e-Research: A Journal of Undergraduate Work

Volume 1
Number 2 *Vol 1, No 2 (2010)*

Article 3

September 2014

On the ranking of the disease susceptibility locus in family-based candidate gene studies: a simulation-based analysis

Lisa A. Brown

Cyril Rakovski

e-Research: A Journal of Undergraduate Work, Vol 1, No 2 (2010)

On the ranking of the disease susceptibility locus in family-based candidate gene studies: a simulation-based analysis

Lisa A. Brown, Cyril S. Rakovski

Abstract

The ranking of the p-value of the true causal single nucleotide polymorphism in the ordered list of individual SNP p-values is an important factor for achieving success in the ultimate objective of association studies - identifying deleterious genetic variants. Thus, we undertake a study to assess the implications of complex, multimarker correlation structure, sample size and disease models on the ranking of the causal SNP. We carry out an extensive family-based candidate gene simulation study to analyze the position of the disease susceptibility locus in the complete list of individual SNP p-values ordered according to their statistical significance. We simulate data based on the haplotype distributions of ten randomly selected genes extracted from the HapMap database, various sample sizes (600,1000 and 2000) that current association studies employ, and disease models that mimic the characteristics of complex human disorders.

We conclude that the average ranking of the causal SNP for sample sizes 600, 100 and 200 of 10.97, 9.65, and 8.34 are dramatically distant from the most significant and intuitively appropriate top position. This result is even more pronounced for genes with high average correlation and large number of common SNPs. Moreover, the gain of the DSL ranking when comparing sample sizes 600 to 1000 and 1000 to 2000, averaged over disease models, causal SNPs and genes, was approximately 1.3. These outcomes both reveal the importance of the sample size and quantify the magnitude required to unequivocally determine the identity of the DSL in family-based candidate gene studies.

Our results show the overwhelming importance of large sample sizes in the localization of deleterious SNPs even under simple disease models. These conclusions possess pronounced importance for the design and result interpretation of candidate gene, next generation high-density genome-wide association studies, as well as for the construction and implementation of association tests based on the distribution of the most significant (minimum p-value) test statistics.

Keywords: Fine Mapping, FBAT, Causation

Background

There has been a continual effort in finding the best analytical methods for identifying disease causal mutations in the human genome. The chronological progression of the research methodology involves linkage analysis [Chioza, et al. 2009; King, et al. 2000], candidate gene studies [Mollaki, et al. 2009; Murphy, et al. 2009; Rakovski, et al. 2007b] and currently, genome-wide association scans based on hundreds of thousands of single nucleotide polymorphisms (SNPs) with case-control [Barrett, et al. 2009; Beecham, et al. 2009; Check Hayden 2009; Himes, et al. 2009; Palmer, et al. 2009; van de Mortel, et al. 2000] and family-based data [Malarstig, et al. 2009; Van Steen and Lange 2005]. The ultimate objective of detecting and assessing the effects of the disease susceptibility loci (DSL) is predicated not only on the existence of significant corresponding test statistics p-values but also on their ranking in the sorted list of all (commonly SNP-specific) p-values. This is especially true with genome-wide

association scans due to the severity of the multiple testing problems which is generally handled by significance level adjustment methods that fail to reorder the test statistics. Thus, the emergence of the true DSL at the top of the significance list is pivotal for the success and replicability of genome-wide association scans.

In current large-scale studies, the physical locations of the genotyped SNPs are usually far apart and linkage disequilibrium (LD) is not an issue. On the other hand, in family-based and case-control candidate gene association studies we come across high level of LD between the pairs of SNPs in candidate regions containing a putative deleterious allele which causes multiple significant p-values to arise from association testing. In fact, we generalize a related question about the induced test statistics correlation for pairs of SNPs in LD [Pritchard and Przeworski 2001]. The problem of the position of the DSL in the significance ranking in candidate gene studies arises both from multiple testing and complex multimarker correlation structure represented by the region's haplotype distribution. Finally, the analysis of next generation genome-sequencing data [Check Hayden 2009; Pennisi 2009] will include a combination of both challenges, a dramatic increase of the magnitude of the multiple testing problem and an additional complexity of high correlations between closely positioned SNPs. In general, it is intuitive to expect that the most significant p-value will originate from the true causal SNP, but as with genome-wide association studies, this may not always be true. Lastly, there is a class of powerful testing strategies that are based on the distribution of the most significant (minimum p-value) individual marker test statistic that make the implicit assumption that the statistic with the minimum p-value corresponds to the causal SNP [Kimmel, et al. 2007; Rakovski, et al. 2007a]. We have undertaken a family-based simulation study to explore the validity of this assumption.

Methods

We simulated and analyzed family data consisting of family trios with complete data and discordant sib pairs with missing parental data. The results for both family designs were practically identical with respect to the question of interest; thus, in the subsequent presentation we report results for discordant sib-pairs data analyses only. With respect to data analyses approaches, we used the two natural tools for detecting difference in allele frequencies at a particular locus between related cases and controls, conditional logistic regression and the classical family-based test for association [van de Mortel, et al. 2000]. Since the results obtained under the implementation of both methods were extremely similar, we only report the outcomes of our work for conditional logistic regression. As customary, we employ additive coding of the marker genotypes by assigning values of 0, 1, and 2 to AA, Aa, and aa, respectively. In the conditional logistic regression model, we treated each family as a stratum containing the affected-unaffected pair of offspring and fitted all single covariate models to assess the unadjusted effect sizes of all markers in the candidate region.

Simulation

We randomly selected 10 genes from autosomal chromosomes of the human genome, extracted the resequenced unphased genotypes of the CEPH families from the publicly accessible ENCODE data from the HapMap database [Tanaka 2009] and calculated the corresponding haplotype distributions using the EM algorithm [Dempster, et al. 1977] via its implementation in the FBAT software package [Girirajan, et al. 2009]. We used R version 2.9.2 [Thauvin-Robinet, et al. 2009] for both the simulation of the genotype data and for the analysis of these data via FBAT and conditional logistic regression. Summary statistics of the ten genes are shown in Table 1.

Table 1. Summary characteristics of the 10 genes.

Gene	Common SNPs*	Mean MAF	Mean $\mathbf{r}^2$	Size**	Common Haplotypes***
1	18	0.14	0.38	21	4
2	20	0.18	0.24	13	6
3	30	0.30	0.42	11	5
4	41	0.22	0.24	22	6

5	14	0.15	0.25	10	4
6	61	0.24	0.39	19	6
7	96	0.23	0.32	60	4
8	29	0.30	0.37	34	4
9	46	0.25	0.89	38	5
10	36	0.18	0.37	17	5

*Number of SNPs with minor allele frequency (MAF)> 0.05.
**Measured in kilobase pairs (kb).
***Number of haplotypes with frequency >0.05.

We randomly paired haplotypes to create each of the parents and simulated pairs of offspring through Mendilian transmissions of haplotypes. Further, in consecutive simulation steps, we used six different disease models that mimic the small effect sizes of complex human disorders [Rakovski, et al. 2007b] to assign disease status to each offspring. We implemented two additive, two dominant and two recessive models defined through triplets of penetrance function with details shown in Table 2. For each of the ten genes and under each disease model, we simulated populations of families with pairs of offspring and created datasets of sizes 600, 1000 and 2000 by ascertainment of families with discordant offspring disease status. In each of the 1000 simulated dataset the causal SNP was randomly chosen and the parental genotypes were removed from the data and subsequent analyses. Summary statistics of the ten genes are shown in Table 2.

Table 2. Disease models used in the simulation study.

	Penetrance functions		
	f_0	f_1	f_2
Model 1	0.005	0.01	0.015
Model 2	0.01	0.02	0.03
Model 3	0.005	0.01	0.01
Model 4	0.01	0.02	0.02
Model 5	0.005	0.005	0.01
Model 6	0.01	0.01	0.02

Results

Our results show the importance of the three analyzed factors: gene, disease model and sample size on the ranking of the DSL. Moreover, we present the outcome of our simulation study that describes the three-way interaction of these factors. The details on the ranking of the DSL p-value for each sample size, gene and disease model combinations are displayed in Tables 3, 4 and 5.

Table 3. Ranking* of the DSL averaged over randomly chosen causal SNPs for sample size 600.

	Gene									
Disease Model**	1	2	3	4	5	6	7	8	9	10
1	4.35	4.41	5.29	5.80	4.05	11.25	16.09	4.73	20.34	7.48
2	3.97	4.06	5.71	7.38	3.50	13.73	18.03	4.94	20.10	8.23
3	2.88	3.12	4.71	4.70	2.82	10.95	13.22	3.56	19.24	6.49
4	3.07	3.48	4.72	4.95	2.72	11.02	13.71	3.44	17.87	6.51
5	8.77	8.55	13.69	17.13	6.88	25.00	40.23	11.45	22.15	16.08
6	9.20	8.70	13.88	17.96	7.02	26.79	38.89	12.23	23.11	14.14

*Ties among ranked p-values were resolved by assigning averages.
** These six disease models are described in the Simulation Design section.

Table 4. Ranking* of the DSL averaged over the randomly chosen causal SNPs for sample size 1000.

	Gene									
Disease Model**	1	2	3	4	5	6	7	8	9	10
1	2.53	3.64	4.39	5.05	3.26	7.86	14.59	3.12	18.37	6.92
2	3.26	3.34	4.08	4.56	2.92	8.99	14.79	2.63	18.07	7.07
3	1.93	2.83	4.04	3.66	2.09	8	8.72	3.23	17.39	5.35
4	2.09	2.54	3.91	3.40	2.05	6.81	10.24	3.09	17.44	5.16
5	9.04	8.65	10.34	15.26	7.04	22.97	40.44	8.98	22.26	16.19
6	8.45	7.6	10.74	17.18	6.43	25.28	43.37	10.04	22.14	13.26

*Ties among ranked p-values were resolved by assigning averages.

** These six disease models are described in the Simulation Design section.

Table 5. Ranking* of the DSL averaged over the randomly chosen causal SNPs for sample size 2000.

	Gene									
Disease Model**	1	2	3	4	5	6	7	8	9	10
1	2.18	3.21	3.66	3.63	1.77	7.53	9.02	2.47	18.54	5.78
2	1.73	2.62	3.79	3.61	1.96	6.89	9.64	3.03	17.36	5.51
3	1.64	2.49	3.70	3.48	1.67	5.90	7.45	2.60	17.15	4.79
4	1.56	2.00	3.48	3.44	1.74	7.25	7.30	2.92	16.99	5.09
5	7.96	7.26	9.54	16.41	6.06	20.73	29.68	8.98	21.50	12.73
6	8.93	8.64	9.17	13.77	5.98	18.30	34.54	7.31	22.18	14.29

*Ties among ranked p-values were resolved by assigning averages.

** These six disease models are described in the Simulation Design section.

Table 3 shows the average ranking of the DSL p-value for sample size of 600. The three consistently least favorable results are associated with genes 6, 7 and 9 with rankings averaged over all disease models and causal SNPs (both unknown quantities) of 16.45, 23.36 and 20.47 respectively. These results are foreseeable as the above-mentioned genes possess the three largest numbers of common SNPs, 61, 96 and 46 and three of the top four biggest average correlations. In particular, the average correlation of the ten genes is 0.33 while the average correlation within gene 9 is 0.89. In contrast, the lowest rank of 2.72 was obtained for the combination of disease model 4 and gene 5; this is the smallest candidate region with very low average correlation. On the other hand, disease models 3 and 4 attained the highest rankings, averaged over all DSLs and genes, of 7.17 and 7.15. Models 1 and 2 performance followed closely with ranks of 8.38 and 8.96 while models 5 and 6 could only achieve average scores of 16.99 and 17.19. However, these differences are expected due to the differences in the underlying distinctions in the effect sizes. Further, looking within disease models of the same type, the average increase taken over models 1, and 2, 3 and 4, and 5 and 6 is only 0.3. Lastly, the ranking of the DSL for sample size 600 averaged over all simulation setting parameters is 10.97.

Table 4 shows the average ranking of the DSL p-value for sample size of 1000. Again, genes 6, 7 and 9 attain the largest average rankings of 13.32, 22.02 and 19.28 respectively. The lowest rank of 1.93 was obtained for the combination of disease model 3 and gene 1; this is the second smallest candidate region which also possesses low average correlation. Moreover, as anticipated the order for all disease models is constant across sample sizes, models 3 and 4 attained rankings of 5.72 and 5.67, models 1 and 2 followed with 8.38 and 8.96 while models 5 and 6 average achieve 16.99 and 17.19. Further, looking within disease models of the same type, the average increase

taken over models 1, and 2, 3 and 4, and 5 and 6 is only 0.1. Lastly, the ranking of the DSL for sample size 1000 averaged over all simulation setting parameters is 9.65.

Table 5 shows the average ranking of the DSL p-value for sample size of 2000. Similarly to the previous sample sizes, genes 6, 7 and 9 attain the largest average rankings of 11.10, 16.30 and 18.95 respectively. The lowest rank of 1.55 was obtained for the combination of disease model 4 and gene 1. Models 3 and 4 attained rankings of 5.10 and 5.17, models 1 and 2 followed with 5.78 and 5.61 while models 5 and 6 average achieve 14.08 and 14.31. Further, looking within disease models of the same type, the average increase taken over models 1, and 2, 3 and 4, and 5 and 6 is only 0.05. Lastly, the ranking of the DSL for sample size 1000 averaged over all simulation setting parameters is 8.34.

Interestingly, the ascendancy of the DSL across the ranking list as sample size increases is much slower than anticipated. We observe that the average ranking improves by only 1.3 as we increase the sample sizes from 600 to 1000 and from 1000 to 2000. On the other hand, expectedly, based on their inherent characteristics, genes 6, 7, and 9 attain the lowest ranking regardless of sample size. Even sample size of 2000 is not sufficient to overcome the detrimental effect of large number of SNPs and high average correlation present within these genes. Further, for all simulated settings there are no observed DSL rankings of less than 2 for sample size 600 and there is only one such ranking (1.925) for sample size 1000. In contrast, for sample size 2000, we see multiple (7) average rankings of less than 2. Clearly, however, these conclusions are not particularly inspiring or helpful for the localization and identification of the causal SNP.

Discussion

We implemented a simulation-based study to explore the behavior of the ranking of the p-value of the DSL in family-based candidate gene association studies. We used classical FBAT as well as the Wald test arising from conditional univariate (SNP-specific) logistic regression models to carry out the statistical analyses. We did not use tagging SNPs because the deleterious variant might be absent from the set and that diverts the focus of the study in a different direction.

There are several important conclusions with far-reaching consequences produced by our study. For all of the studied scenarios, the overall ranking of the DSL is dramatically distant from the intuitively anticipated top position. The departure is especially remarkable with smaller sample sizes and for genes with high average correlation and large number of common SNPs. This is a somewhat surprising result that contradicts the intuitive expectation that the DSL will likely yield the most significant p-value regardless of the correlation structure, gene span and sample size. The gain in the DSL ranking when comparing sample sizes 600 to 1000 and 1000 to 2000, averaged over disease models and genes, is approximately 1.3 .This reveals the magnitude of the sample sizes needed to precisely determine the location and identity of the DSL in family-based candidate gene studies. This problem could be further exacerbated if we are testing multiple genes matching a linkage peak, performing a second stage genome-wide scan or next generation genome sequence analysis.

Lastly, minimum p-value-based association methods make the implicit assumption that the most significant test statistic is driven by and associated with the true causal SNP. We have shown that it is rarely the case and in the vast majority of the cases the most significant test corresponds to a SNP that is in LD with the causal SNP. This means that common practice to report the most significant results would potentially omit the DSL and make the findings unlikely to replicate in subsequent studies, especially in populations with different LD structure.

Conclusions

Our results underline the necessity of large sample sizes in the localization of deleterious SNPs even under simple disease models. These conclusions possess pronounced importance for the design and result interpretation of

candidate gene, next generation high-density genome-wide association studies, as well as for the construction and implementation of association tests based on the distribution of the most significant (minimum p-value) test statistics.

Acknowledgments

We thank Drs. Daniele Struppa, Michael Fahy and Janeen Hill for their support of this work.

References

Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H and others. 2009. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet.

Beecham GW, Martin ER, Li YJ, Slifer MA, Gilbert JR, Haines JL, Pericak-Vance MA. 2009. Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. Am J Hum Genet 84(1):35-43.

Check Hayden E. 2009. Genome sequencing: the third generation. Nature 457(7231):768-9.

Chioza BA, Aicardi J, Aschauer H, Brouwer O, Callenbach P, Covanis A, Dooley JM, Dulac O, Durner M, Eeg-Olofsson O and others. 2009. Genome wide high density SNP-based linkage analysis of childhood absence epilepsy identifies a susceptibility locus on chromosome 3p23-p14. Epilepsy Res.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. 39(1):1-38.

Girirajan S, Chen L, Graves T, Marques-Bonet T, Ventura M, Fronick C, Fulton L, Rocchi M, Fulton RS, Wilson RK and others. 2009. Sequencing human-gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. Genome Res 19(2):178-90.

Himes BE, Hunninghake GM, Baurley JW, Rafaels NM, Sleiman P, Strachan DP, Wilk JB, Willis-Owen SA, Klanderman B, Lasky-Su J and others. 2009. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. Am J Hum Genet 84(5):581-93.

Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. 2007. A randomization test for controlling population stratification in whole-genome association studies. Am J Hum Genet 81(5):895-905.

King AL, Yiannakou JY, Brett PM, Curtis D, Morris MA, Dearlove AM, Rhodes M, Rosen-Bronson S, Mathew C, Ellis HJ and others. 2000. A genome-wide family-based linkage study of coeliac disease. Ann Hum Genet 64(Pt 6):479-90.

Malarstig A, Buil A, Souto JC, Clarke R, Blanco-Vaca F, Fontcuberta J, Peden J, Andersen M, Silveira A, Barlera S and others. 2009. Identification of ZNF366 and PTPRD as novel determinants of plasma homocysteine in a family-based genome-wide association study. Blood 114(7):1417-22.

Mollaki V, Georgiadis T, Tassidou A, Ioannou M, Daniil Z, Koutsokera A, Papathanassiou AA, Zintzaras E, Vassilopoulos G. 2009. Polymorphisms and haplotypes in TLR9 and MYD88 are associated with the development of Hodgkin's lymphoma: a candidate-gene association study. J Hum Genet.

Murphy A, Tantisira KG, Soto-Quiros ME, Avila L, Klanderman BJ, Lake S, Weiss ST, Celedon JC. 2009. PRKCA: a positional candidate gene for body mass index and asthma. Am J Hum Genet 85(1):87-96.

Palmer ND, Langefeld CD, Ziegler JT, Hsu F, Haffner SM, Fingerlin T, Norris JM, Chen YI, Rich SS, Haritunians T and others. 2009. Candidate loci for insulin sensitivity and disposition index from a genome-wide association analysis of Hispanic participants in the Insulin Resistance Atherosclerosis (IRAS) Family Study. Diabetologia.

Pennisi E. 2009. DNA sequencing. No genome left behind. Science 326(5954):794-5.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1):1-14.

Rakovski C, Xu X, Laird N. 2007a. A new permutation test for family-based association studies. p 16.

Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM. 2007b. A new multimarker test for family-based association studies. Genet Epidemiol 31(1):9-17.

Tanaka T. 2009. [HapMap project]. Nippon Rinsho 67(6):1068-71.

Thauvin-Robinet C, Franco B, Saugier-Veber P, Aral B, Gigot N, Donzel A, Van Maldergem L, Bieth E, Layet V, Mathieu M and others. 2009. Genomic deletions of OFD1 account for 23% of oral-facial-digital type 1 syndrome after negative DNA sequencing. Hum Mutat 30(2):E320-9.

van de Mortel TF, Laird P, Jarrett C. 2000. Client perceptions of the polysomnography experience and compliance with therapy. Contemp Nurse 9(2):161-8.

Van Steen K, Lange C. 2005. PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. Hum Genomics 2(1):67-9.