down-weighted. At marker k the weighted sum of squares within groups $wSSW_k$ and between groups $wSSB_k$ may be computed as follows, where for simplicity we have dropped the k superscript, and $\bar{T}_l$ is the weighted group mean of the similarity scores, $\bar{T}..$ is the weighted grand

mean, and $m_l = n_l(n_l - 1)/2$ is the number of similarity scores in group l :

$$wSSW = \sum_{l=1}^L \sum_{j < i} w_{l(ij)} (y_{l(ij)} - \bar{T}_l)^2 \quad (6)$$

$$wSSB = \sum_{l=1}^L m_l (\bar{T}_l - \bar{T}..)^2 \quad (7)$$

Note that if there are consensus quality scores then the weights will be identical at each variant, yielding a test statistic identical to KBAT.

In the simpler case where the variant quality is the same across individuals, the individual variant test statistics remain the same as in KBAT, but components of variant test statistic k in the sums of the K -marker test statistic can be weighted by the variant quality score of variant k . In the locus-specific AMELIA test statistic the single (converted) quality score q_k at each variant k is inserted as a weight into both the numerator and denominator of (5):

$$\frac{\sum_{k=1}^K q_k SSB_k}{\sum_{k=1}^K q_k SSW_k}. \quad (8)$$

In this form, variants that have a low probability of being a true variant contribute a lower weight than the others to the locus-based test statistic.

For all methods, only those variants that have a MAF below a prespecified threshold (0.05) are included in the analysis, where the MAF is calculated from the cases sample. This choice of MAF calculation is arbitrary and alternatives such as calculating MAF only from the control sample or from the entire population may be used instead. None of the MAF calculation methods is ideal, which leads to an arbitrary choice for calculation. The p values for KBAT and AMELIA are obtained via random permutations of the case-control status. Initially, 1,000 permutations are run to calculate the p values, and if one of the two p values is below 0.02, then 10,000 permutations are executed. This is sufficient for power studies, but for data analysis purposes a larger number of permutations may be preferred in order to obtain more precision in the p values; with 10,000 permutations a p value of 0 indicates that the p value is $< 10^{-4}$. For whole-genome analyses the software is designed to analyse each chromosome in parallel, as well as each region (per chromosome) in parallel.

Simulations

A series of simulations is explored in which the causal variants are randomly chosen such that the individual MAF is below $\delta = 0.02$ and the total MAF of the causal variants does not exceed $Q = 0.05$. We assume that there are n individuals (cases + controls), and that the genotype score at a variant is the number of minor alleles present for the individual. Genotype data are simulated to mimic the allele frequencies and pairwise LD structure from a 50-kb region on chromosome 1 (data from 1,000 Genomes pilot study, August 2009 Release, 68 CEU samples) by using the hapsim R package [18]. Variant quality scores

Table 1. Parameter settings for OR in the power studies

ORs for risk variants only	1.5	2	2.5	3	4
ORs for risk and protective variants	1.5, 0.667	2, 0.5	2.5, 0.4	3, 0.333	4, 0.25

The complete set of parameter values for OR. An OR setting with two values indicates that the casual variants are evenly distributed between the two OR values.

are sampled with replacement from the variant phred-scaled quality scores from the same 50-kb region dataset.

A logistic regression simulation model is employed where the genotype score of each causal variant k has an effect β_k on the phenotype, allowing flexibility in the direction and/or magnitude of the causal variant effects. We considered simulations in which all causal variants are risk-increasing, and the setup in which both protective and risk variants are present in the region of analysis. When only risk variants are present in the region, we assign identical ORs to each of these variants, i.e. $\beta_k = \beta = \log(\text{OR})$ for each causal variant k , where the OR is one of the values from table 1. In the scenarios where both protective and risk variants are present in the region, the effect directions are randomly distributed among the causal variants such that there is an even distribution of protective and risk variants. That is, assuming that there are m causal variants, we randomly assign a negative effect $\beta_k = \beta_1 = -\log(\text{OR})$ to m_1 of these variants and a positive effect $\beta_k = \beta_2 = \log(\text{OR})$ to the remaining $m_2 = m - m_1$ variants, where m_1 is the greatest integer less than or equal to $m/2$ (e.g. if $m = 11$ then $m_1 = 5$ and $m_2 = 6$). The parameter settings for OR are chosen from the same set considered for the risk variant simulations, and the complete set of OR parameter settings is provided in table 1. In addition to these parameter values, we also set $\text{OR} = 1$ to find the empirical type 1 error probability. Qualities are incorporated into the simulation as inverse weights to the genotype scores. In these simulation setups oversampling is necessary in order to achieve the desired case-control sample size.

For the non-consensus quality score scenario we focus on three series of simulations that differ in the range of values for the quality scores of variants. First, we allow all positive phred-scaled quality scores (filter out quality scores of zero), but restrict the causal variant phred-scaled quality score to the value 10, which is the usual variant quality threshold used in quality control of the data before analysing for association; this filtering is referred to as Q10. This is equivalent to the causal variants

each having a correct variant call probability of 0.90. These simulations incorporate more information than the usual approach of filtering out variants with quality lower than 10, but follow the assumption that causal variants will be called at a high quality (at Q10 causal variants are not likely to be filtered out under this assumption).

In the next simulation scenario, variants are filtered in the usual manner such that quality scores are restricted to phred-scaled scores ≥ 10 (correct variant call probability ≥ 0.90), and the quality scores of causal variants follow the same distribution as those that are noncausal.

This is the usual Q10 filtering for data before association analysis, which is why we investigate this setup. Due to the high probabilities at all variants there is expected to be little gain in accounting for quality scores, but it is an interesting comparison between the collapsing and allele-matching approaches.

The final simulation series examined for non-consensus quality scores is to allow a more relaxed filtering of the variants such that variant quality scores are filtered to include only phred-scaled scores ≥ 1 (correct variant call probability ≥ 0.205); this filtering is referred to as Q1. This setup takes advantage of more information than the Q10 filtering since more variants are now included in the analysis. Furthermore, it removes the possibility of excluding a causal variant that may be called at a lower quality.

Since AMELIA takes a different form when there are consensus variant quality scores, a set of simulations is also examined for this situation where the causal variants are set to have phred-scaled quality score 10 and all variants with positive quality scores are included in the analysis.

For non-consensus quality scores we initially explore sample sizes of 1,000 (500 cases/500 controls) and 2,000 (1,000 cases/1,000 controls) in the high-quality causal variant setup, but find that allele-matching tests are quite sensitive to the sample size; with $n = 1,000$ there is not a clear pattern in the power differences between the collapsing and allele-matching approaches, but when $n =$

Fig. 1. Distributions of converted quality scores used for simulating data. The distribution of positive converted quality scores (probabilities of correct variant call) for a 50-kb region (left panel). The middle and right panels show the (converted) quality scores filtered at Q1 and Q10, respectively.

2,000 an obvious trend of an increase in power by the allele-matching tests is apparent. For this reason, we focus on $n = 2,000$ in the remaining simulation scenarios.

Results

In various simulation scenarios the power of the two original methods were compared with their extensions that account for incorporating variant quality scores. A 50-kb region from chromosome 1 has been examined, which appears to be representative of a typical region (fig. 1, left panel; online suppl. fig. 1). In this region there were 184 polymorphic variants of which 138 had MAF $< 5\%$, and the minimum MAF was 0.00793. The converted quality score distributions for the Q1 and Q10 filtered region are given, respectively, in the middle and right panels of figure 1.

There are several general conclusions that may be reached from these simulation results. When the direction of effect differs among the causal variants, the collapsing methods had a large loss in power, while the allele-matching tests were more robust. As expected, all tests had higher power when the sample size was increased, but the allele-matching tests were particularly

sensitive. When the causal variants were only of high quality and the data were unfiltered, ARIEL and AMELIA displayed a clear increase in power over their unweighted counterparts, while there was similar performance between the weighted and unweighted approaches without these conditions. For the collapsing methods we also calculated permutation p values, using 1,000 permutations, and found that the powers based on these p values were less than 5% below the parametric p values (results not shown). The type 1 error (at 5% threshold) for the collapsing methods were between 4 and 7%, with the highest error probabilities occurring under the Q1 filtering scenario, while the allele-matching approaches were found to be much more conservative with empirical type 1 errors near 0.5% (table 2).

First, we consider the non-consensus quality score simulations in which there is minimal filtering, but causal variants meet the Q10 filtering standard, so that each causal variant has the probability of a correct variant call set to 0.90. Results for total sample sizes of 1,000 and 2,000 (fig. 2) suggest that the allele-matching tests were more sensitive to sample size than the collapsing approaches. For both sample sizes, and at all OR magnitudes, the collapsing approaches lost approximately half of their power when both effect directions were present compared to

Fig. 2. Power comparisons for the collapsing method, ARIEL, KBAT, and AMELIA with high-quality causal variants. Power comparisons of the four methods at different OR magnitudes for common direction of effect (left panel) and presence of both risk and protective variants (right panel). The simulated data has non-consensus variant quality scores and the causal variants are set to have phred-scaled score quality 10 (probability of correct call is 0.90). The total sample size (cases + controls) is 1,000 (upper) and 2,000 (lower).

Table 2. Type 1 errors for the four methods

Scenario	Collapsing	ARIEL	KBAT	AMELIA
N = 1,000, NC high-quality causals	0.060 (0.045, 0.075)	0.063 (0.048, 0.078)	0.0030 (0, 0.006)	0.0050 (6.3×10^{-4} , 0.0094)
N = 2,000, NC high-quality causals	0.052 (0.042, 0.062)	0.060 (0.050, 0.070)	0.0050 (0.0019, 0.0081)	0.0060 (0.0026, 0.0094)
N = 2,000, NC Q10 filtered	0.041 (0.032, 0.050)	0.041 (0.032, 0.050)	0.0070 (0.0033, 0.011)	0.0060 (0.0026, 0.0094)
N = 2,000, NC Q1 filtered	0.071 (0.060, 0.082)	0.065 (0.054, 0.076)	0.0050 (0.0019, 0.0081)	0.0060 (0.0026, 0.0094)
N = 2,000, C high-quality causals	0.050 (0.040, 0.060)	0.047 (0.038, 0.056)	0.0040 (0.0012, 0.0068)	0.0020 (4.2×10^{-5} , 4.0×10^{-3})

Empirical type 1 errors (and 95% confidence intervals) for the four methods using a threshold of 5%, under each of the simulation scenarios considered. NC denotes non-consensus quality scores, while C denotes consensus quality scores.

Fig. 3. Power comparisons for the collapsing method, ARIEL, KBAT, and AMELIA for Q10 and Q1 filtering of variants. Power comparisons of the four methods at different OR magnitudes for common direction of effect (left panel) and presence of both risk and protective variants (right panel). The simulated data has non-

consensus variant quality scores, where (upper) all variants have phred-scaled score quality ≥ 10 (probability of correct call is ≥ 0.90) and (lower) all variants have phred-scaled score quality ≥ 1 (probability of correct call is ≥ 0.205). The total sample size (cases + controls) is 2,000.

when there was a common direction of effect of the same magnitude. In contrast, the allele-matching tests did not have as large of a power loss when both protective and risk variants were present. For the smaller sample size (500 cases/500 controls), despite the greater power loss from different directions of effect, the collapsing methods outperformed the allele-matching tests if $OR \leq 2$, and power gains for AMELIA were only observed when there were different effect directions and $OR \geq 2.5/OR \leq 0.4$. However, when the sample was increased to 1,000 cases/1,000 controls, there was a clear performance improvement for the allele-matching tests, particularly AMELIA, which was up to 15% more powerful than KBAT in the scenarios considered, demonstrating that incorporating the variant quality scores in this scenario leads to a noticeable power improvement for the allele-matching test.

Continuing with the non-consensus quality score simulations, we next consider Q10 filtering of all variants,

which decreased the number of non-monomorphic variants to 65, of which 32 had MAF below 5%. This setup corresponds to the filtering typically performed before an association analysis and results in uniformly high probabilities (≥ 0.90). ARIEL and AMELIA had very similar performances to their unweighted counterparts (fig. 3, upper panel). Similar to the previous simulations, the allele-matching tests were found to be more powerful than collapsing approaches, most notably when there were both protective and risk variants present. When there was a common direction of effect and $OR \leq 2.5$, the collapsing methods were slightly more powerful than the allele-matching approaches.

In order to make use of more of the variants and to include the possibility of lower-quality causal variants, in the next set of non-consensus quality score simulations we used Q1 filtering, so that all probabilities were ≥ 0.205 . This resulted in 119 non-monomorphic variants of which

Fig. 4. Power comparisons for the collapsing method, ARIEL, KBAT, and AMELIA for consensus quality score data. Power comparisons of the four methods at different OR magnitudes for common direction of effect (left panel) and presence of both risk and protective variants (right panel). The simulated data has consensus variant quality scores, where the causal variants are set to have phred-scaled score quality 10 (probability of correct call is 0.90). The total sample size (cases + contols) is 2,000.

78 had MAF below 5%. Allowing small correct variant call probabilities for the causal variants downweights them in the extended method analyses. ARIEL and AMELIA tended to have a reduced power compared to the unweighted methods by at most 5% (fig. 3, lower panel). When both protective and risk variants were present the allele-matching tests were most powerful at all effect sizes, with power gains of 20–60% depending on the parameter settings. The collapsing approaches appeared to be approximately 10% more powerful when there was a common direction of effect for all causal variants and the OR was low.

An interesting observation is that in comparing the powers based on data filtered such that all variants have phred-scaled quality score above 10 (fig. 3, upper panel) against those with the more relaxed filtering of keeping variants with scores beyond 1 (fig. 3, lower panel), the allele-matching approaches had a 10–20% power increase with the more relaxed filtering compared to the more conservative filtering, provided that both protective and risk variants were present and the $OR \geq 2.5/OR \leq 0.4$. On the contrary, under the same conditions, the collapsing approaches had a 10–15% increase under the more conservative Q10 filtering. This may partially be due to the fact that under Q10 filtering there were fewer variants in the analysis, and thus fewer noncausal variants included, suggesting that the collapsing approaches may not be as robust to the inclusion of noncausal variants as the allele-matching approaches. This is consistent with the result

that in the simulations with minimal filtering and high-quality causal variants (fig. 2) there are more variants in the analysis and the collapsing approaches had lower power than in either of the two quality filtering scenarios.

In the set of simulations where there are consensus quality scores, and causal variants are of high quality, there is a similar pattern to those of non-consensus. In particular, when causal variants have opposite directions of effect the allele-matching tests had a dramatic power increase over the collapsing approaches (fig. 4). Also, AMELIA and ARIEL were consistently more powerful than KBAT and the collapsing method, respectively, but the increase in power was below 5%.

Discussion

We have proposed ARIEL and AMELIA as two methods for low-frequency/rare variant association analyses that incorporate genotype uncertainty via variant quality scores. ARIEL extends a regression framework collapsing method and has the underlying assumption that variants within the functional unit analysed are all protective or all risk-causing, while AMELIA is an extension of an allele-matching approach that does not require such knowledge of the risk alleles and is robust to the presence of LD. Although case-control data have been the focus of the methods described here, extensions to quantitative trait phenotypes are relatively straightforward. The quality-

incorporated extensions tend to be more powerful than their unweighted counterparts, most notably when the causal variants are of high quality. It is worth noting that, under the specific circumstances when the unweighted methods have a higher power than the weighted methods, the difference tends to be at most 5%.

When the sample size is small (500 cases/500 controls) and all causal variants have effects in the same direction, the collapsing methods have a slightly higher power than the allele-matching tests for all OR settings. However, for larger sample sizes (1,000 cases/1,000 controls), this remains to be true only for low to moderate ORs. When both protective and risk variants are present, the collapsing methods have a dramatic power loss, while the allele-matching approaches are more robust and noticeably more powerful. The allele-matching tests tend to have increased power when there is less stringent variant quality filtering.

In summary, when the causal variants are of high quality, there tends to be an advantage in applying AMELIA to unfiltered data. When the quality of the causal variants varies according to the filtering threshold, as in the Q10 and Q1 filtering simulations the allele-matching approaches are more powerful under the less stringent filtering, and the difference in power from incorporating uncertainty was below 5% for the cases studied. In contrast, the collapsing methods are less powerful under the Q1 filtering when the causal variants have effects in both directions, but remain less powerful than the allele-matching methods. However, if the sample size is small, ARIEL is most powerful when there is a common direction of effect of any magnitude or there are low to moderate effects in both directions ($OR \leq 2/OR \geq 0.5$). Since in reality the quality scores of the causal variants are unknown, the most powerful approach appears to be application of the allele-matching test that incorporates uncertainty to minimally filtered data (e.g. Q1 filtered).

For the purpose of comparing the extended methods to their original counterparts we have focused on simulations based on a 50-kb region on chromosome 1. The proposed methods have also been applied in whole-exome analyses of several hundred samples and both scale well to the genome-wide level, taking between a few minutes and less than 24 h to run (depending on the number of permutations). The simulation studies have focused on sample sizes of 1,000 (500 cases/500 controls) to 2,000 (1,000 cases/1,000 controls) and for all methods we observe the expected increase in power with increasing sample size, with the most dramatic increase in power for the allele-matching approaches. Both methods would be

severely underpowered at smaller sample sizes, in keeping with power constraints of complex trait studies.

Both common and rare variants were included in the original implementation of KBAT [15], but we introduced a MAF threshold so that the analyses are based on the same variants analysed by collapsing methods. Furthermore, by focusing only on low-frequency/rare variants there is a gain in computational speed for KBAT and AMELIA which are both computationally intensive due to the large score matrices and permutation testing.

Among the alternative weighting schemes suggested for KBAT [15], only the allele-matching scheme requires no knowledge of the directions of effect since scores are based on counts of common alleles between pairs of individuals in the same group (e.g. cases or controls). Via simulation studies, the allele-matching scores have been illustrated to be robust to a mixture of effect directions among rare variants. This robustness is likely due to the testing of differences in the similarity scores between the groups at each variant, where the scores are not influenced by the coding of the alleles. Other weighting schemes with this property that may be explored are scoring a pair of individuals as 1 if they are exactly the same genotype, and 0 otherwise, or assigning a score of 1 to pairs of individuals with at least one allele in common, and 0 otherwise. These two schemes are analogous to a recessive or dominant coding, respectively.

The field of complex trait genetics has recently moved into a new era of next-generation sequence-based association studies. The development of an appropriate analytical toolbox will be important for the successful identification of novel genetic susceptibility loci. Different methods will be more powerful under different allelic architecture scenarios and the current paucity of empirical data and positive control examples makes it difficult to predict what those architectures will be. Accounting for uncertainty arising from sequencing or imputation approaches in downstream analyses is poised to increase power across the board. We introduce ARIEL and AMELIA as statistical approaches that help fill this gap and enable powerful rare variant analysis.

Acknowledgements

This work was supported by the Wellcome Trust (098051).

The ARIEL and AMELIA packages are freely available from the web sites: <http://www.sanger.ac.uk/resources/software/ariel/> and <http://www.sanger.ac.uk/resources/software/amelia/>.

References

- 1 Cohen J, et al: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004;305:869–872.
- 2 Ji W, et al: Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 2008;40:592–599.
- 3 Nejentsev S, Walker N, Riches D, Egholm M, Todd J: Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009;324:387–389.
- 4 Li B, Leal S: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
- 5 Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;40:695–701.
- 6 Madsen B, Browning S: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;5:e1000384. DOI: [10.1371/journal.pgen.1000384](https://doi.org/10.1371/journal.pgen.1000384).
- 7 Morris A, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;34:188–193.
- 8 Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;70:42–54.
- 9 Zawistowski M, et al: Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 2010;87:604–617.
- 10 Price A, et al: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;86:832–838.
- 11 Bhatia G, et al: A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 2010;6:e1000954. DOI: [10.1371/journal.pcbi.1000954](https://doi.org/10.1371/journal.pcbi.1000954).
- 12 Shriner D, Vaughan L: A unified framework for multi-locus association analysis of both common and rare variants. *BMC Genomics* 2011;12:89.
- 13 Ionita-Laza I, Buxbaum J, Laird N, Lange C: A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 2011;7:e1001289. DOI: [10.1371/journal.pgen.1001289](https://doi.org/10.1371/journal.pgen.1001289).
- 14 Neale B, et al: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;7:e1001322. DOI: [10.1371/journal.pgen.1001322](https://doi.org/10.1371/journal.pgen.1001322).
- 15 Mukhopadhyay I, Feingold E, Weeks D, Thalamuthu A: Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol* 2010;34:213–221.
- 16 Lawrence R, Day-Williams AG, Elliot KS, Morris AP, Zeggini E: CCRaVAT and QuTie: enabling analysis of rare variants in large-scale case control and quantitative trait association studies. *BMC Bioinformatics* 2010;11:527.
- 17 Durbin R, et al: A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–1073.
- 18 Montana G: HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics* 2005;21:4309–4311.