***Authors have addressed some of my previous comments sufficiently but not the following ones:***

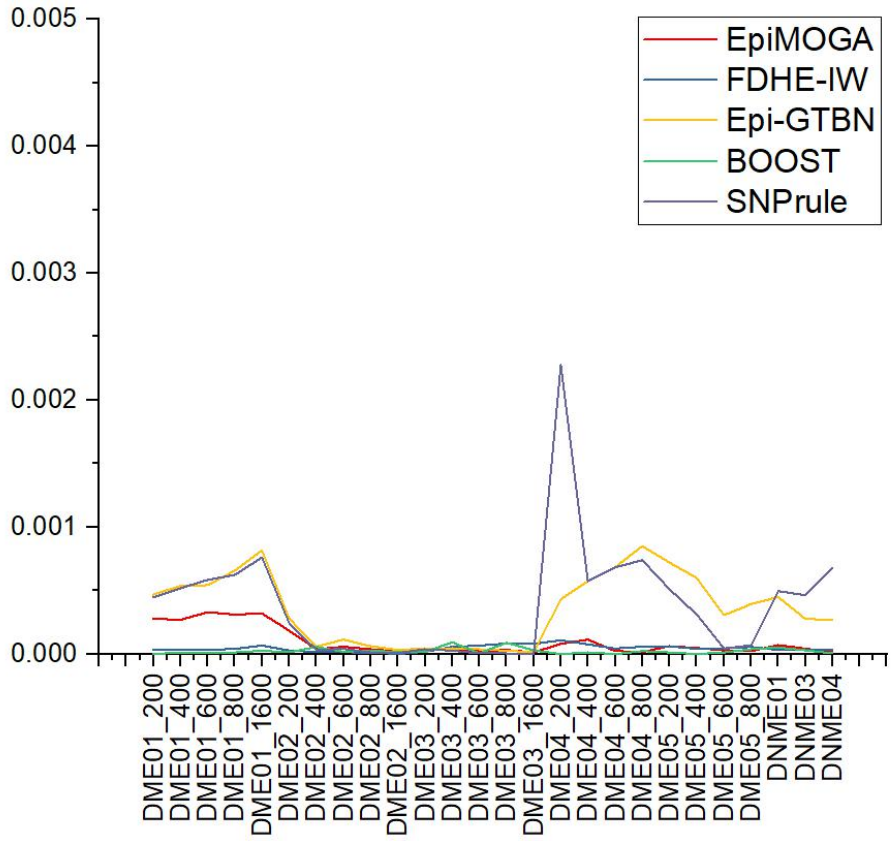
***2. On the false positives. just do not see why "it is unreasonable to choose false positives as the evaluation index of the method". The argument that "the number of negatives is much larger than the number of positives" is not valid: Even in a typical GWAS for a complex trait, we expect this asymmetry. Is this one of the reasons that we correct for multiple hypothesis testing?***

**Response:**

We are sorry that we did not explain clearly about the false positive rate. To solve this problem, we calculated the false positive rate for different methods on all 26 simulated datasets and the following paragraphs about the false positive rate have been added to the Supplementary Material (Section 6).

**"The false positive rate of simulated experimental results**

For comparison methods further, we calculated the false positive rate (FPR) of different methods on all 26 simulated datasets. Supplementary Figure 4 shows the false positive of different mthods on different datasets.




**Supplementary Figure 4.** Line chart of false positive rates of different methods on different datasets.

As shown in the Supplementary Figure 4, we found that the EpiMOGA method showed a very low false positive rate in most datasets. At the same time, it can be seen that the maximum false positive rate is less than 0.0025. Therefore, on the whole, the difference of false positive rate among different methods is not large.

In addition, false positivity is related to the number of solutions. Reducing the number of solutions can reduce the false positive rate, but also affect the detection efficiency Power. For example, in the datasets DME01_400, DME01_600 and so on, FDHE-IW and BOOST showed lower false positive rate, but at the same time, the detection efficiency Power of this method was far lower than that of other methods, as shown in the supplementary table 3.

**Supplementary table 3.** The Power and FPR of different methods

| | FDHE-IW | | BOOST | | EpiMOGA | | SNPrule | | Epi_GTBN | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Power | FPR | Power | FPR | Power | FPR | Power | FPR | Power | FPR |
| DME01_400 | 0.01 | 3.43E-05 | 0.04 | 1.212E-05 | 0.55 | 2.71E-04 | 0.11 | 5.19E-04 | 0.35 | 5.25E-05 |
| DME01_600 | 0.01 | 3.23E-05 | 0.03 | 1.010E-05 | 0.58 | 3.29E-04 | 0.17 | 5.84E-04 | 0.39 | 5.43E-04 |
| DME01_800 | 0.04 | 4.04E-05 | 0.2 | 1.212E-05 | 0.62 | 3.11E-04 | 0.19 | 6.24E-04 | 0.38 | 6.59E-04 |

Therefore, we chose a more comprehensive evaluation index F-measure to evaluate the detection accuracy in this study."

3. ***On the QC component. "The main step leading to the decrease of the SNP number is not the Hardy-Weinberg test, but the Chi-square test of SNP loci": what is this Chi-square test? Test of what? In additon on "The reason why we chose this approach is that we cannot directly process the original hundreds of thousands of data using MATLAB, due to the limitation of hardware equipment. Similar operations can be found in other methods such as Epi_GTBN and FHSA-SED": efficient computing is something worth discussing.***

***Regarding the HWE, if you choose a HWE p-value of 0.0001, you will screen out ~100 SNPs for no reasons other than what is expected based on multiple hypothesis tesing.***

**Response:**

As we introduced in the manuscript, the AD dataset has more than 600,000 SNP loci, which means that the combination of 2-SNPs may exceed 18 billion. The data dimension is too high for us to handle directly with our ordinary computers. Therefore, after reviewing relevant literature, we choose to complete the pre-screening of SNP through Chi-square test. The null hypothesis of the Chi-square test is that the single SNP locus is independent of Alzheimer's disease.

We are very sorry about the HWE p-value. After carefully checking of the source code, we found that this may be a clerical error when we wrote the paper. The p-value in the source code is set to 0.05. And we also checked the code and manuscript again and again, and we were sure that the p-value in HWE was 0.05. Thank you for helping us find this problem. We have modified "p-value from the Hardy-Weinberg test less than 0.0001" to "p-value from the Hardy-Weinberg test less than 0.05." in the revised manuscript.

4. ***On the gap between simulation and application settings. "The purpose of the simulation***

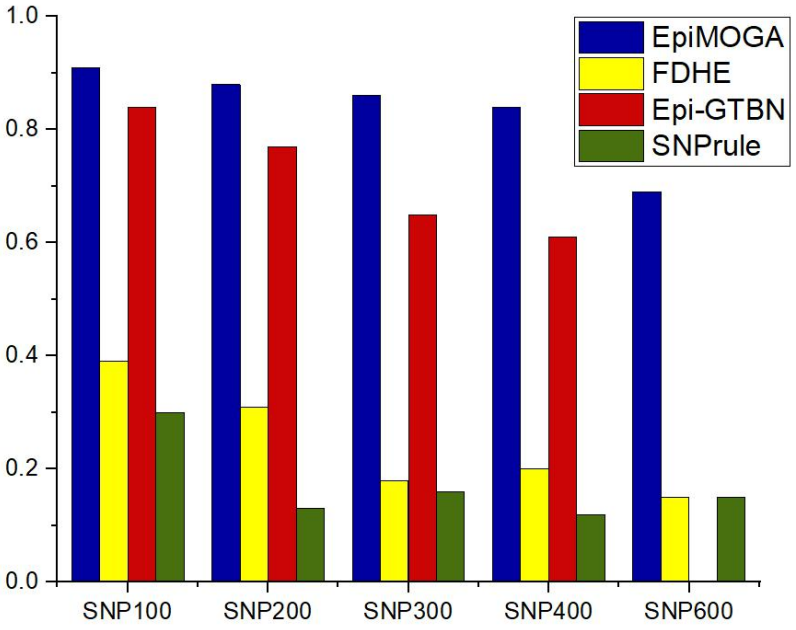
***experiment is not to completely simulate the actual situation but to compare the detection effects of different methods under the same conditions": Yes, it is often hard to make the simulation setting represent the real data setting because a) we don't know the truth of the real data to start with and b) computation can be an issue in simulation studies since many replicates must be studied. However, I find author's answer disappoting. Yes, the purpose of the simulation studies is to evaluate different methods under the same condition, but "the same condition" should be as comprehensive and as realistic as possible, so that by the time you apply the methods to a real dataset, you could say something about the performance of the difference methods.***

**Response:**

We are sorry that our answer did not satisfy you. We misunderstood your meaning. Now, we have added simulation experiments about the number of SNP loci to make the "condition" be more comprehensive. The following paragraphs about the influence of SNP quantity on detection efficiency have been added to the Supplementary Material (Section 7).

**"Analysis on the influence of SNP quantity on detection efficiency**

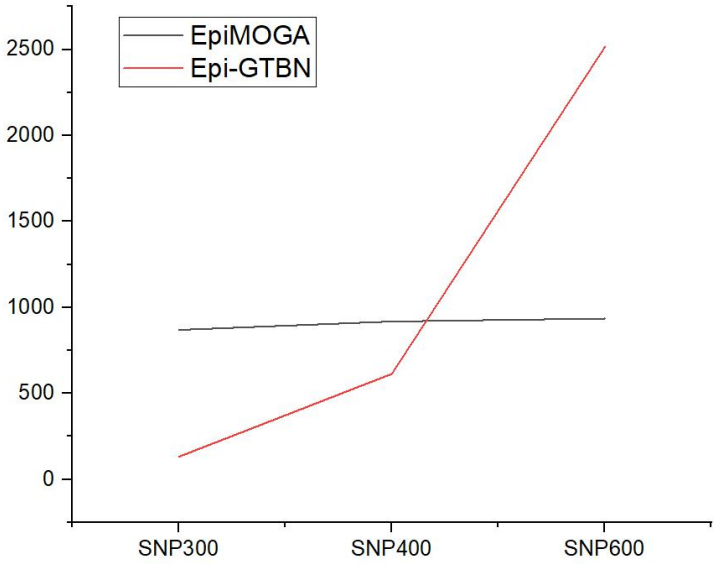
The number of SNP sites is also one of the important factors affecting the performance of methods. We designed a simulation experiment to analyze the influence of the number of SNP sites on different methods. Without changing other parameters, we used the DME model 2 to generate 5 datasets with SNP number of 100, 200, 300, 400 and 600 when sample size is 600. Supplementary figure 5 shows the detection efficiency of the different methods on this five datasets.




**Supplementary Figure 5.** Detection efficiency comparisons between EpiMOGA and other comparative methods on DME models with 5 different SNP number.

In supplementary Figure 5, as the number of SNP loci increase, the difficulty of detection increases significantly, and the detection efficiency of all methods has reduced. However, the detection efficiency of EpiMOGA on almost all datasets is better than FDHE-IW and SNPrule.

Although Epi-GTBN has maintained a relatively good detection efficiency, the detection time has increased rapidly, which will take a long time to complete the detection of the SNP600 and SNP800 datasets. We compared the detection time of EpiMOGA and Epi-GTBN on different datasets, as shown in supplementary Figure 6.




**Supplementary Figure 6.** The detect time of EpiMOGA and Epi_GTBN on different datasets.

In supplementary Figure 6, the abscissa represents the datasets with different number of SNP loci, and the ordinate is the detection time of a single dataset in seconds. As shown on the figure, the detection time of EpiGTBN is positively correlated with the number of SNPs. The detection time of Epi_GTBN exceeds 40 minutes on a SNP600 single dataset, and it will takes at least 3 days to complete the detection of all 100 datasets. At the same time, we also find that the detection time of Epi_GTBN on a single dataset of 800SNP is more than 90 minutes, what means that Epi_GTBN will take at least 7 days to complete all tests. In comparison, the detection time of EpiMOGA has increased slowly, maintaining a relatively stable state. On the SNP600 dataset, the detection efficiency of EpiMOGA almost reaches 70%, which is even better than the detection result of Epi_GTBN on SNP400 dataset. Meanwhile, the detection time of a single SNP600 dataset is only one third of that of Epi_GTBN.

Therefore, it is reasonable to believe that EpiMOGA is more suitable for the detection on a large number of SNP loci datasets than other methods.”