AN EVALUATION OF POWER TO DETECT LOW-FREQUENCY VARIANT ASSOCIATIONS USING ALLELE-MATCHING TESTS THAT ACCOUNT FOR UNCERTAINTY

E. ZEGGINI* and J.L. ASIMIT

Wellcome Trust Sanger Institute, Hinxton, CB10 1HH, UK
**E-mail: Eleftheria@sanger.ac.uk*

There is growing interest in the role of rare variants in multifactorial disease etiology, and increasing evidence that rare variants are associated with complex traits. Single SNP tests are underpowered in rare variant association analyses, so locus-based tests must be used. Quality scores at both the SNP and genotype level are available for sequencing data and they are rarely accounted for. A locus-based method that has high power in the presence of rare variants is extended to incorporate such quality scores as weights, and its power is compared with the original method via a simulation study. Preliminary results suggest that taking uncertainty into account does not improve the power.

Keywords: Allele-Matching; Rare variants;Locus-based method; Quality scores; Sequencing

1. Introduction

There is an increasing interest in the role of rare variants in multifactorial disease etiology, while the evidence that rare variants are associated with complex traits is steadily expanding. Although any individual rare variant exists in low frequencies, the frequency with which any rare variant is present makes them collectively common. Under the multiple rare variant hypothesis (MRV), the effects of multiple rare variants with moderate to high penetrance combine to increase the risk of most common inherited diseases [1]. At the other extreme is the common disease common variant (CDCV) hypothesis, which states that most common complex diseases are due to a few common variants with moderately small effects [2]. The most likely scenario is that a combination of both common and rare variants contribute to disease risk.

In most genome-wide association (GWA) studies only variants with minor allele frequency (MAF) greater than 1-5% are followed up, and the focus tends to be on identifying common disease variants that are associated with complex diseases. However, this approach is limited since only 5-10% of the heritable component of disease is explained by the many genetic variations identified as having strong evidence of disease association in GWA studies. This suggest that a fruitful direction is to search for associations with multiple rare variants [3].

By design, SNP genotyping panels often focus on common SNPs, so that they only contain a relatively small number of rare variants. This leads to a common issue in rare variant analyses, in that on most platforms there is an insufficient number of rare variants (Table 1).

There appears to be a clear difference in the effects of rare variants in comparison to SNPs of higher frequency, with rare variants having stronger effects. According to the odds ratios (OR) for common and rare variants identified in published studies, most common-disease associated variants have ORs between 1.1 and 1.4 with only a few above 2, while the majority of the identified rare variants to date have an OR greater than 2 and a mean of 3.74 [1]. In

Table 1: Approximate low frequency/ rare variant GWAS platform content.

Platform	Affymetrix 500k	Affymetrix 6.0	Illumina 370k	Illumina 550k	Illumina 610k	Illumina 1.2M
MAF<0.05	55k	106k	9k	32k	35k	62k
MAF<0.01	17k	35k	1k	7k	8k	22k

addition, causality may more easily be fine-tuned by identifying rare variants. For most GWA-identified loci, there is difficulty in assigning causality since high LD complicates the use of association mapping to precisely determine which variant is functionally relevant. There are even more complications when elucidating the effects of SNPs that map to genomic regions with no clear role. The problem may be simplified by searching for disease-associated rare variants in known functional genomic regions, such as genes. In addition, it might be easier to at least infer causality at a locus that contains both common and rare disease-associated variants.

In the analysis of the association of rare variants and disease, there is a loss of power due to genotype misspecification. Quality scores are available for genotype and sequence-derived data, but in rare variant analyses, the information is not usually put to use. In addition, the 1000 Genomes reference set contains variants with MAF as low as .01, which makes the imputation of rare variants now possible. A probability distribution for the genotype at each variant may be estimated using the imputation method of choice. We propose methods for rare variant analyses that take advantage of the extra information contained in quality scores derived from sequencing and probability distributions resulting from imputation.

In section 2 we introduce an Allele Matching Empirical Locus-specific Integrated Association test (AMELIA), which is a nonparametric and robust test that accounts for genotype uncertainty. It is an extension of a Kernel-Based Association Test (KBAT) [4], which has been demonstrated to have high power in the presence of rare variants. In section 3 the powers of AMELIA and KBAT are briefly compared in a short simulation study, while a concluding discussion is provided in section 4.

2. Allele-Matching Tests

Before providing the details of AMELIA, we first discuss the original method, KBAT. The kernel-based association test (KBAT) [4] tests for a joint association of multiple SNPs (correlated or independent) with a categorical phenotype, without any assumptions on the directions of individual SNP effects. In simulation studies done by the authors, KBAT was found to generally have more power than other multi-marker approaches (Zglobal[5] and MDMR[6]), especially in the presence of rare causal SNPs. First, similarity scores $y_{l(ij)}$ between individuals i and j in group l (e.g. 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. Let g_i be the genotype score at a specific SNP, which is conveniently defined as the number of reference alleles at the SNP, since knowledge of the risk allele is irrelevant. At a given SNP, for individuals $i \neq j$ in group l with respective genotypes $g_l(i)$ and $g_l(j)$, the

similarity score is defined by

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

By defining the kernel in this way, there is no need to have knowledge of the risk allele at each SNP. Similarity scores that depend on knowledge of the risk allele are also explored in [4]. This is general to any number of $L \geq 2$ groups, where group l consists of n_l individuals.

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

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

where μ is the general effect for pairs of individuals, α_l is the group specific treatment effect, and to test for disease association the null hypothesis is $H_0 : \alpha_1 = \alpha_2$. The single SNP test statistic at marker k is the ratio of the between group sum of squares SSB_k and the within group sum of squares SSW_k, and the K-marker KBAT test statistic is

$$\frac{\sum_{k=1}^{K} SSB_k}{\sum_{k=1}^{K} SSW_k}. \quad (2)$$

Rather than summing over the K single SNP test statistics (ratios), the K-marker test statistic takes the form of (2), which was found to have a higher power when the SNPs are correlated (see [4]). Clearly the similarity scores $y_{l(ij)}$ are not independent Normal random variables, so that neither the single SNP test statistics nor the KBAT test statistic (2) may be approximated by an F-distribution. Thus, permutation is required to obtain the p-value for each locus.

Our extensions that incorporate genotype uncertainty due to quality scores at the SNP and genotype level or imputation are introduced as AMELIA. Here, we focus on the incorporation of the two levels of quality scores. Quality scores of SNPs and genotypes can be accounted for by using weights. Phred quality scores at both the SNP and genotype level are transformed into the probability of a correct call as follows, $1 - 10^{-q/10}$), where q is the quality score. This transformation is employed in order to account for the fact that the phred quality scores are not linear and to avoid down-weighting SNPs that are actually of acceptable quality. For example, quality scores of 30 and 90 both translate to probabilities near 1, and by using the phred quality scores as weights the SNP with score 30 would contribute little weight when it is not really of poor quality.

First, (transformed) genotype quality scores are incorporated into the analysis by fitting a weighted ANOVA model at each SNP k, where the weight for the pair of individuals (i,j) in group l is a function of the genotype quality scores $q^k_{l(i)}$ and $q^k_{l(j)}$, with the simplest weight function being $w^k_{l(ij)} = q^k_{l(i)} + q^k_{l(j)}$. Note that for a more suggestive notation for the quality incorporation into the analysis we use $q^k_{l(i)}$ to denote the transformed genotype quality score. In the original method, KBAT, each of the similarity scores contributes a unit weight to the SNP-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 above 1, while those with two poor scores are down-weighted to contribute a weight

below 1. 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 $\bar{T}_{l\cdot}$ is the weighted group mean of the similarity scores, $\bar{T}_{\cdot\cdot}$ is the weighted grand mean, and $m_l = n_l(n_l - 1)/2$ is the number of similarity scores in group l:

$$wSSW = \sum_{l=1}^{L} \sum_{i=2}^{m_l} \sum_{j<i} w_{l(ij)}(y_{l(ij)} - \bar{T}_{l\cdot})^2 \tag{3}$$

$$wSSB = \sum_{l=1}^{L} m_l(\bar{T}_{l\cdot} - \bar{T}_{\cdot\cdot}) \tag{4}$$

Components of SNP test statistic k in the sums of the K-marker test statistic can be weighted by the SNP quality score(s) of SNP k. In the case that there is a common SNP quality score Q_k across all individuals (score at a SNP is based on reads from all individuals), the weight for SNP k in the sums is simply the (transformed) single SNP quality score Q_k. If the quality scores at a SNP differ among individuals (score at a SNP based on multiple reads from single individual), then the weight may be taken as the sum of these scores at the SNP. In the latter case, the K-marker test statistic is

$$\frac{\sum_{k=1}^{K} Q_k SSB_k}{\sum_{k=1}^{K} Q_k SSW_k}. \tag{5}$$

In this form, SNPs that have a low probability of being a true variant contribute a lower weight than the others.

2.1. *Implementation*

In order to increase the speed of the permutations, as suggested in [4], the similarity scores between all possible pairs of individuals are computed, regardless of which cohort they belong to. Then, in the permutation stage, the similarity scores for the permuted case-control samples may be quickly extracted without further computation. However, for large cohorts ($N > 1000$), this causes both AMELIA and KBAT to be memory-intensive, requiring additional memory allocation to run. For example, when $N = 1000$ there are 499,500 similarity scores between all possible pairs of individuals, which requires manipulation of a 499,500 $\times$ 499,500 array. The time requirement for both methods also increases with the number of SNPs since a test statistic must be computed at each SNP for each permutation.

3. Simulation Study

A brief simulation study has been run to compare the powers of KBAT [4] and our version of AMELIA that accounts for quality scores. Genotype and quality score data are simulated based on data from the pilot study of 1000 Genomes (68 individuals). More specifically, we use the `haplosim` function of the `hapsim` [7] R package to simulate a population of haplotypes that possess the same allele frequencies and pairwise LD structure as a specified chromosomal region from the 1000 Genomes data. This approach produces realistic data that includes variants with MAFs down to .01. A cohort of N individuals is formed by randomly pairing up

$2N$ haplotypes sampled from a population of 40000 simulated haplotypes. SNP and genotype quality scores were generated by randomly sampling with replacement from the scores observed in the 1000 Genomes data. In the simulations considered there is only one causal SNP, which has a MAF close to a certain frequency, and is chosen randomly among the possible SNPs that satisfy this criterion. More complicated simulations involving multiple causal SNPs are to be explored in the near future.

Case-control status is generated by using a multiplicative model for the genotype relative risks to compute the probability of disease given the genotype at the causal SNP and its relative risk (RR) (for details see [4]). This probability is then used to generate a Bernoulli random variable that ascertains an individual as a case when its value is 1, and a control otherwise. For this reason, it is necessary to over-sample (say, $5N$) the number of individuals to ensure that the desired number of cases is attained.

In order to obtain the p-value in an efficient manner, we first obtained p-values based on 1000 permutations. If this p-value was below .02, additional permutations were run to update the p-value on the basis of 10,000 permutations. This procedure of updating the p-value continues up to a maximum of 1,000,000 permutations, if necessary.

In order to compare the two tests in a scenario similar to that of [4], rather than testing the whole region we also test regions of 11 SNPs formed from the causal SNP and 10 randomly selected SNPs among the 20 SNPs that form a neighborhood around the causal SNP (10 upstream and 10 downstream from the causal SNP) (termed the neighborhood region).

3.1. *Results*

In this brief simulation study, a 150 KB region from chromosome 1 of the 1000 Genomes data was considered, which contains 342 SNPs. This region was chosen slightly arbitrarily, but also because it has a genome-average recombination rate of approximately 1Mb/cM. All SNPs were retained, except for those with a SNP quality of 0. We assumed a single low frequency causal SNP (MAF=.02, RR=2), and 500 cases and 500 controls were simulated over 1000 replications.

Table 2: Power results (5% level of significance) for AMELIA and KBAT when there is one rare causal SNP and there are 500 cases and 500 controls.

region	AMELIA	KBAT
whole	.0871	.0953
neighborhood	.1731	.2161

When jointly testing all SNPs within a region there is a slight loss of power with the use of AMELIA in comparison to KBAT. However, both methods have a relatively low power when there are many SNPs in the region. In a comparable scenario examined in [4], where the region contains only 10 SNPs and the causal SNP has a MAF of .108 with RR=1.25 the power of KBAT was .323. In our neighborhood simulations comparing AMELIA and KBAT we obtain powers of similar magnitude (see Table 2). Thus the low powers for the entire region tests are

likely due to the fact that our region of 150kb contains almost 350 SNPs, which are all jointly tested. This illustrates a caveat of this multi-marker testing approach.

In order to examine type I error, a null simulation in which we set the relative risk as 1 is also examined. However, we only consider the neighborhood region due to the extremely low power observed for the entire region. At the 5% level both methods are found to be quite conservative, with AMELIA and KBAT having respective type I errors of .00502 and .00401.

4. Discussion

In the short simulation study presented here, a decrease in power has been observed by incorporating quality scores of SNPs and genotypes as in AMELIA, with the difference largest for a small number of SNPs. The relatively low power of the two methods may be due to the fact that almost 350 SNPs are being tested jointly, of which there is only one causal SNP. This may suggest that this multi-marker approach may be best suited for smaller regions, or after some filtering to reduce the number of SNPs that are jointly tested. For example, when the focus is on low-frequency variants, the analysis may include only those with a MAF below a certain threshold, such as 0.05. It is noted that the replications that were identified only by KBAT tend to have a causal SNP with a high SNP quality score. In such situations it may be that by allowing for uncertainty that is not present, power to detect the signal is inadvertently diluted. In the simple simulations examined, the power of AMELIA appears to be lower than KBAT, and both tests are conservative with similar error rates. We are extending our methods further to achieve greater power.

References

1. Bodmer W and Bonillna C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**: 695-701.
2. Pritchard JK and Cox NJ. (2002). The allelic architecture of human disease genes: common disease common variant ... or not? *Human Molecular Genetics* **11**: 2417-2423.
3. Schork NJ, Murray SS, Frazer KA, and Topol EJ. (2009). Common vs rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development* **19**: 212-219.
4. Mukhopadhyay I, Feingold E, Weeks DE, and Thalamuthu A. (2010). Association Tests Using Kernel-Based Measures of Multi-Locus Genotype Similarity Between Individuals. *Genetic Epidemiliogy* **34**: 213-221.
5. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, and Thibodeau SN. (2005). Nonparametric Tests of Association of Multiple Genes with Human Disease. *American Journal of Human Genetics* **76**: 780-793.
6. Wessel J and Schork NJ. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. American Journal of Human Genetics 79: 792-806.
7. Montana G. (2005). HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics* **21**: 4309-4311.