META-ANALYSIS OF A RARE-VARIANT ASSOCIATION TEST.

THOMAS LUMLEY, JENNIFER BRODY, JOSÉE DUPUIS, ADRIENNE CUPPLES

Abstract. Genome-wide assocation studies have often been carried out by meta-analysis rather than by pooling individual-level data. For one-dimensional parameter estimates and the corresponding tests of association these meta-analyses lead to essentially no loss of information relative to pooling individual data. The situation is different for multi-parameter tests, such as the omnidirectional rare-variant tests being used in resequencing studies. In this paper we consider one popular rare-variant test, a version of the sequence kernel association test. We show that meta-analyses based on the *p*-value or test statistic from each contributing study are importantly less efficient than an analysis pooling individual data, but that a more sophisticated meta-analysis retains full efficiency. The meta-analysis is based on a reformulation of the test that links it to tests used in survey analysis.

Keywords: Rao-Scott test; score test; sequence kernel association test; genetic epidemiology; DNA sequencing

1. Introduction

Many large genome-wide association studies have been performed by *ad hoc* international collaborations that are unwilling or unable to share individual-level genetic data, and so have used meta-analysis to combine study-specific estimates. In these studies, each test or estimate is typically for the additive component of the genetic association between phenotype and a single SNP, a setting where meta-analysis is fully efficient[Lin and Zeng, 2010a,b]. Genetic epidemiology is now moving on to DNA resequencing studies, which generate large numbers of very rare sequence variants. It is not possible to test each variant individually, and many tests for the collective effect of rare variants in an exon, gene, or set of genes have been developed [eg Wu et al., 2011, Madsen and Browning, 2009, Hoffman et al., 2010, Morgenthaler and Thilly, 2007, Li and Leal, 2008]. Some of these, the 'unidirectional tests', are still based on a one-dimensional summary statistic, and can be written as score tests for a single regression parameter[Lin and Tang, 2011], leading to fully-efficient meta-analyses. Other tests are intended to detect any combination of positive and negative effects. These have sometimes been called 'bidirectional' tests; we prefer the term 'omnidirectional' to emphasize that the test is sensitive to departures from the null in any direction in a high-dimensional space of possible alternatives.

For the omnidirectional tests, the test statistic is typically not even asymptotically a sufficient statistic and there can be substantial information loss in meta-analysis. In this paper we consider the variance-component score test case of the sequence kernel association test[SKAT, Wu et al., 2011], and show that meta-analysis based on the p-value or the test statistic loses substantial information, but that a more complicated meta-analysis with nearly full efficiency is possible. Our proposed meta-analysis requires pooling individual-variant score statistics and the genotype covariance matrix. The individual-variant test statistics are routinely shared for meta-analysis in genome-wide association studies, and the genotype covariance matrix contains no information about genotype:phenotype associations, so data sharing for these quantities should be no more difficult to arrange than in GWAS studies.

In section 2 we formulate the SKAT test in terms of regression and define a meta-analysis. Section 3 describes some generic approaches to meta-analysis of test statistics or p-values and compares these to the SKAT meta-analysis in simulation. In section 4 we comment on the scope for efficient meta-analysis of other rare-variant tests.

2. A regression formulation of the SKAT variance component test

The SKAT variance score test is designed as an omnibus test for a collection of possibly-rare sequence variants. It was developed as an example of the very general 'sequence kernel association test' and also justified as a score test in a random-effects model. There is another way to formulate the variance component score test that illuminates its relationship to more familiar omnibus tests and shows how an efficient meta-analysis can be constructed.

Consider the generalized linear model for additive effect of k variants $G_1, \ldots, G_k$:

$$g(E[Y]) = \alpha + \sum_{i=1}^{k} \beta_i G_i.$$

The approach taken in the SKAT variance component test is to model the vector β as a random sample from an unknown effect-size distribution P, scaled by a variance τ^2. If τ^2 is zero, the variants have no effect on the phenotype, but if $\tau^2 > 0$ there is an effect. Wu et al. [2011] show that the score test of the null hypothesis $\tau^2 = 0$ does not depend on P and is a special case of their kernel-machine assocation test.

The standard omnibus Wald test for all $\beta_i = 0$ is based on

$$\hat{\beta} V^{-1} \hat{\beta}^T = z R^{-1} z$$

where V^{-1} is the covariance matrix of $\hat{\beta}$, R is the correlation matrix, and z is the vector of z-statistics. For rare variants, this covariance matrix is not well estimated, and the standard asymptotic approximation is likely to be singular In a logistic or survival model, $\hat{\beta}$ is also quite likely to be infinite.

The problem of poorly-estimated V is an old one in survey statistics, where the effective degrees of freedom of a large survey may be surprisingly small. For example, a two-year wave of the National Health and Nutrition Examination Survey (NHANES) has a sample size of roughly 10,000, but only about 15 degrees of freedom for estimating V.

A solution [Rao and Scott, 1981] is to use a score test, thus avoiding the need for $\hat\beta$, and to weight the contributions of each parameter using only the diagonal of the Fisher information, not the whole matrix. That is, the test statistic is the (weighted) sum of squares of z-statistics from score tests of each variant taken separately. In linear regression the score statistic is just the single-variant regression coefficient divided by its standard error; in logistic regression the score statistic can be computed from a single iteration of iteratively reweighted least squares.

The simplest form of the SKAT statistic is thus an unweighted sum of squares of per-variant association tests

$$Q=\sum_{i=1}^k z_i^2.$$

This corresponds to the test of Wu et al. [2011] with what they refer to as Madsen–Browning weights. More generally, the SKAT test incorporates a vector of weights w based on allele frequency and, in principle, on functional annotation, giving

$$Q=\sum_{i=1}^k 2p_i(1-p_i)w_i^2z_i^2.$$

The Madsen–Browning weights are proportional to $p_i^{-1/2}(1-p_i)^{-1/2}$, where p_i is the observed minor allele frequency for variant i, but Wu et al prefer weights proportional to $(1-p_i)^24$

In the context of rare genetic variants, the impact of using just the diagonal of $\hat V^{-1}$ for weighting is that redundant information from correlated SNPs is given more weight in the SKAT test than in the Wald test. In the survey context the improvement in small-sample behaviour by using only the diagonal of the Fisher information in weighting is substantial, even when $\hat\beta$ and $\hat V^{-1}$ do exist.

The resulting test statistic would have a scaled χ^2_p distribution if the variants were truly independent and Madsen–Browning weights were used. If the variants are not independent, the distribution depends on the correlation matrix. Let C be the observed genotype covariance matrix (which is proportional to $\hat V$), and let w the vector of weights, then the distribution of Q is

$$\sum_{i=1}^k z_i^2 \sim \sum_{i=1}^k \lambda_i\chi^2_1$$

where λ_i are the eigenvalues of wCw^T. Note that it is not necessary that all the eigenvalues are non-zero, either theoretically or in practice, and that rescaling the weights w by an arbitrary factor will rescale the eigenvalues by the square of this factor, resulting in the same p-value.

It might appear that the instability of the estimate of V makes it impossible for this procedure to work: the individual eigenvalues will certainly not be well-estimated. Considering the Satterthwaite approximation to the distribution makes the test look more plausible. The Satterthwaite approximation to the distribution in the simplest case, with Madsen-Browning weights, is

$$\sum_{i=1}^{k} z_i^2 \dot{\sim} a\chi_q^2$$

where

$$a = \frac{1}{k}\sum_{i,j=1}^{k} r_{ij}^2$$

and

$$q = \frac{k^2}{\sum_{i,j=1}^{k} r_{ij}^2}.$$

So, to a good approximation, the asymptotic distribution depends only on the number of variants observed and the *average* linkage disequilibrium between variants, both of which are relatively well estimated.

2.1. **Meta-analysis.** Computing the SKAT test using our new formulation requires the estimation of per-variant score test statistics and of the genotype covariance matrix, using summary statistics from each cohort. In the simplest case of linear regression with no adjustment variables we proceed in the following steps

(1) Each study computes and shares its genotype covariance matrix C_m. These are averaged, with weights proportional to the sample size of the study. From the genotype covariance matrix (or separately), the overall minor allele frequency is computed for each variant and sent back to the studies. It is important to ensure that variants not seen in a study are included in the computations, with zero copies, rather than treated as missing.
(2) The score test statistics are computed as

$$\begin{aligned}\hat{\beta}_{im} &= \frac{\sum_{j=1}^{n_m} G_{imj} Y_{mj}}{2n_m p_i(1-p_i)} \\ s_{im}^2 &= \frac{1}{n_m}\hat{\sigma}_m^2 2p_i(1-p_i)\end{aligned}$$

where G_{imj} and Y_{mj} are the genotype and phenotype for individual j, study m, and variant i; n_m is the sample size in study m and $\hat{\sigma}^2_m$ is the variance of the phenotype. That is, the variance and the denominator of $\hat{\beta}$ are computed using the full-data minor allele frequency, not the study-specific minor allele frequency. Note that if a study does not observe variant i, the regression coefficient reduces to $\hat{\beta}_i = 0$. The studies then share $\hat{\beta}_i$ and s^2_i.

(3) A standard precision-weighted meta-analysis is done to obtain $\hat{\beta}_i$ and $\hat{s}^2_i$, and thus z_i.

(4) The test statistic Q is computed as the weighted sum of z^2_i and the null distribution is computed from the averaged genotype covariance matrix.

Modifications for case-control data are straightforward and code is supplied in the Supplementary Information. In the presence of additional adjustment variables the meta-analysis is less exact, being conservative when the adjustment variables are correlated with phenotype and anticonservative when they are correlated with genotype, although empirically it still agrees closely with results based on complete individual data.

Allowing for sampling weights to accommodate complex multi-phenotype sampling plans is also straightforward: regression with sampling weights is supported in most general-purpose statistics packages, and the estimated regression coefficients and standard errors can be combined exactly as above.

2.2. **Numerical analysis issues.** Code for the linear combination of χ^2_1 is readily available in R, using an Applied Statistics algorithm that inverts the characteristic function[Davies, 1980], or a saddlepoint approximation[Kuonen, 1999], both of which are nearly exact. The CompQuadForm package, function `davies()` inverts the characteristic function; the function `pchisqsum()` in the survey package provide a more consistent interface to different ways of calculating the tail probabilities. For other languages, the Satterthwaite approximation described above, which behaves well for moderate p-values but is anticonservative in the extreme tail, is trivial to implement and is standard in survey statistics.

Even though the matrix C is positive semi-definite by construction, computing the eigenvalues to finite numeric precision may give some small negative values or even complex values. Complex eigenvalue estimates can be avoided by using an eigenvalue routine designed for symmetric matrices. Small negative eigenvalues present no problem for computing the p-value using the Satterthwaite, saddlepoint, or Davies methods. Alternatively, small and complex eigenvalues can simply be dropped as they do not contribute meaningfully to the p-value computation. For example, the SKAT software [Wu et al., 2011] uses only the eigenvalues within a factor of 10^5 of the largest eigenvalue.

3. Other approaches to the meta-analysis

Simply adding up SKAT test statistics from each cohort gives a statistic whose distribution is the sum of the M linear combinations of χ^2_1s. That is, if Q_m is the SKAT test statistic in cohort m, and ν_m is a cohort weight (perhaps proportional to sample size), the overall test statistic would be $Q^\star = \sum_{m=1}^M \nu_m Q_m$ with sampling distribution

$$\sum_{m=1}^{M}\sum_{i=1}^{k_m} z_i^2 \sim \sum_{m,i} \lambda_{im}\chi^2_1$$

If there were no overlap in variants between cohorts, $Q^\star$ would be the same as the SKAT test statistic based on complete individual data, and also the same as our proposed meta-analysis. When there is overlap between cohorts, $Q^\star$ will give a valid, but less powerful test. For example, in the extreme case where the variants are all independent and have the same allele frequencies between cohorts, $Q^\star$ will have the same non-centrality parameter as Q, but more degrees of freedom.

A very general approach to meta-analysis of complex test statistics is to base the meta-analysis on p-values, either with a log transformation[Fisher, 1932] or an inverse-normal transformation[Stouffer et al., 1949]. If p_{Qm} is the p-value for cohort m, then

$$F = -2\sum_{m=1}^{M} \log p_{Qm}$$

has a χ^2_{2k} null distribution and

$$S = \sum_{m=1}^{M} \Phi^{-1}(p_{Qm})$$

has a Normal null distribution. In both cases weights can be used, eg proportional to square root of sample size; the inverse-normal transformation still gives a Normal null distribution and the log transformation gives a linear combination of χ^2 distributions as the null.

To compare these meta-analysis approaches with our proposed SKAT meta-analysis we conducted a simulation study. We simulated 40kB of DNA sequence for each of 4000 people using MaCS[Chen et al., 2009], dropped variants with minor allele frequency over 1%, and divided the data into three cohorts of size 1000, 2000, and 1000. Half the variants had no effect, for the other half true genotype effects were simulated from the Beta(1,25) distribution corresponding to the default weights for the SKAT test, and randomly assigned positive or negative signs. Phenotypes were generated according to a Normal distribution, and in addition to the genotype effects there was a cohort mean of (0, 0.2, 0.4) for the three cohorts respectively. The simulations were repeated 5000 times.

We compared the SKAT test on complete data, adjusted for cohort, the proposed meta-analysis, the meta-analysis based on $Q^\star$ with $\nu_m = (1, 2, 1)$, a normal transformation of

p-values with weights $(1, \sqrt{2}, 1)$, and log transformations without weights and with weights $(1, 2, 1)$. The saddlepoint approximation was used for the linear combination of χ^2 in all cases. The results are in Table 1.

Table 1. Simulated power of SKAT test on full data, the proposed meta-analysis, and four meta-analyses using only the per-study test statistic or p-value. Based on 5000 simulations, 4000 individuals in three cohorts, and 157 variants with minor allele frequency less than 1% simulated using MaCS

	Power		
	at $\alpha = 0.001$	at $\alpha = 0.01$	at $\alpha = 0.05$
SKAT test	0.09	0.22	0.41
proposed meta-analysis	0.09	0.22	0.41
sum of test statistics ($Q^\star$)	0.03	0.12	0.28
weighted inverse-normal transform	0.04	0.13	0.27
log transform	0.03	0.11	0.26
weighted log transform	0.03	0.12	0.28

The simulation shows that our proposed meta-analysis does retain essentially all the information in the data, and that the techniques based on only the test statistic or p-value from individual studies are substantially less efficient. It is interesting that the power is so similar for the four less-efficient techniques; the differences shown in the table are larger than the Monte Carlo error but are small in practical terms.

Table 2. Simulated power of SKAT test adjusted for principal components, on full data, the proposed meta-analysis, and four meta-analyses using only the per-study test statistic or p-value. Based on 5000 simulations, 4000 individuals in three cohorts, and 157 variants with minor allele frequency less than 1% simulated using MaCS, 3500 variants used to compute principal components.

	Power		
	at $\alpha = 0.001$	at $\alpha = 0.01$	at $\alpha = 0.05$
SKAT test	0.13	0.26	0.46
proposed meta-analysis	0.12	0.25	0.45
sum of test statistics	0.07	0.16	0.33
weighted inverse-normal transform	0.08	0.20	0.37
log transform	0.08	0.18	0.34
weighted log transform	0.06	0.17	0.35

To investigate the impact of other adjustment variables we simulated 400kB of DNA sequence for each of 4000 people using MaCS [Chen et al., 2009] and computed five principal

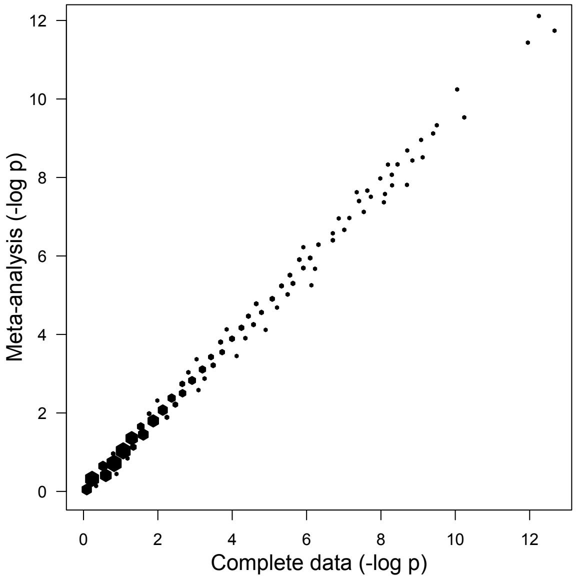

Figure 1. Agreement of complete-data and meta-analysis p-values (on $-\log_{10} p$ scale), in analysis adjusted for study and five study-specific principal components of genetic variation

components of genetic variation. We then dropped variants with minor allele frequency over 1%, and divided the data from the first 157 of the remaining rare variants, matching the number used in the previous simulation, into three cohorts of size 1000, 2000, and 1000. Phenotypes were generated the same way as in the previous simulation, and again we used 5000 repetitions. Table 2 again shows very good agreement between the SKAT test on complete data and the proposed meta-analysis, and the power loss from using only the test statistic or p-value. Figure 1 plots the $-\log_{10} p$-value from the complete-data test and the meta-analysis, showing that the agreement is good uniformly, not just on average.

4. Discussion

We have shown that the loss of efficiency from meta-analysis based only on p-values can be substantial for the omnidirectional SKAT variance component test, in contrast to the situation with one-dimensional parameters and tests familiar from GWAS and clinical-trial meta-analysis. We have also shown how the SKAT variance component score test can be meta-analyzed without without loss of efficiency, sharing only the sort of information that is routinely shared in GWAS analyses.

Our paper does not provide a general approach to efficient meta-analysis for other multi-variant test statistics, and we do not believe such an approach exists in general. The loss of power we have demonstrated is likely to be greater for tests based on model selection, shrinkage, or cross-validation if they are run without special tuning. These tests attempt to optimize a bias:variance tradeoff within the individual study, but since meta-analysis will tend to reduce variance but not bias, the optimal bias:variance tradeoff will be different for estimates intended for meta-analysis.

In general, efficient analysis of DNA sequence data will require either careful attention to constructing estimators that have finite-dimensional sufficient statistics, or a more liberal approach to data pooling.

References

Gary K. Chen, Paul Marjoram, and Jeffrey D. Wall. Fast and flexible simulation of dna sequence data. *Genome Research*, 19(136-142), 2009.

Robert B. Davies. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):pp. 323–333, 1980.

R. A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 4 edition, 1932.

Thomas J. Hoffman, Nicholas J. Marini, and John S. Witte. Comprehensive approach to analyzing rare genetic variants. *PLoS One*, 5(11), 2010.

D. Kuonen. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):pp. 929–935, 1999.

B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*, 83(3):311–321, 2008.

D. Y. Lin and Z. Z. Tang. A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics*, 89:354–367, 2011.

D. Y. Lin and D. Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010a.

D. Y. Lin and D. Zeng. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology*, 34:60–66, 2010b.

Bo Eskerod Madsen and Sharon R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), 2009.

S. Morgenthaler and W. G. Thilly. A groupwise association test for rare mutations using a weighted sum statistic. *Mutation Research*, 615:28–56, 2007.

J. N. K. Rao and A. J. Scott. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76(374):pp. 221–230, 1981.

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, and Jr R. M. Williams. *The American soldier, Volume I: Adjustment during Army life.* Princeton University Press, Princeton, NJ, 1949.

M. C. Wu, S. Lee, T. Cai, Y. Li, M Boehnke, and X. Lin. Rare variant association testing for sequencing data using the sequence kernel association test (skat). *American Journal of Human Genetics*, 89(82-93), 2011.