British Journal of Medicine & Medical Research
4(6): 1413-1422, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

A Simple Scalable Association Hypothesis Test Combining Gene-wide Evidence from Multiple Polymorphisms

Dhananjay Vaidya[1*], Lisa R. Yanek[1], Rasika A. Mathias[1], Taryn F. Moy[1], Diane M. Becker[1] and Lewis C. Becker[1]

[1]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA.

Authors' contributions

This work was carried out in collaboration between all authors. Author DV designed the study, performed statistical analyses and wrote the first draft of the manuscript. Author LRY maintained the study database and contributed to the data analysis. Author TFM managed the participant recruitment and measurements in the study. Author RAM made critical revisions to the manuscript. Authors DMB and LCB obtained funding or the study and made critical revisions to the manuscript. All authors read and approved the final manuscript.

Original Research Article

Received 24th July 2013
Accepted 20th September 2013
Published 10th December 2013

ABSTRACT

Aims: In single-nucleotide polymorphism (SNP) scans, SNP-phenotype association hypotheses are tested, however there is biological interpretation only for genes that span multiple SNPs. We demonstrate and validate a method of combining gene-wide evidence using data for high-density lipoprotein cholesterol (HDLC).
Methodology: In a family based study (N=1782 from 482 families), we used 1000 phenotype-permuted datasets to determine the correlation of z-test statistics for 592 SNP-HDLC association tests comprising 14 genes previously reported to be associated with HDLC. We generated gene-wide p-values using the distribution of the sum of correlated z-statistics.
Results: Of the 14 genes, *CETP* was significant ($p=4.0\times10^{-5}$ <0.05/14), while *PLTP* was significant at the borderline ($p=6.7\times10^{-3}$ <0.1/14). These p-values were confirmed using empirical distributions of the sum of χ^2 association statistics as a gold standard (2.9×10^{-6} and 1.8×10^{-3}, respectively). Genewide p-values were more significant than Bonferroni-corrected p-value for the most significant SNP in 11 of 14 genes (p=0.023). Genewide p-

**Corresponding author: Email: dvaidya1@jhmi.edu;*

values calculated from SNP correlations derived for 20 simulated normally distributed phenotypes reproduced those derived from the 1000 phenotype-permuted datasets were correlated with the empirical distributions (Spearman correlation = 0.92 for both).
Conclusion: We have validated a simple scalable method to combine polymorphism-level evidence into gene-wide statistical evidence. High-throughput gene-wide hypothesis tests may be used in biologically interpretable genomewide association scans. Genewide association tests may be used to meaningfully replicate findings in populations with different linkage disequilibrium structure, when SNP-level replication is not expected.

Keywords: Bonferroni; hypothesis tests; combining evidence.

1. INTRODUCTION

Genomewide scans of DNA polymorphisms that may not be functional, but may be linked to functional genetic differences are often used to study the association genes with phenotypes [1]. Although interpretable biological hypotheses can typically be formulated regarding whole genes or functional gene-domains, hypothesis tests are currently reported for individual polymorphisms [1]. The inspection of multiple hypothesis tests is appropriately recognized as a problem, currently resolved by using more stringent thresholds for significance as suggested by Bonferroni [2], i.e., corrected threshold = threshold/number of tests; alternatively: corrected p-value = calculated p-value*number of tests. Unfortunately, Bonferroni's calculation for independent hypothesis tests is overly conservative for multiple polymorphisms within the same gene, which are expected to be in linkage disequilibrium. Instead of the Bonferroni correction, several methods have been proposed of applying sliding penalties to association statistics ranked from most significant to least significant [3-6]. After any of these corrections, the significant hypotheses are still regarding specific polymorphisms, rather than evidence for the association for the gene or functional genetic domain. Such a finding cannot be replicated in a different population with a different structure of linkage disequilibrium, different polymorphisms and different haplotypes even though an association of that gene with the phenotype has the same biological relevance in the other population.

Another set of techniques used to combine association statistics is the summation of statistics. This summation may be on the logarithmic scale, e.g., in the case of Fisher's product of p-values [7], the logarithm of which is distributed as χ^2 when appropriately scaled, or may be on the original scale [8]. However, when the summed statistics are correlated, the distribution is not analytically known, and must be empirically derived using a large enough number of permutations [9]. Wille et al. [10] have suggested a method of summing association statistics in order of statistical significance, deflating the remaining marker-level statistic at every step, based on their correlation with the statistics for the markers already summed. However, the asymptotic normality of association statistics is assumed, and this is generally not present at the first few steps, which include the most significant results.

Therefore, we demonstrate and validate a method to combine polymorphism-level hypotheses test statistics into a single gene-wide sum statistic testing of the hypothesis of association, which has the potential to obviate the above problems.

2. METHODOLOGY

2.1 Mathematical Basis for the Calculation of Combined P-values

Consider a number of hypothesis-test p-values p_i that are to be combined into a single p-value. For valid statistical tests, the probit transformation of p-value is a standard normal variate:

$\text{Probit}(p_i) = Z_i$

We use the theorem [12]: if $(Z_1, Z_2, ..., Z_k) \sim N(0, \sum_{kxk})$, then $SUM(Z_i) \sim N[0, SUM(\sum_{kxk})]$, where $\sum_{kxk}$ is the variance-covariance matrix, and "SUM" is the sum of the elements of the matrix.

We thus obtain a single standard normal variate that pertains to all of the hypothesis tests that need to be combined, and this is converted back into a p-value.

The variance-covariance matrix (which is equivalent to the correlation matrix for standard normal variates) is obtained by estimating the association of the polymorphisms on randomly permuted datasets. The permuted phenotypes also serve to test null hypotheses.

Note regarding a simplifying assumption: The correlation between one-sided test-statistics of the genotype-phenotype correlation is expected for SNPs in linkage disequilibrium. However, p-values for hypothesis tests are two-sided, because either tail of the hypothesis represents a biologically relevant genotype-phenotype association.

2.2 Demonstration Hypotheses and Dataset

We used 14 genes previously shown to be associated with HDL cholesterol (HDLC) levels summarized in a review article by Pirruccello and Kathiresan [13] as our demonstration hypotheses: *ABCA1, ANGPTL4, APOA145C3, CETP, FADS1-2-3, GALNT2, HNF4A, LCAT, LIPC, LIPG, LPL, MVK, PLTP, TTC39B*. The hypotheses are of the form "polymorphisms within the gene are associated with HDLC levels". For comparability we also tested the hypotheses "the most significant reported SNP by Pirruccello and Kathiresan in the gene was associated with HDLC levels". If the SNP was not genotyped we used the imputed genotype score using MACH [14].

To test these hypotheses, we used data from European-Americans enrolled in GeneSTAR, a study of families identified from probands (N=482) who were admitted with documented coronary artery disease (CAD including myocardial infarction or angina with angiographically proven stenoses >50%, or that required percutaneous intervention or bypass graft surgery). This study was approved by the Institutional Review Board, and all research participants gave informed consent. Siblings of probands, offspring of the siblings, and co-parents of the offspring, all of whom were apparently healthy were enrolled into the study (N=1782). The individuals have been genotyped using the Illumina Human 1M chip. Only genotyped (non-imputed) SNPs with sample minor allele frequency > 5%, annotated to be within the 14 hypothesized genes were included for analysis.

Additive SNP associations with fasting serum HDLC were estimated using a mixed model with family as the random effect, with a sex-SNP interaction and main effect, including age, smoking, statin use and two population stratification adjustment variables obtained using

EIGENSTRAT [15] as covariates. A 2-degree of freedom chi-squared test simultaneously testing the SNP main effect and the sex-SNP interaction was considered as the hypothesis test.

For determining the correlation between p-values, phenotype-permuted datasets were generated by permuting measurements within families, and permuting measurements between families of the same size. This maintains the underlying heritability of the phenotype, but breaks any genotype-phenotype correlations. For the purpose of this analysis, the number of permuted datasets was 1000. Using Fisher's z-transform of the correlation coefficient, this allows a minimum precision of ±0.06 in the estimate of the correlation coefficient at 95% confidence. Estimates of correlation coefficients were also made using 20 simulated normally distributed phenotypes, these have a minimum precision of ±0.44 at 95% confidence for truly uncorrelated variables. However, if the true correlation is higher the 95% precision is also much better (e.g., for a true population correlation of 0.9, the precision interval is 0.76 to 0.96).

The mathematical theorem and calculation of combined gene-wide p-value are described above. For comparison the Bonferroni-corrected p-value is calculated as (nominal p-value)*(number of tests). If the calculation results in a value >1, the Bonferroni-corrected p-value = 1.

3. RESULTS AND DISCUSSION

3.1 Results

Table 1 shows the association results for Pirruccello and Kathiresan's [13] lead SNP in each of the 14 genes, along with the replication-wide Bonferroni corrected p-value. Of 14 SNPs 3 were genotyped, the others were imputed. The lead SNP reported by Pirruccello and Kathiresan [13] in 3 genes, namely *CETP*, *HNF4A*, and *LIPC* are significant at the Bonferroni-corrected level in GeneSTAR, while *PLTP* is additionally significant at the nominal level.

Table 1. SNP with the most significant p-value in genes reported by Pirruccello and Kathiresan [13] and p-value for that SNP in the GeneSTAR study

Gene	lead SNP in [13]	SNP p-value in [13]	p-value GeneSTAR
ABCA1	rs1883025	1.00×10^{-9}	0.10
ANGPTL4	rs2967605	1.00×10-^{8}	0.56 (i)
APOA145C3	rs964184	1.00×10^{-12}	0.20 (i)
CETP	rs173539	4.00×10^{-75}	2.48×10^{-7} (i)
FADS1-2-3	rs174547	2.00×10^{-12}	0.51
GALNT2	rs4846914	4.00×10^{-8}	0.44 (i)
HNF4A	rs1800961	8.00×10^{-10}	4.5×10^{-4} (i)
LCAT	rs2271293	9.00×10^{-13}	0.20 (i)
LIPC	rs10468017	8.00×10^{-23}	2.6E-03 (i)
LIPG	rs4939883	7.00×10^{-15}	0.95 (i)
LPL	rs12678919	2.00×10^{-34}	0.21 (i)
MVK	rs2338104	1.00×10^{-10}	0.66 (i)
PLTP	rs7679	4.00×10^{-9}	0.014 (i)
TTC39B	rs471364	3.00×10^{-10}	1.00

(i) Imputed genotype in GeneSTAR.

Table 2 shows the number of genotyped SNPs, the minimum p-value for any genotyped SNP, the Bonferroni-corrected p-values correcting for SNPs in the gene, and the Bonferroni-corrected p-values for the whole gene-replication study (592 SNPs) are tabulated. Gene-wide p-values obtained using the correlated chi-2 towards greater significance (12/14, p=0.009 nonparametric sign rank test) than the within-gene Bonferroni-corrected minimum p-value.

Table 2. Most significant SNP and gene-wide association p-values in candidate genes related to the HDLC phenotype

Gene				Gene-wide p-values		
	# SNPs	Minimum p-value for any SNP	within-gene-minimum Bonferroni-p for SNP	Using permuted χ^2 distribution	Using MVN SUM z-statistic (1000 permutations)	Using MVN SUM z-statistic (20 permutations)
ABCA1	85	8.3×10^{-4}	0.070	0.23	0.36	0.36
ANGPTL4	4	0.035	0.14	0.084	0.083	0.075
APOA145C3	75	6.9×10^{-3}	0.52	0.21	0.30	0.31
CETP	25	3.9×10^{-7}	9.8×10-6	2.9×10^{-6}	3.97×10^{-5}	9.03×10^{-5}
FADS1-2-3	40	0.021	0.86	0.30	0.30	0.32
GALNT2	82	8.1×10^{-3}	0.67	0.30	0.24	0.23
HNF4A	49	8.0×10^{-3}	0.39	0.36	0.34	0.33
LCAT	5	0.079	0.40	0.14	0.12	0.13
LIPC	84	0.039	1	0.49	0.56	0.56
LIPG	16	0.13	1	0.61	0.69	0.68
LPL	27	0.032	0.86	0.30	0.25	0.30
MVK	7	0.22	1	0.57	0.54	0.54
PLTP	16	2.7×10^{-3}	0.044	0.012	0.007	0.004
TTC39B	77	3.4×10^{-3}	0.26	0.37	0.59	0.58

For both of the proposed simplified methods, using 1000 simulated phenotypes and 20 simulated phenotypes, respectively, to estimate the correlation matrices, the p-values obtained are rank-correlated with the gold standard empirical χ^2 p-values with a Spearman correlation coefficient of 0.92 ($p=4\times10^{-6}$), both simplified methods having a Spearman rank-correlation of 1.00 with each other. For the simplified method using 1000 simulated phenotypes, gene-wide p-values that were more significant than Bonferroni-corrected p-value for the most significant SNP in the 11 of 14 genes (p=0.023).

The null-p distributions of the 14 genes for 1000 simulated phenotypes are presented in the histograms in Fig. 1. The variance inflation factors (lambda) for the 14 genes ranged from 0.90 to 1.11.

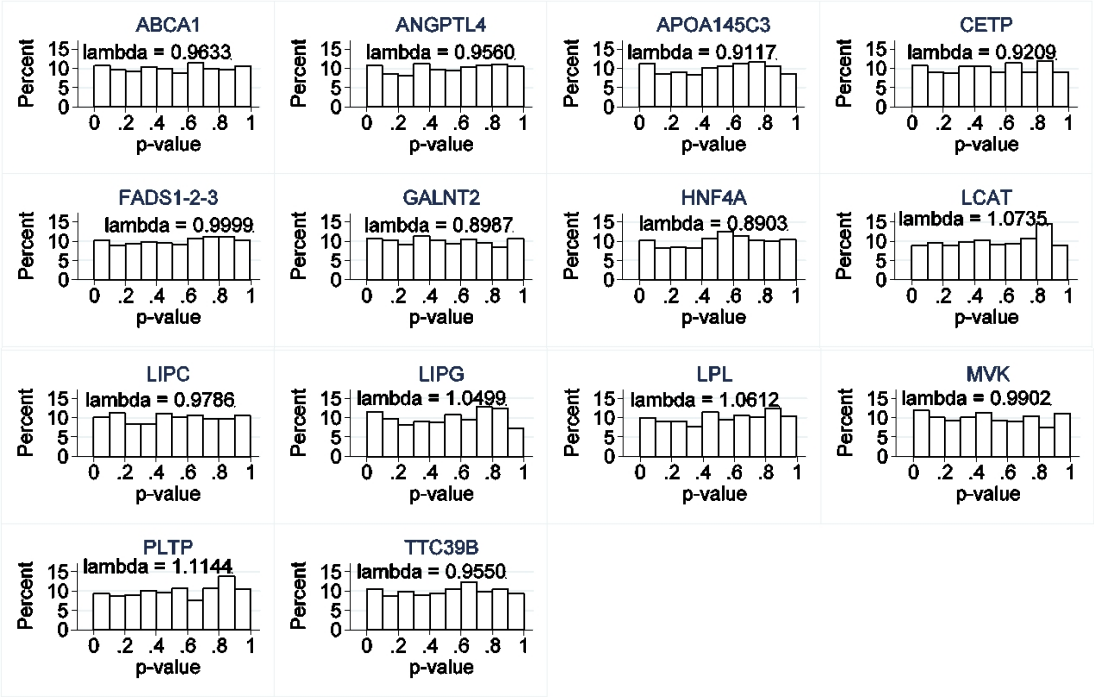

Fig. 1. Gene-wide p-value histograms for null-hypothesis tests and variance inflation factors (lambda) for the 14 genes.

The gene-wide p-values where correlation was estimated using only 20 simulated normally distributed phenotypes are plotted against the permutation estimated p-values in Fig. 2. The Spearman rank correlation of these p-values with the permutation test p-values is 1.0, and the p-values are also numerically quite close together, lying over the line of identity.

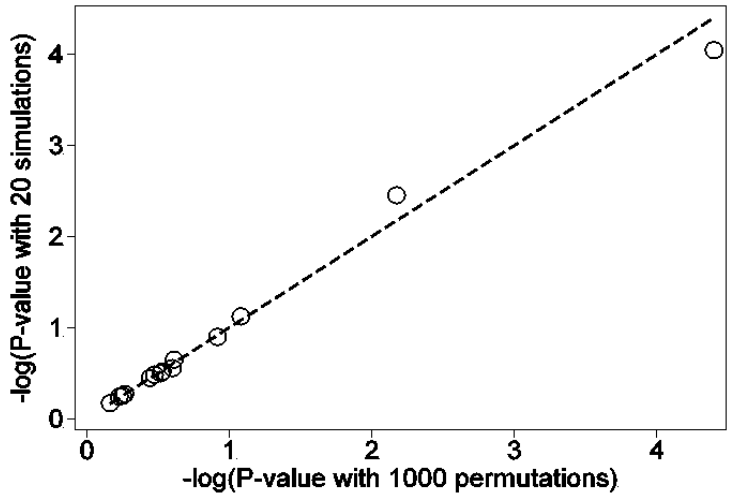

Fig. 2. Correlation of gene-wide p-values derived from 20 simulated normally distributed phenotypes vs. permutation test with 1000 permutations for the 14 genes.

3.2 Discussion

We have demonstrated a method for the calculation of a more interpretable gene-wide p-value using the theorem regarding the calculation of correlated p-values. We have shown approximate validity of the calculation in terms of p-value distributions and variance inflation factors (lambda) in spite of a major simplification in the calculation, namely, the use of the distribution of the sum of correlated normally distributed z-variables, rather than the distribution of the sum of correlated chi-squared variables. We have shown that this analysis can be implemented even in complicated study sampling designs requiring mixed model analysis, using permutation tests or appropriate simulated phenotypes to determine the correlation in p-values.

If we do not hypothesize a particular direction for the SNP-phenotype association, a chi-squared statistic, which is large whether the association is inverse or direct, is calculated and compared against the chi-squared distribution. It is possible to summate χ^2 statistics, however in case of SNPs that are in linkage disequilibrium, the χ^2 statistics are expected to be correlated. This sum of χ^2variables cannot be compared against the distribution of uncorrelated χ^2. This distribution is quite difficult to compute [11], hence we use permutation of phenotype to generate the summated test statistic distribution. The number of permutations allowed the p-value to be determined to 3 significant digits. For all genes except *CETP*, 10,000 permutations were adequate, but for *CETP* 10,000,000 permutations were needed. This is used as the gold standard for comparison with the simplifying assumption where a multivariate normal distribution is used.

As a positive control for the method, we have demonstrated that gene-wide p-values are able to reproduce association with some genes where strong HDLC phenotype-SNP associations have been previously collated in a review article [13]. The greatest degree of reproducibility is seen for imputed SNPs (4 of 14 SNPs showed nominally significant p-values). This is presumably because the samples within the earlier report, as well as the replication samples are European-Americans, who have a reproducible linkage disequilibrium structure. This is not expected in non-European origin populations.

If genotyped SNPs were used, only SNPs within two genes meet the significance threshold for within-gene Bonferroni correction, and only one gene meet correction for all 592 SNPs in the 14 genes examined. Gene-wide p-values using our method also show two genes to be significantly associated, and one of these is significant after correction for 14 tests. However, the gene-wide p-values were typically lower that the gene-wide Bonferroni corrected p-values, suggesting greater power to detect true associations.

Further we have shown that using 20 simulated variables, the correlation structure of SNP association tests is adequately estimated, as compared to the computation intensive permutation tests. Because this correlation structure could be used for any phenotype, this simplification allows for massive increase in the throughput of the analyses.

Because of our probit transformation, our proposed sum statistic differs from that proposed by Wille et al. [10].

Though we have discussed our method using “gene” as the unit of biological interpretation, the method can be used to combine evidence from any collection of SNPs, a functional domain of gene, a linkage disequilibrium block, or a genetic region bounded by recombination hotspots, i.e., any collection of SNPs which are expected to be in linkage

disequilibrium with non-genotyped functional polymorphisms. Thus our method differs from ranked SNP-specific corrections [3-6].

The hypothesis tested by our method differs from the hypothesis tested by the method by Gauderman et al. [16], who use principal component variables to summarize the implicit haplotypes of the SNPs within the gene. Our approach provides a gene-wide test that can summarize signals from correlated and uncorrelated markers, that may reside in different haplotypes.

We recognize some caveats regarding the use of this method. This method should only be used to test hypotheses of association of a whole gene with a phenotype, where SNPs are not hypothesized to be functional, but only in linkage disequilibrium with unknown functional loci. If there is a biological hypothesis regarding a certain polymorphism, this method does not apply, and will be less powerful than a hypothesis test for the single hypothesized polymorphism. This is because sum of chi-squared statistics for the hypothesized polymorphism and other non-hypothesized polymorphisms, which may not be associated with the phenotype may result in a bias towards the null for the summed chi-squared statistic. The test statistic for this method is an overall p-value, and there is no interpretable beta coefficient. However, even for SNPs in linkage disequilibrium with the true functional polymorphism, there is no mechanistic interpretation of the beta coefficient. Indeed, when the hypothesis is regarding the whole gene, which may have multiple polymorphisms with different degrees of association with the phenotype, no specific beta coefficient is meaningful.

4. CONCLUSION

We have demonstrated and validated a method for the estimation of the overall association of gene with a phenotype by combining the association tests for genotyped polymorphisms within the gene. If the hypothesis is regarding the whole gene and not a specific polymorphism, this method is valid, and tends to be more powerful than the conservative Bonferroni correction for multiple polymorphism tests.

CONSENT

All authors declare that ‘written informed consent was obtained from the patient (or other approved parties) for publication of this case report and accompanying images.

ETHICAL APPROVAL

This study was approved by the Johns Hopkins Institutional Review Board, and all research participants gave informed consent. All authors hereby declare that all experiments have been examined and approved by the appropriate ethics committee and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

ACKNOWLEDGEMENTS

This research was supported by the NIH Grants, R01 HL72518, R01 HL092165, and the Johns Hopkins Institute for Clinical and Translational Research, NIH grant M01 RR00052.

The funding agency had no role in the study design, collection, analysis and interpretation of data or in the writing of the manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Sale MM, Mychaleckyj JC, Chen W-M. Planning and Executing a Genome Wide Association Study (GWAS). In: Park-Sarge O-K, Curry TE, editors. Molecular Endocrinology, Methods in Molecular Biology 590, Humana Press, Totowa, NJ. 2009;403-418.
2. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. In: Studi in Onore del Professore Salvatore Ortu Carboni. Rom, Italy. 1935;13-60.
3. Coneely 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.
4. 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.
5. 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.
6. Seaman SR, Muller-Myhsok B. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet 2005;76:399-408.
7. Fisher RA. 1932. Statistical methods for research workers. London: Oliver and Boyd.
8. Hoh J, Wille A, Ott J. Trimming, weighting and grouping SNPs in human case-control association studies. Genome Res. 2001;11:2115–2119.
9. Unschuld PG, Ising M, Erhardt A, et al. Polymorphisms in the serotonin receptor gene HTR2A are associated with quantitative traits in panic disorder. Am J Med Genet B Neuropsychiatr Genet. 2007;144B:424-9.
10. Wille A, Hoh J, Ott J. Sum Statistics for the Joint Detection of Multiple Disease Loci in Case-Control Association Studies With SNP Markers. Genet Epidemiol. 2003;25:350-359.
11. Kotz S, Balakrishnan N, Johnson NL. Continuous Multivariate Distributions. JohnWiley & Sons, New York; 2000.
12. Borecki IB, Province MA. Genetic and genomic discovery using family studies. Circulation. 2008;118:1057-1063.
13. Pirruccello J, Kathiresan S. Genetics of lipid disorders. Curr Opin Cardiol 2010;25:238–242.
14. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008;40:161-9.
15. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904-909.

16. Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol. 2007;31:383-95.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=358&id=12&aid=2693