Author's response to reviews

Title: Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families

Authors:

Suyeon Park (sooyeon1002@gmail.com)
Sungyoung Lee (biznok@gmail.com)
Young Lee (lyou7688@gmail.com)
Christoph Lange (CLANGE@hsph.harvard.edu)
Sungho Won (won1@snu.ac.kr)

Version: 3 Date: 1 April 2015

Author's response to reviews: see over
Dear Editor,

We very much appreciate the insightful comments of the reviewers. They helped us to improve the manuscript significantly. In the revised version, we have addressed all comments.

With very best wishes,

Reviewers' Comments

Reviewer: 1

Specific concerns:

One limitation of this manuscript, if I understand correctly, is that the methods built on the assumptions are made on ascertainment method as well as the quantity of heritability. This is typically not known by researchers

(Response) We appreciate this thoughtful comment. Heritability and ascertainment conditions are often unknown, and in such a case the proposed method is not useful. We added the following sentences:

the mis-specified heritability and prevalence can lead to the statistical power loss for the proposed methods, but it was found to be not substantial at least in our simulation studies.

However we considered the limited ascertainment conditions and comprehensive simulation studies are still necessary.

The other question is: is this approach more useful if the pedigree structure is more complex? Further, if one is looking at multiple generations in Latino families, how well does the current method work? The authors may wish to explore this case of scenario.

(Response) We are sorry for the uncertainty of this issue. The proposed method can analyze the extended families and our software can handle the extended families. In addition the heterogeneous effects of ascertaining condition on family members become larger in the extended families and the proposed method may provide more power improvement for the extended families. For instance, Tables 2-4 are for nuclear families and Tables 5-7 are for large families with three generations. The empirical power differences between WL and the proposed methods are bigger for large families. We added the following sentences:

the proposed method was conceptually simple and can be applied to the extended families

Furthermore the differences of the empirical power estimates from WL and the proposed methods are larger for Tables 5-7 than Tables 2-4, which indicates that the heterogeneity of ascertainment condition is positively related with family size and the proposed methods become more efficient for extended families

Lastly, with the real-world data presented in the manuscript, what will happen if one simply uses FBAT to conduct the analysis without any level of statistical correction.

(Response) We appreciate this thoughtful suggestion. We included results from FBAT in our real data analyses.
Reviewer: 2

Major Compulsory revisions

One of the main contributions of this work relies on the derivation of the efficiency increasing of the proposed test statistic. This is shown in the supplementary material (Text S1). However, in its represent form, the derivation is not correct: Several minor mistakes occur that makes the derivation difficult to understand. For instance, line 17, a symbol is written instead of vertical dots. Line 19, -2 has to be +2. Line 21, last equality is wrong it should be written as follow:

(Response) We appreciate these comments. We modified our manuscript accordingly.

More importantly, in this text S1, the mean issue is to prove the inequality on page 2 which is not rigorously done in this version. Indeed, the authors have to prove that the last term of the right-hand side of equation line 19 is positive definite. This is not proved in the next equation that allows only to conclude that this term is null. Moreover, the equation line 21 are not justified: why the authors use this approximation which is generally not true?

(Response) We are sorry for this limitation and appreciate this thoughtful comment. The variance of depends on the realization of and there is no globally optimal choice of the offset. Thus alternatively we try to minimize the expected variance of . We added the following sentences

The optimal choices of offset can be identified by minimizing the variance and if we let \( V = I_N \) and

\[
\Sigma = \sigma^2 \left( R - 1_N \left( 1_N' R^{-1} 1_N \right)^{-1} 1_N \right),
\]

the denominator of the non-centrality parameter became

\[
\text{var}(S) = \text{tr} \left\{ \Sigma T T' \right\}
\]

However the amount of variance depends on the realization of , and there is no globally optimal choice of offset. Alternatively we focus on the expected variance of .

Furthermore, ascertainment process was simulated for a half of the sample and another ascertainment process was simulated for the other half, describing an heterogeneous ascertainment process. It would be interesting to simulate other ascertainment processes in order to assess the robustness of the proposed method. I am not sure to understand the interest of the way of simulating the heterogeneity: why do not simulate part of affected individuals.

(Response) We are sorry for the uncertainty of our manuscript. In our simulations, we considered two different scenarios. In scenario 1, we assumed that half of families have at least \( n_{\text{proband}} \) affected offspring and the other half of families have at least \( n_{\text{proband}} \) unaffected offspring. In scenario 2, we assumed that all families have at least \( n_{\text{proband}} \) affected offspring. \( n_{\text{proband}} \) can be 1, 2 or 3. For the former, we considered the effect of heterogeneous ascertaining condition between families while the latter assume that ascertaining conditions are homogeneous. However we agreed that our consideration is still limited and we include the following sentences:
In the second simulation setting, the numbers of extended families were assumed to be 100, 300, 600, and 900, and all families were ascertained if the number of affected family members was larger than or equal to $n_{\text{proband}}$. However, we considered the limited ascertainment conditions and comprehensive simulation studies are still necessary.

Additionally, the statistical model is not detailed in the paper. This would allow to discuss the differences between this model and the one of Won and Lange. Consequently, $\beta$ parameter equal to zero under the null hypothesis, is introduced line 8 without prior definition.

(Response) We are sorry for the uncertainty and appreciate this comment. We modified our manuscript accordingly and added the following sentences to provide a model for our statistics:

We denoted an minor allele frequency (MAF) of a variant in unaffected individuals by $q_A$. We assumed [19] that for a constant $\gamma$,

$$E(X|T) = 2q_A I_n + \gamma T,$$

where $0 < 2q_A + \gamma < 1$. Then, the score for a variant [19, 24] can be defined by

$$S = T'(X - \hat{E}(X)) \text{ and } \hat{E}(X) = I_n \left( I_n' R^{-1} I_n \right)^{-1} I_n' R^{-1} X.$$

The variance of $S$ is

$$\text{var}(S) = \sigma^2 T' V^{-1} \left( R - I_n \left( I_n' R^{-1} I_n \right)^{-1} I_n' \right) V^{-1} T,$$

and we considered the following statistic[18-19, 25]:

$$\frac{T' \left( I_n - \left( I_n' R^{-1} I_n \right)^{-1} I_n' R^{-1} \right) X}{\sqrt{\sigma^2 T' \left( R - I_n \left( I_n' R^{-1} I_n \right)^{-1} I_n' \right) T}} \sim N(0,1) \text{ if } \gamma = 0$$

Moreover the authors have written that the power loss attributable to the misspecified prevalence is not substantial. However, table 8 suggests a significant loss of power. Could the authors explain their conclusion? And what happen if $n_{\text{miss}}$ is bigger than 3?

(Response) We are sorry for this and appreciate this comment. In Table 8, the effect of mis-specified prevalence can be found by comparing the results obtained by setting 0.2 prevalence with those from other choices of prevalences. Results show that their differences are usually small. However the empirical powers for $FQLS_1$ and $FQLS_2$ are quite different, and we added the following sentences:

Table 8 shows that the results obtained by setting prevalence to be 0.1 and 0.3 are similar to the results when the prevalence was set to be 0.2, which indicates that the power loss attributable to the misspecified prevalence is not substantial. Furthermore the empirical power estimates are positively related with $n_{\text{proband}}$ and inversely related with $n_{\text{missing}}$. If $n_{\text{missing}}$ is larger than 3, the power loss may be more substantial.
Finally I am surprised that FQLS2 give better results even when the probands are well designed. Could the authors discuss this results.

(Response) We are sorry for the uncertainty of our manuscript. For the most of family-based association analyses, probands are not clearly defined and we considered this issue in our simulation studies. We assume that families with the number of affected family members larger than \( n_{proband} \) are ascertained, and the uncertainty of probands makes FQLS\(_1\) less efficient. We added the following sentences:

Last our simulation results show that FQLS\(_2\) was slightly better than FQLS\(_1\), and this may be induced by the uncertainty of probands in our simulation studies.

Minor Essential revisions

1. Typo for the probability density function line10-page9
2. Missing “s” kube15-page10
3. Line18-page14 : It is Figure 4 in place of Figure2
4. Line23-page14 : Figure 6 in place of Figure 4

(Response) We appreciate these comments. We changed our manuscript accordingly.