Many thanks for the helpful comments. We have tried our best to address the issues. Please find below a point to point answer.

**Reviewer number 1**

I feel that the paper in general has benefited from the revision by the authors. In my opinion, the current line of approach and structure of the paper makes the message of the authors more clear. At this stage, I still have some concerns (mostly in terms of writing); I have provided some detailed suggestions, questions and comments below.

General comments:

(1) The authors state that Satagopan et al. propose to base hypothesis tests on the observations of the second stage only. However, Satagopan et al. combine data from both stages. The difference with the work in this paper is that Satagopan et al. do not aim to control a particular error rate. Please revise this and adjust the text accordingly.

*Thank you for pointing this out. We changed the text in the introduction accordingly.*

(2) In the introduction, the authors write that using a fixed p-value threshold to select markers for the second stage will result with high probability in the selection of some markers even if the global null is true. They state that this is a drawback because resources could be saved by stopping the experiment for futility. While this reasoning seems sensible, I still don’t clearly see a difference with the situation of the FDRS design. Since this also involves a p-value cutoff, it will also be possible that some markers will be selected even if the global null is true. Do the authors mean that the cutoff of the FDRS design is in general (much) smaller than gamma1 making this situation less likely (although not impossible)? More precise arguments are needed for the statements made on this matter.

*Indeed, for the FDRS design under the global null hypothesis the trial will continue with a probability of $\alpha_1$. In contrast, if a fixed p-value threshold is used, the probability to continue approaches 1 as the number of hypotheses increases (and assuming the test statistics are sufficiently independent). This statement is now also given in the introduction.*

(3) On page 5, it is then stated that when using an alpha1-level for the BH procedure, the experiment is stopped under the global null with probability $1 - \text{alpha1}$. Does alpha1 represent the FDR level or the p-value threshold? If alpha1 is the latter, I would think that under the global null, the experiment is stopped with probability $(1 - \text{alpha1})^m$ (for independent tests). If alpha1 represents the
FDR (I assume it does), I'm not sure how 1-alpha1 is obtained.

Indeed, alpha1 represents the FDR. Note that under the global null hypothesis (i.e., assuming that for all hypotheses the null hypothesis holds) the FDR is equal to the familywise error rate, i.e. the probability to reject any null hypothesis (this holds because in this setting V/R is equal to 1 whenever a hypothesis is rejected such that FWE = P(V > 0) = E(V/ max{R, 1}). We extended the explanation in the introduction.

(4) On page 2, in section "The test problem", it is stated that the BH procedure controls the FDR under positive regression dependency. In section "FDR control" on page 6, a more elaborate discussion is given on dependency assumptions (weak dependency etc.) and FDR control. It might be good to move this to "The test problem" as this is a general discussion on FDR control that applies to all approaches considered here. Then at the start of the section "FDR control", the authors could indicate that results are derived for independent tests but that given the fact that genetic data are considered, the performance under correlation is studied through simulations.

We changed the text accordingly.

Results: (5) In their report, the authors describe a heuristic explanation on why the the FDR is inflated for small m2 for the FNS procedure. I think it would be good to make clear that this inflation is to be expected in case of m2 and maybe further explain this in the appendix.

We added a heuristic explanation for the inflation.

(6) For the FDRS procedure, the authors mention that the FDR control is inflated for large pi0 and refer to additional file 1. Is it possible that this file only contains the corrected results (no inflation)?

Thank you for pointing this out. Indeed, the additional file contains only the results of the modified procedure where no inflation is observed. We corrected the text accordingly.

(7) The heading of table 4 and 5 state that results are shown for independent scenarios; this should be corrected.

We corrected the text accordingly.

Discussion:

(8) The comparison with one-stage designs is briefly mentioned in the discussion; it might be good to add the main findings on this. Furthermore, the place where this is mentioned interrupts the flow of the discussion. I suggest to discuss this
after the discussion of adaptive procedures.

We moved the paragraph as suggested and added a general statement regarding the efficiency of the single compared to the two stage designs. To state the results in more detail would require to give (quite lengthy) details of the single stage testing procedures and we refer to the additional file for details.

(9) On page 12, last line, it is mentioned that additional file 3 contains material on adaptive procedures. I don’t see why this should be mentioned at this point as this is only discussed in the following paragraph and interrupts the flow of the discussion. Please remove this sentence.

We deleted this sentence as suggested.

Minor comments: (10) Page 3, first paragraph, there is a typo: "eventhough" instead of "even though".

(11) Page 6, last paragraph, line 4: please remove "with fixed selection boundary" as this is already mentioned in the first sentence of this paragraph.

(12) Page 9, there is no point to end the sentence "In the following, we assume..."

(13) Page 12, second last line: it states "singel" instead of "single". However, as stated in a previous comment, I would suggest to move the discussion on one-stage designs to a different paragraph.

(10)-(13): We corrected the sentences.

Reviewer number: 2 Reviewer’s report:

1. In abstract, “Results: For both selection rules the multiple testing procedures control the FDR in the considered scenarios.” and “Conclusion: The proposed hypothesis tests provide a tool for FDR control for the considered two-stage designs.” These will give the impression that FNS and FDRS with integrated approach will control FDR. You should point out these only hold “with modifications” (requirement on $m_s$) and the conclusion are drawn through simulations.

In the revised abstract we state more clearly that the results are based on simulations. We do not refer to the different testing procedures in the abstract as we cannot give sufficient detail in this place. Instead we moved the description of the modified test procedure for the FDRS design to the main methods section such that it becomes clearer that this is the proposed procedure.

2. It is somewhat misleading to say that $\gamma_1$ converges to fixed value in FNS (page 7). The argument in your appendix need the assumption that the proportion of 2nd stage tests $(m_2/m)$ converges. Then it is not FNS anymore but rather
a fix proportion. In practice, \( m_2 \) is usually dictated by resource available and likely to be constrained. Thus, your argument that \( \gamma_1 \) is fixed asymptotically is not practically meaningful. It seems to me you don’t really need \( \gamma_1 \) to be fixed because results are mostly based on simulation.

*We agree that the asymptotic results has limited relevance for the practical application. However, we believe that the arguments give a heuristic motivation for the proposed test procedures and therefore decided to leave them in the manuscript (see also replies to Referee 1)*

3. I am not sure the R code for computing sequential p-value is publically available already. If not, authors should make it (the part that works with normal distributed data) available. This is because one big barrier for practitioner to use the proposed methods will be the unfamiliarity/difficulty of computing sequential p-values.

*We made the code available on our website and added a corresponding reference.*

4. Thanks for performing the adaptive procedures. First of all, are the adaptive procedures applied to FDRS with \( m_s > 5 \) or without \( m_s \) restriction. Second, author should include the adaptive result in the appendix. They are improvement over the vanilla BH, at least in the independent cases.

*All adaptive results are presented for \( m_s = 6 \). We clarify this in the additional file and insert results as suggested.*

5. Additional file 2 where FDR results are plotted, there should be more descriptions alongside. It is impossible for reader to know what they are for without looking back main text. Authors could move some/most of the description for additional file 2 in the main text into the actual file.

*We removed table 1 from the manuscript to additional file 1 (former additional file 2) and added some more descriptions.*

6. Some FDR plots looks strange, page 18 of Additional file 2, the bottom two plots on the left column (FDRS) have unusual fluctuations, any explanation why this might happen?

*By mistake, for \( m = 100000 \) the number of simulation runs was too small in some scenarios. We updated the simulations.*

Minor:

1. Page 8, last sentence of the first paragraph, 'the FDR is still controlled at level \( \alpha \) if the FDR is still controlled at level \( \pi_0 \alpha \) it is more correctly to say \( \pi_0 \alpha \).
2. Many grammar errors and typos can be found throughout the paper; I suggest authors proofread more carefully. For example: 1). page 3, However these approaches do not make full use of the available data because the first stage observations are used for selection only. The placement of the comma is awkward. However, these approaches do not make full use of the available data because the first stage observations are used for selection only. is more correct. 2). page 7, in principle also hypotheses that have not been selected in the interim analysis can in principle be rejected in the final test. Using two in principle in one sentence seems redundant. 3) page 13, We thank the two Reviewer for many helpful suggestions.

We changed the minor points as suggested and the manuscript was edited by a native speaker.