Am. J. Hum. Genet. 74:765–769, 2004

Report

A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other

Dale R. Nyholt

Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia

In this report, we describe a simple correction for multiple testing of single-nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD) with each other, on the basis of the spectral decomposition (SpD) of matrices of pairwise LD between SNPs. This method provides a useful alternative to more computationally intensive permutation tests. A user-friendly interface (SNPSpD) for performing this correction is available online (http://genepi.qimr.edu.au/general/daleN/SNPSpD/). Additionally, output from SNPSpD includes eigenvalues, principal-component coefficients, and factor "loadings" after varimax rotation, enabling the selection of a subset of SNPs that optimize the information in a genomic region.

SNPs in disease-related genes are increasingly being used as candidates in the search for causative variations. Both theoretical (Long and Langley 1999; Service et al. 1999; Zollner and von Haeseler 2000; Akey et al. 2001; Bader 2001; Morris and Kaplan 2002) and empirical studies (Clark et al. 1998; Terwilliger and Weiss 1998; Escamilla et al. 1999; Martin et al. 2000) have produced contradictory results on whether haplotypes of two or more SNPs provide greater power than individual SNPs to find useful linkage disequilibrium (LD) between a causative mutation and linked marker loci. Moreover, variability in LD across the genome, the large dependence of the strength of association on allele-frequency differences between the disease variant and the SNP (e.g., Ohashi and Tokunaga 2001), and questions regarding the suitability of the "common disease common variant" (CDCV) hypothesis (i.e., depending on the ascertainment method) (Pritchard and Cox 2002) all suggest that an initial investigation of a candidate gene or interval should test many SNPs individually for association. However, unless the selected SNPs are all in complete LD with each other, such multiple testing will increase the false-positive (type I error) rate under nominal significance thresholds (e.g., $\alpha = 0.05$). On the other hand, when background LD exists between SNPs but they are assumed to be completely independent, then the Šidák correction—which is approximated by the popular Bonferroni correction (Šidák 1968, 1971)—would markedly overcorrect for the inflated false-positive rate, resulting in a reduction in power. Here we describe a simple correction for multiple testing of SNPs in LD with each other, on the basis of the spectral decomposition (SpD) of matrices of pairwise LD between SNPs. This method provides a useful alternative to more computationally intensive permutation tests.

It has previously been shown that the collective correlation among a set of variables can be measured by the variance of the eigenvalues (λs) derived from a correlation matrix (e.g., Cheverud et al. 1983, 2001). As detailed by Cheverud (2001), high correlation among variables leads to high λs. For example, if all variables are completely correlated, the first λ equals the number of variables in the correlation matrix (M) and the rest of the λs are zero. In this case, the variance of the λs is at its maximum, and it is equal to the number of variables in the matrix. Conversely, if no correlation exists among variables, all of the λs will be equal to one, and the set of λs will have no variance. Hence, the variance of the λs will range between zero, when all the variables are independent, and M, where M is the total number of variables included in the matrix. Therefore, the ratio of observed eigenvalue variance, $\mathrm{Var}(\lambda_{obs})$, to its maximum ($M$) gives the proportional reduction in the num-

Received November 5, 2003; accepted for publication January 29, 2004; electronically published March 2, 2004.

Address for correspondence and reprints: Dr. Dale R. Nyholt, Queensland Institute of Medical Research, Post Office Royal Brisbane Hospital, Brisbane QLD 4029, Australia. E-mail: daleN@qimr.edu.au

0002-9297/2004/7404-0018$15.00

Table 1

LD Matrix with Pairwise Correlations (Δ) and Eigenvalues (λs)

Locus	Distance from T5991C (bp)	Pairwise Correlation									
		T5991C	A5466C	T3892C	A240T	T93C	T1237C	G2215A	I/D	G2350A	4656(CT)$_{3/2}$
T5991C	0	***7.84***									
A5466C	25	.99	***1.60***								
T3892C	1,599	.83	.82	***.24***							
A240T	5,251	.99	.98	.82	***.21***						
T93C	5,398	.99	.98	.82	1.00	***.08***					
T1237C	10,979	.59	.58	.69	.61	.61	***.03***				
G2215A	15,108	−.61	−.60	−.71	−.63	−.63	−.86	***.01***			
I/D	16,945	.61	.60	.71	.63	.63	.86	−1.00	***.00***		
G2350A	17,372	.61	.60	.71	.63	.63	.86	−1.00	1.00	***.00***	
4656(CT)$_{3/2}$	26,796	.58	.57	.69	.60	.60	.81	−.94	.94	.94	***.00***

Note.—Pairwise correlations (Δ) are given below the diagonal, and the 10 eigenvalues (λs) associated with this matrix are given along the diagonal (*bold*, *italic*).

ber of variables in a set, and the effective number of variables (M_{eff}) may be calculated as follows:

$$M_{eff} = 1 + (M - 1)\left(1 - \frac{\mathrm{Var}(\lambda_{obs})}{M}\right).$$

The common LD measure Δ is also the correlation coefficient for a 2 × 2 table (Hill and Robertson 1968), where

$$\Delta = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})^{1/2}}$$

and the notation for estimated haplotype and marker allele frequencies in the 2 × 2 table is as follows:

	SNP 2		
SNP 1	Allele 1	Allele 2	Total
Allele 1	π_{11}	π_{12}	π_{1+}
Allele 2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Consequently, λs for the LD correlation (Δ) matrix may be calculated by principal-components analysis or, more generally, by spectral decomposition (SpD), and the approach of Cheverud (2001) may be applied to obtain the effective number of independent SNPs (M_{eff}) represented in the matrix.

Although M_{eff} could easily be calculated using standard statistical packages and/or free software in the public domain, we developed a user-friendly Web interface (SNPSpD) because we believe a wide variety of researchers may have use for this approach, which simply requires users to upload a MERLIN-format pedigree and map file (Abecasis et al. 2002). The uploaded files are run through a slightly altered version of Gonçalo Abecasis's LDMAX program—part of the GOLD Command Line Tools package [gold-1.1.0.tar.gz] (Abecasis and Cookson 2000)—which uses the expectation-maximization–based approach of Excoffier and Slatkin (1995) to estimate haplotype frequencies in case-control or family data. Using these haplotype frequencies, LDMAX calculates a number of pairwise LD statistics. A Perl script then creates a matrix of pairwise Δ measures, from which SNPSpD calculates λs by SpD, by use of the EIGEN function of R (v1.7.1) (R Development Core Team 2003). SNPSpD output includes the matrix of SNP-SNP Δ measures, M, λs, Var (λ_{obs}), M_{eff}, and a Šidák-corrected significance threshold (for M_{eff} tests) required to keep the type I error rate at 5%.

To investigate the performance of the M_{eff}-Šidák correction we utilized two real data sets. The first data set consisted of 10 highly associated SNPs, spanning ~27 kb within the *angiotensin-I converting enzyme* (*ACE*) gene (Keavney et al. 1998), and the second data set consisted of 23 SNPs, spanning ~794 kb within the T-cell antigen receptor (TCR) α/δ locus (Moffatt et al. 2000). The results of SNPSpD were validated by permutation (e.g., Westfall and Young 1993).

For the Keavney data set, 88 founders were utilized. For each permutation, 44 founders were randomly selected (without replacement) and labeled "cases," and the remaining 44 founders were labeled "controls." This selection process maintained each founder's haplotype and, hence, the LD information between each SNP. For each permuted case-control sample (replicate), a χ^2 test of homogeneity was used to compare genotype frequencies between the permuted case and control populations for each SNP. Thus, for each replicate, a total of 10 χ^2 values were produced. This process was repeated 50,000 times. Finally, the number of replicates in which at least

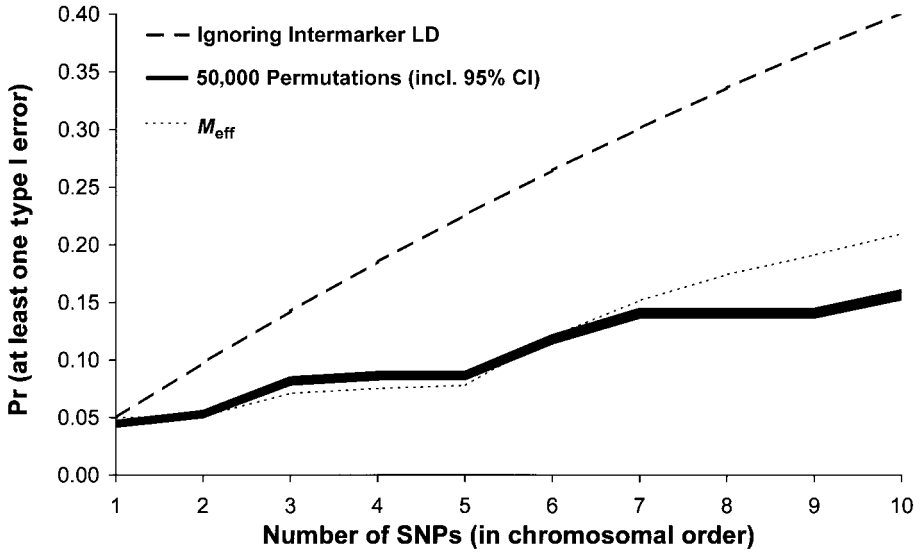

Figure 1 Probability (Pr) of a type I error plotted against the number of SNPs (in chromosomal order) tested in the Keavney et al. (1998) data. The graph shows the expected increase in the false-positive rate for completely independent SNPs [i.e., $1 - (1 - \alpha)^M$] (*thick dashed line*), by use of an M_{eff}-Šidák correction [i.e., $1 - (1 - \alpha)^{M_{eff}}$] (*thin dashed line*), and from 50,000 permutations (*thick solid line*), for $\alpha = 0.05$.

one SNP had a χ^2 value with $P \leq .05$ [i.e., $\chi^2 \geq 5.991476$; df 2] were counted to estimate the probability of a type I error. For example, the number of replicates producing at least one χ^2 value ≥ 5.991476 were 2,235, 2,655, 4,100, 4,328, 4,328, 5,909, 7,042, 7,042, 7,042, and 7,844 for the first 1, 2, 3, 4, 5, 6, 7, 8, 9, and all 10 SNPs (in chromosomal order), respectively.

Permutations were performed in R, utilizing the SAMPLE and CHISQ.TEST functions. Permuting 50,000 replicates took 26 min for the Keavney data set and 62 min for the Moffatt data set, whereas our SNPSpD Web interface took only 12 s and 14 s, respectively. Considering the fact that the R permutations were performed on a 2.8 GHz Xeon (Linux v2.4.20) server with exclusive CPU use, whereas the SNPSpD interface was run on our 300 MHz Sun4 SPARC 10 (SunOS 5.8) Web server, the SNPSpD approach was well over 100 times faster than the R permutations.

The 10 SNPs in the Keavney data set produced an M_{eff} of 4.59, representative of high intermarker LD (see table 1). Figure 1 shows the probability (Pr) of a type I error plotted against the number of SNPs tested for the Keavney et al. (1998) data set. Compared with the permuted rate, a Šidák correction ignoring intermarker LD (standard-Šidák correction) would clearly overcorrect for the inflated type I error rate, whereas the M_{eff}-Šidák rate, although slightly conservative in the presence of higher order intermarker LD (i.e., very strong LD across >2 SNPs) provides a good approximation to the permuted rate. For example, in terms of the significance threshold required to keep the type I error rate at 5% if all 10 SNPs were individually tested for association with ACE levels, the standard-Šidák [i.e., $1 - (1 - \alpha)^{1/M}$], M_{eff}-Šidák [i.e., $1 - (1 - \alpha)^{1/M_{eff}}$], and permutation-based corrections would specify thresholds of $P \leq .005$, $P \leq .011$, and $P \leq .015$, respectively.

Analysis of the 23 SNPs in the Moffatt data set indicated low levels of intermarker LD with an M_{eff} of 22.53 and resulted in thresholds to keep the type I error rate at 5% of $P \leq .0022$, $P \leq .0023$, and $P \leq .0028$ for the standard-Šidák, M_{eff}-Šidák, and permutation-based corrections, respectively.

It is worth noting that >50,000 permutations would be required to avoid rounding highly significant P values ($P < .00002$). Consequently, to correct for multiple testing of SNPs in LD with each other, our SNPSpD approach provides a simple and useful alternative to more computationally intensive permutation tests. Furthermore, by providing an estimate of the number of independent tests (M_{eff}), the SNPSpD approach allows researchers to apply any flavor of multiplicity correction they prefer—for example, the modified Bonferroni procedures of Holm (1979), Hochberg (1988), and Hommel (1988) or the more recently proposed false-discovery-rate (FDR) approach of Benjamini and Hochberg (1995).

Coincidentally, during the preparation of this manuscript, Meng et al. (2003) described a method based on the SpD of matrices of pairwise LD between markers to select a subset of SNPs that optimize the information in a genomic region. Although there are some parallels be-

tween the approach of Meng et al. (2003) and that presented here, our study, unlike that of Meng et al. (2003), not only is primarily concerned with the correction for multiple testing when using multiple SNPs in LD with each other but also provides important validation of the use of an SpD-based approach to correct for such nonindependence. That said, to complete the usefulness of our SNPSpD interface, we have extended analyses to include results after varimax rotation. Specifically, we report λs, proportions of variance, and principal-component coefficients after varimax rotation (an orthogonal rotation method that minimizes the number of variables that have high loadings on each factor, thus simplifying the interpretation of the factors). Furthermore, we maximize interpretability of these results by flagging the SNP(s) contributing the *most* to each rotated factor (i.e., group of SNPs). These flagged SNPs may be viewed as "haplotype-tagging SNPs." Indeed, even in data with strong LD, the rotated factors correspond well with haplotypes obtained via traditional methods. For example, the seven haplotypes reported in the Keavney et al. (1998) study correspond to the seven factors produced by SNPSpD after varimax rotation.

Finally, because the user may then easily select SNPs to represent either each factor, the factor(s) with the largest M_{eff} λs, or the factor(s) explaining a selected proportion of variance, we believe many researchers will appreciate the convenience of our SNPSpD Web interface.

Acknowledgments

The author thanks Professor Nicholas G. Martin, Dr. Grant W. Montgomery, and, in particular, Dr. David L. Duffy, for many helpful discussions, and Dr. Martin Farrall, for generously sharing the ACE data set. Special thanks go to David C. Smyth for assisting with the development of the SNPSpD Web interface. This research was supported in part by a National Health and Medical Research Council (NHMRC) Peter Doherty Fellowship and an NHMRC (Australia) grant 241916.

Electronic-Database Information

The URLs for data presented herein are as follows:

SNP Spectral Decomposition (SNPSpD) Web Interface, http://genepi.qimr.edu.au/general/daleN/SNPSpD/

LDMAX Program (Part of the GOLD Command Line Tools Package [gold-1.1.0.tar.gz]), http://www.sph.umich.edu/csg/abecasis/GOLD/download/index.html

TCR α/δ Locus Data of Moffatt et al. (2000), http://www.well.ox.ac.uk/asthma/public/TCR/index.shtml

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Abecasis GR, Cookson WO (2000) GOLD: graphical overview of linkage disequilibrium. Bioinformatics 16:182–183

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291–300

Bader JS (2001) The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics 2: 11–24

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc [Ser B] 57:289–300

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87: 52–58

Cheverud J, Rutledge J, Atchley W (1983) Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. Evolution 37:895–905

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612

Escamilla MA, McInnes LA, Spesny M, Reus VI, Service SK, Shimayoshi N, Tyler DJ, Silva S, Molina J, Gallegos A, Meza L, Cruz ML, Batki S, Vinogradov S, Neylan T, Nguyen JB, Fournier E, Araya C, Barondes SH, Leon P, Sandkuijl LA, Freimer NB (1999) Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. Am J Hum Genet 64:1670–1678

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Statist 6:65–70

Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75:383–386

Keavney B, McKenzie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. Hum Mol Genet 7:1745–1751

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM (2000) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet 67:383–394

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG (2003) Selection of genetic markers for association analyses, using

linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130

Moffatt MF, Traherne JA, Abecasis GR, Cookson WO (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR α/δ locus. Hum Mol Genet 9:1011–1019

Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233

Ohashi J, Tokunaga K (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J Hum Genet 46:478–482

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant…or not? Hum Mol Genet 11:2417–2423

R Development Core Team (2003) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, http://www.R-project.org (accessed March 1, 2004)

Service SK, Lang DW, Freimer NB, Sandkuijl LA (1999) Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728–1738

Šidák Z (1968) On multivariate normal probabilities of rectangles: their dependence on correlations. Ann Math Statist 39:1425–1434

——— (1971) On probabilities of rectangles in multivariate normal Student distributions: their dependence on correlations. Ann Math Statist 41:169–175

Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 9:578–594

Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. John Wiley & Sons, New York

Zollner S, von Haeseler A (2000) A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am J Hum Genet 66:615–628