European Journal of Human Genetics (2014) 22, 414–418

www.nature.com/ejhg

ARTICLE

Correction for multiple testing in a gene region

Audrey E Hendricks*[,1], Josée Dupuis[1,2], Mark W Logue[1,3], Richard H Myers[4] and Kathryn L Lunetta[1]

Several methods to correct for multiple testing within a gene region have been proposed. These methods are useful for candidate gene studies, and to fine map gene-regions from GWAs. The Bonferroni correction and permutation are common adjustments, but are overly conservative and computationally intensive, respectively. Other options include calculating the effective number of independent single-nucleotide polymorphisms (SNPs) or using theoretical approximations. Here, we compare a theoretical approximation based on extreme tail theory with four methods for calculating the effective number of independent SNPs. We evaluate the type-I error rates of these methods using single SNP association tests over 10 gene regions simulated using 1000 Genomes data. Overall, we find that the effective number of independent SNP method by Gao *et al*, as well as extreme tail theory produce type-I error rates at the or close to the chosen significance level. The type-I error rates for the other effective number of independent SNP methods vary by gene region characteristics. We find Gao *et al* and extreme tail theory to be efficient alternatives to more computationally intensive approaches to control for multiple testing in gene regions.
European Journal of Human Genetics (2014) **22,** 414–418; doi:10.1038/ejhg.2013.144; published online 10 July 2013

Keywords: correlated; permutation; Bonferroni; type-I error; effective; independent

INTRODUCTION

Methods to control for multiple comparisons within gene regions are used for various kinds of analyses including candidate gene studies, and higher order analyses such as single-nucleotide polymorphism (SNP)–SNP interaction analysis among pairs or groups of genes. The Bonferroni correction is simple and widely used, but is usually overly conservative due to high correlation among SNPs in a chromosomal region. Permutation provides a type-I error rate that asymptotically approaches the chosen significance level when the units being permuted are exchangeable under the null hypothesis. However, permutation is computationally intensive especially for high-throughput analyses or simulations. A computationally efficient option, which is less conservative than the basic Bonferroni correction, is to calculate the effective number of independent SNPs (M_{effs}) in a gene region and to use this value in the Bonferroni correction.[1–5] One further option is to use extreme tail theory to explicitly calculate the probability of detecting a test statistic as large as or larger than the observed maximum test statistic in the gene region.[6]

Although evaluations of methods to control for multiple testing exist,[6,7,8] the evaluations often have drawbacks. Some were done using 20 or fewer markers, well below the number typically seen in gene regions of a few hundred kb. Others did not simulate gene region variation between replicates, which may be particularly problematic for methods such as M_{eff} methods and extreme tail theory, which are dependent on the correlation within the gene region.

In 2008, Moskvina and Schmidt compared extreme tail theory with the first M_{eff} method developed by Cheverud[1] and modified by Nyholt[5] across scenarios with both a small (40) and large ($\sim$6000) number of SNPs.[8] They found that extreme tail theory produced a more accurate estimate for M_{eff} while still being computationally efficient. Cheverud and Nyholt's method is known to be overly conservative,[3,4] and Moskvina and Schmidt did not compare extreme tail theory with some of the more recent M_{eff} methods as we do here. Thus, the relative merits of extreme tail theory versus more recently proposed M_{eff} methods remain untested.

First, in order to gain a more complete understanding of M_{eff} methods, we compare the methods using simple examples where the number of independent SNPs is known. Using logistic regression to assess the association between case status and SNPs, we compare the type-I error rates for the M_{eff} methods and extreme tail theory in gene region simulations where we strive to overcome some of the drawbacks of previous studies by varying replicate linkage disequilibrium (LD) structure, and including many SNPs in each region. To our knowledge, no study has compared extreme tail theory with the most promising M_{eff} methods.

METHODS

Methods to compute the effective number of independent SNPs

Unless otherwise indicated, we calculated the eigenvalues, λ, using eigenvector decomposition in R[9] with the genotypic correlation of additively coded genotypes using Pearson's correlation coefficient. We outline the M_{eff} methods below where M represents the total number of markers in the region and M_x is the M_{eff} calculated by method x.

Cheverud[1] and Nyholt[5]

$$M_{\text{Chev}} = 1 + (M-1)\cdot\left(1 - \frac{\text{var}(\lambda)}{M}\right) \quad (1)$$

where $\text{var}(\lambda)$ is the variance of the eigenvalues.

Cheverud first developed equation (1) to calculate the M_{eff} dependent on the variation of the eigenvalues calculated using genotypic correlation. Nyholt then modified Cheverud's method by using the allelic correlation to calculate the eigenvalues rather than the genotypic correlation.

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; [2]Bioinformatics Program, Boston University, Boston, MA, USA; [3]Department of Biomedical Genetics, Boston University School of Medicine, Boston, MA, USA; [4]Department of Neurology, Boston University School of Medicine, Boston, MA, USA
*Correspondence: Dr AE Hendricks, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. Tel: +44 1223 837 175; Fax: +44 1223 494 919; E-mail: ah16@sanger.ac.uk
Received 8 May 2012; revised 27 April 2013; accepted 11 May 2013; published online 10 July 2013

Li and Ji[4]

$$M_{\text{Li and Ji}} = \sum_{m=1}^{M} I[\lambda_m \geq 1] + (\lambda_m - |\lambda_m|) \quad (2)$$

In 2005, Li and Ji developed a method that separates the eigenvalues into two components representing: (1) the correlation between SNP genotypes (the integers of the eigenvalues) and (2) the independent contribution of each SNP (the remainders of the eigenvalues).[4] Li and Ji sum these components over all of the eigenvalues to estimate the M_{eff}.

In practice, decomposing a correlation matrix into its eigenvalues can sometimes yield very small negative numbers. Taking the floor of these negative eigenvalues for Li and Ji's method can provide inaccurate results. Thus, when implementing Li and Ji's method we used the absolute value of the eigenvalues.

Gao *et al*[3]

$$\frac{\sum_{m=1}^{M_{\text{Gao}}} \lambda_m}{\sum_{m=1}^{M} \lambda_m} \geq c \quad (3)$$

where c is user defined.

Gao *et al*'s method estimates the M_{eff} as the number of eigenvalues needed to explain a prespecified proportion of the sum of all of the eigenvalues.[3] They suggest that a threshold of 0.995 works well in most situations, although a higher or lower threshold would likely perform better depending on the LD structure of the gene region. We use Gao's recommendation of $c = 0.995$ in our implementation.

Galwey[2]

$$M_{\text{Galwey}} = \frac{\left[\sum_{m=1}^{M} \sqrt{\lambda_m}\right]^2}{\sum_{m=1}^{M} \lambda_m} \quad (4)$$

In 2009, yet another equation to calculate the effective number of independent SNPs was proposed by Galwey. As previously mentioned, eigenvalue decomposition sometimes yields very small negative numbers. As the square root in Galwey's original equation cannot use negative numbers, Galwey suggests changing all negative eigenvalues to zero. Here, we use the absolute value of the eigenvalues, which we found produces identical results to setting all negative eigenvalues to 0 as Galwey suggested.

Extreme tail theory

To our knowledge, using extreme tail theory to control for multiple comparisons in a gene region was first described by Conneely and Boehnke in 2007[6] and was further evaluated by Moskvina and Schmidt in 2008.[8]

Assuming a multivariate normal distribution for test statistics under the null hypothesis of no association, we calculated the probability of observing a maximum test statistic as large as or larger than a certain threshold.

P_{ET} is the probability of observing at least one test statistic whose absolute value is as large or larger than a critical value, Z^*; M is total number of markers or test statistics; m is the marker or test statistic indicator; N is total number of subjects; i is subject indicator; Y is the phenotype; x_m is SNP m; Z_m is test statistic m.

$$\begin{aligned} P_{\text{ET}} &= P(\max(|\underline{Z}|) \geq Z^*) = 1 - P(\max(|\underline{Z}|) < Z^*) \\ &= 1 - P(|Z_1|, \ldots, |Z_M| < Z^*) = 1 - \int_{-Z_1^*}^{Z_1^*} \cdots \int_{-Z_M^*}^{Z_M^*} \Phi(Z_m)\mathrm{d}z \end{aligned} \quad (5)$$

where Φ is the multivariate normal probability density function.

As shown in equation (5), P_{ET} depends on the joint distribution of M statistics $(Z_1,\ldots, Z_M)$ where $Z \sim N(0, \Sigma)$. In 2007, Conneely and Boehnke showed that, under the null hypothesis of no association, the covariance of the test statistics, Σ, can be calculated directly from the correlation between SNPs $(X_1,\ldots, X_M)$. We use this to find the critical value of the test statistic that corresponds to the multivariate probability, P_{ET}, in equation (5).

Simulation design

Simple examples: known number of independent SNPs. We created SNP correlation matrices to have an independent block structure where the SNPs within each block are perfectly correlated. First, we created independent blocks each with an equal number of SNPs ($m = 2$, 5, and 20) and varied the number of blocks from 1 to 10. We then simulated two or three independent blocks each containing a different number of SNPs between 0 and 100. To highlight the implications of the later scenario, we simulated many pairs of blocks ($N_{\text{blockpairs}} = 1$, 20, and 50) so that each pair contained a total of 10 SNPs distributed unequally between the two blocks.

As we created situations consisting of independent blocks where the number of independent SNPs is known, we can rewrite the M_{eff} equations in terms of the block structures. Rewriting the equations gives us insight into the tendencies of each method. These simplifications are shown in the Supplementary material.

Type-I error: association model and replicates. Instead of using the total number of gene region SNPs in the Bonferroni correction to control for multiple testing, the M_{eff} can be used instead. While using the M_{eff} in the Bonferroni correction is less conservative than using the total number of SNPs, the resulting type-I error may not be equal to the chosen significance level. Therefore, we performed a simulation study to investigate the ability of the M_{eff} and extreme tail theory approaches to retain the specified type-I error rate. For each gene region SNP, we tested the null hypothesis that the SNP, under an additive genetic model, is not associated with case status using a LRT statistic from a logistic regression. In addition, for each simulation scenario, we calculated the type-I error rate of the traditional Bonferroni correction using the total number of gene region SNPs.

To simulate case status for each replicate and simulation type, we randomly paired 4000 haplotypes from a population (discussed below) to create 2000 subjects, randomly labeling one half as cases and the other half as controls. We used 10 000 replicates to evaluate the type-I error rate.

Gene region simulations. To gain further understanding of each method's performance over a wide-variety of gene regions including many rare variants, we applied all methods on simulated data from the 10 gene regions from the HapMap ENCODE resequencing and genotyping project. For each region, we simulated 1000 cases and 1000 controls for 10 000 replicates using Hapgen[10,11] from an initial sample of the 112 CEU haplotypes from the 1000 Genomes Pilot data for which phased haplotype data was available at the time of this publication. Hapgen introduces variation in the gene region between haplotypes while still retaining the general LD and minor allele frequency (MAF) characteristics of the gene region. Basic information about the 10 gene regions is provided in Table 1.

As Conneely *et al*'s approach is computer intensive as the number of marker's increases, we broke up each of the Encode gene regions into two subregions. We visually choose the split locations to have the smallest amount of correlation between the two subregions. We then calculated the extreme tail theory adjusted P-value for each subregion and used a Bonferroni correction, choosing the minimum of the two P-values and multiplying by two for the extreme tail theory P-value for the entire region. We calculated the effective number of independent SNPs for each entire gene region, as well as for each subregion so as to better compare the methods with extreme tail theory. We used a logistic regression model to detect the marginal association of each SNP in the region with case status. We chose the SNP with the lowest P-value and adjusted for multiple comparisons within each gene region using the M_{eff} methods or extreme tail theory. We compared the type-I error rate of the M_{eff} methods and the extreme tail theory method, overall 10 gene regions, and repeated the analysis after removing all SNPs with MAF < 5%.

RESULTS

Simple examples: known number of independent SNPs

Cheverud's method overestimated the M_{eff}, often to a large extent, both when the number of independent blocks was varied while the number of SNPs within each block stayed the same and when there were unequal groupings of SNPs within the independent blocks

Table 1 Number of SNPs in 1000 genome simulations

				No. of SNPs			
	Name	*Genomic region*	*Split*	*MAF < 1%*	*1% ≤ MAF ≤ 5%*	*MAF > 5%*	*Total*
1	2p16.3	Chr2	1	3	124	493	620
		51512208–52012208	2	7	223	764	994
			Total	10	347	1257	1614
2	2q37.1	Chr2	1	12	147	651	810
		234156563–234656627	2	7	173	812	992
			Total	19	320	1463	1802
3	4q26	Chr4	1	24	260	717	999
		118466103–118966103	2	9	112	724	845
			Total	33	372	1441	1844
4	7p15.2	Chr7	1	28	146	370	542
		26924045–27424045	2	20	101	465	586
			Total	48	247	835	1128
5	7q21.13	Chr7	1	15	150	592	755
		89621624–90121624	2	13	237	536	786
			Total	28	387	1128	1541
6	7q31.33	Chr7	1	5	188	485	676
		126368183–126865324	2	11	154	572	737
			Total	16	342	1057	1413
7	8q24.11	Chr8	1	24	171	493	686
		118882220–119382220	2	22	90	375	487
			Total	46	261	868	1173
8	9q34.11	Chr9	1	78	131	712	919
		130725122–131225122	2	3	105	254	362
			Total	81	236	966	1281
9	12q12	Chr12	1	60	139	706	903
		38626477–39126476	2	10	199	683	892
			Total	70	338	1389	1795
10	18q12.1	Chr18	1	27	130	476	631
		23719231–24219231	2	20	66	503	589
			Total	47	196	979	1220

(Figure 1). This inflation was predicted by the simplifications of the equations (supplementary material), as well as by Nyholt who has stated that the method is overly conservative when there is strong LD. Nyholt thus recommends removing all redundant SNPs (ie SNPs with an $r^2 = 1$) before calculating the M_{eff}. Nonetheless, it is useful to see the extent to which the method overestimates the M_{eff} in these examples, especially as the other methods produced M_{eff} that were much closer to the number of independent SNPs.

As can be observed in Figure 1, Galwey's method accurately estimated the M_{eff} as the number of equally sized blocks varied, but underestimated the M_{eff} when the block groups were of unequal size. Li and Ji's method accurately estimated the M_{eff} in both scenarios. Finally, Gao *et al*'s method estimated the number of independent SNPs in most scenarios, but slightly underestimated the number of independent SNPs when the total number of SNPs was large (data not shown). These results were further supported by the mathematical simplifications of the M_{eff} formulas (supplementary section).

Simulations

As seen in Figure 2 and in Supplementary Tables S1 and S2, the results were consistent across all 10 gene regions. Galwey's method and Li and Ji's method both had inflated type-I error, the Bonferroni adjustment and Cheverud method had deflated type-I error, and Gao *et al*'s method and extreme tail theory produced type-I error closest to the true level of 0.05. As suggested previously by Gao *et al* in another paper,[13] we saw type-I error rates closer to 0.05 for Gao *et al*'s method, as well as the other M_{eff} methods when each gene region was split into two to calculate M_{eff}. Finally, most methods produced type-I error rates closer to 0.05 when only common SNPs (MAF > 0.05) were included in the analysis. Overall, Gao *et al*'s method and extreme tail theory produced type-I error rates close or slightly above the true level for the simulation scenarios containing all variants. The type-I error of both methods improved when only including common SNPs with Gao *et al*'s method being slightly below the true level of 0.05 for some gene regions.

Run time

The time needed to calculate M_{eff} or extreme tail theory increased as the number of SNPs within each region increased. Extreme tail theory was slightly slower than the M_{eff} methods (Supplementary Table S3).

DISCUSSION

Although we compared these methods using a logistic regression with a dichotomous trait, we expect similar performance for continuous traits as long as the model assumptions are met.

Here, we use $c = 0.995$ as Gao *et al* suggest. Changing c will most often change the M_{eff} calculated for a particular gene region. Smaller values of c will require fewer eigenvalues to reach the threshold and

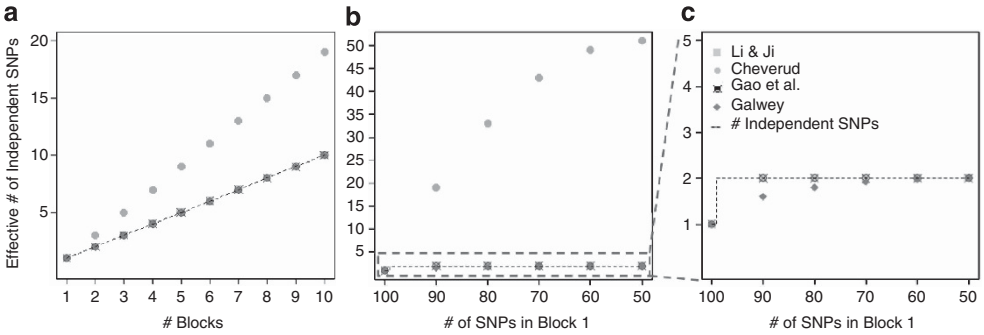

Figure 1 (**a**) M_{eff} for different numbers of two SNP blocks. (**b**) M_{eff} for a varying number of SNPs within two unequal blocks: full plot; (**c**) zoomed.

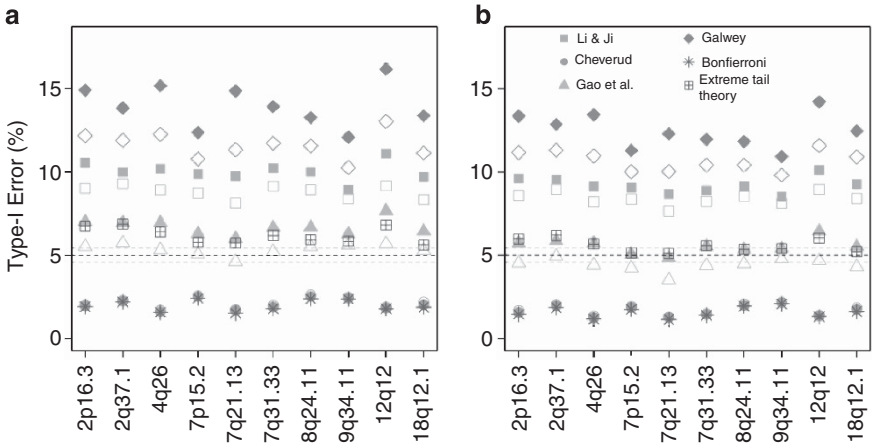

Figure 2 Type-I error rates across 10 gene regions simulated with 1000 Genomes. Filled symbols represent M_{eff} calculated over the entire gene region while open symbols represent M_{eff} calculated on and then combined from the two subregions within each gene region. Black dotted line is 5% significance threshold; gray dotted lines represent the 95% interval around the 5% significance threshold. (**a**) All SNPs; (**b**) SNPs with MAF > 5%.

thus M_{eff} will be smaller while larger values of c will require more eigenvalues to reach the cutoff resulting in a larger M_{eff}.

Most studies now implement imputation based on the HapMap samples or the 1000 Genomes Project. We expect the studies that use imputed data would perform similarly to results shown here. Further, Gao showed that his method performed well in imputed data.[13]

As the number of markers for which we adjust increases to the number of GWAS markers on a chromosome or in the genome, we expect the performance of the methods to decline. This was seen in the 10 gene region simulations using the 1000 Genomes data in which each gene region had between 1000 and 2000 SNPs (Table 1). Although the M_{eff} methods do not implicitly have a computational limit on the number of SNPs, the methods appear to perform better when the large regions are broken into smaller units (Figure 2). Further, as the region size increases, the genotypic correlation matrix used to estimate the correlation of the statistics in extreme tail theory and the eigenvalues for the M_{eff} methods will start to pick up some macro level correlation, such as that due to population stratification, in addition to the gene level correlation. Thus, the methods will be less able to accurately adjust for the correlation in the gene regions. One solution is to break-up the chromosome or the genome into manageable units similar to those used here, such as gene regions or 1 Mb sections. Others have explored using M_{eff} methods on a genome-wide scale and find using smaller units to be both computationally practical and effective at retaining the appropriate type-I error.[8,13] We expect that adjusting for multiple comparisons using broken up sections would perform very similarly to the results presented in Figure 2.

Recently, researchers have begun conducting association studies using rare variants from the sequence data. Our analysis using 10 gene regions simulated using the 1000 Genomes data is a good example of what may occur when rare variants are included in the analysis. Although we do see a slightly improved performance when only common variants are included in the analysis, it appears that including rare variants in the analysis is still possible without too much change in the overall results for the adjustment for multiple comparisons. However, it is worth noting that the sample size needed to detect association between a SNP and case status at a genome-wide level is prohibitively high (in the 100 000s) for rare variants with a moderate effect.[14] Thus, many researchers are using methods to analyze rare variants that collapse or consider multiple variants together, so that each gene or region is treated as a single variable.[12,15,16] These genes or regions could thus be used as the unit of measure for the M_{eff} methods or the extreme tail theory method. More research is needed to compare the performance in this particular scenario.

We found that extreme tail theory and Gao *et al* produce a type-I error rate close to or at the chosen significance level. Thus, we recommend using either extreme tail theory or Gao *et al*'s method to control for multiple testing in gene regions when the gold standard of permutation is not feasible.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

A portion of this research was conducted using the Linux Clusters for Genetic Analysis (LinGA) computing resource funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center and by contributions from individual investigators. We would like to thank Mayetri Gupta for her thoughtful comments on this research.

1 Cheverud JM: A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 2001; **87**: 52–58.
2 Galwey NW: A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genet Epidemiol* 2009; **33**: 559–568.
3 Gao X, Starmer J, Martin ER: A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* 2008; **32**: 361–369.
4 Li J, Ji L: Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 2005; **95**: 221–227.
5 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.
6 Conneely KN, Boehnke M: So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet* 2007; **81**: 1158–1168.
7 Salyakina D, Seaman SR, Browning BL, Dudbridge F, Muller-Myhsok B: Evaluation of Nyholt's procedure for multiple testing correction. *Hum Hered* 2005; **60**: 19–25, (discussion 61–12).
8 Moskvina V, Schmidt KM: On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* 2008; **32**: 567–573.
9 Team RDC: *R: A Language and Environment for statistical Computing*. Vienna, Austria: R Foundation for statistical Computing, 2009.
10 Spencer CC, Su Z, Donnelly P, Marchini J: Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009; **5**: e1000477.
11 Li N, Stephens M: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 2003; **165**: 2213–2233.
12 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89**: 82–93.
13 Gao X, Becker LC, Becker DM, Starmer JD, Province MA: Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol* 2010; **34**: 100–105.
14 Bansal V, Libiger O, Torkamani A, Schork NJ: Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010; **11**: 773–785.
15 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.
16 Price AL, Kryukov GV, de Bakker PI *et al*: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86**: 832–838.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)