Reviewer’s report

Title: Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk

Version: 2 Date: 9 August 2011

Reviewer: Nandita Mitra

Reviewer’s report:

Major Compulsory Revisions:
This well-written manuscript by Simonson et al. presents results from three separate analytic approaches to analyzing GWAS data from the Framingham SNP Health Association Resource data set. The statistical approaches are very clearly and thoroughly described and provide an excellent tutorial for polygenic prediction and polygenic heritability analysis of GWAS data for complex traits. The manuscript also includes comprehensive descriptions of QC processes for GWAS data in addition to a nice description of 10-fold cross validation. A few points need to be addressed regarding the results of the CVD analysis in addition to a few other concerns as listed below.

1. Two methods that utilize combinations of unrelated (and even un-associated) SNPs to explain the overall variability in a CVD related phenotype using GWAS data are presented. This is an interesting approach but one wonders whether this approach of combining seemingly unrelated SNPs is more meaningful than a pathway based approach. Some discussion of this issue should be provided.

2. Along those same lines, I am a bit confused by the CVD polygenic prediction analysis in which, in the discovery phase, subsets of SNPs are defined based on p-values (0-.1, .1-.2, .2-.3, .3-.4 etc). Why is this a meaningful way to group SNPs rather than biological effects or pathway membership? Also, very surprisingly, the number of SNPs appears to be very evenly distributed between these p-value categories. Also, it would help to describe in more detail why the findings in Table 4 are of interest. I am having a hard time understanding why SNPs that had p-values between .2 and .3 in single SNP association tests would then be significantly associated (p=0.004) in combination. What does this mean, why is this significant, and how can this information be used? On the other hand, SNPs that were nominally significant on single SNP analysis are not significant in combination. It seems that other groupings would also be significant such as taking some SNPs from each group. The seemingly endless ways to group SNPs should be addressed.

3. In the GWAS analysis, it is stated that data were “primarily from the Affymetrix 500k mapping array”. Since other platforms may have been used, was imputation needed to combine different platforms. Further, since 500k is relatively small in this era of 2million snp chips, was imputation considered with say the hapmap or 1000 genomes as a backbone to increase the number of snps that were
4. In the description of polygenic prediction, it is stated that PLINK’s SNP scoring routine is used. What specific score or weight was used? It has been shown via simulations that with large numbers of SNPs, when using scoring methods, there is a large chance of both false positives and false negatives. This is a serious concern here where approximately 25,000 SNPs are combined using the scoring method.

Minor Compulsory Revisions:
1. Figure 1 seems to replicate all of the information in Table 4 and is unnecessary.
2. On the bottom of page 8, “confounds” should read “confounders”.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.