**REVIEWER 1:**

*The authors use the term "mutation(s)" and/or "polymorphism(s)" which creates confusion since sometimes it is used to indicate "a change" while in other disciplines it is used to indicate "a disease-causing change". HGVS recommends the use of the term "variant(s)"*

We thank the reviewer for the suggestion and we made changes of the terms on lines xxx, xxx and xxx.

*The authors use the term "pathogenic variant" which creates confusion. While a non-expert concludes the variant described "causes disease", the expert probably means "causes disease when in a specific context". To prevent confusion the HGVS recommends the use of a neutral term like "affects function".*

We thank the reviewer for the suggestion and we made changes of the terms on lines xxx, xxx and xxx.

**REVIEWER 2:**

*This is an interesting manuscript describing an approach to linkage analysis of rare variants. The authors are correct to point out that, due to their low heterozygosity, rare variants tend to provide low amounts of information to linkage analysis. So, an approach to combine several closely spaced variants (in the same gene) into a "super-locus" is likely to be much more powerful than an analysis based on individual rare variants.*

We thank the reviewer for the positive comments.

*There is no doubt that the authors' approach is more powerful than marker-by-marker analysis. However, similar approaches have previously been taken and the authors should compare their approach to the following older method: For some closely spaced variants, a very small recombination fraction is specified between any two adjacent markers and these recombination fractions are fixed, not estimated. The only thing estimated is the recombination fraction between the disease locus (to the left of the set of marker loci) and the nearest marker. In this manner, a 2-point analysis is effectively carried out, but the calculations are multipoint. The advantage of this approach is that no haplotypes need to be estimated as the analysis procedure will automatically form haplotypes in the course of the likelihood calculations, and missing information is properly taken into account. The authors' approach requires estimation of haplotypes and imputation of missing data, both of which may introduce biases.*

We thank the reviewer for pointing out the alternative approach and we’ve included a power comparison in the text. [FIXME: comment on power.]. We also clarified that [FIXME: explain that phase ambiguity is taken account for by taking weighted average of LOD over all possible phase configurations] and that [FIXME: what imputation means in this context]. Additionally [FIXME: comment on type I error, and within-gene recombination].

**REVIEWER 3:**

*The proposed approach is simple and practical and may be useful. However, the method is described too briefly to be fully understandable. A graphical presentation of how the collapsing is done, and how the resulting haplotypes are labeled, would be very useful to make the method easier to grasp.*

The manuscript was written in brevity due to the length constraint by the “short report” format of EJGH. We think the reviewer’s recommendation on using a graphical illustration to the method is excellent, thus we added Figure 2 as suggested, with more details on the CHP method in the figure captain.

*Additional details should be provided, such as allele-frequency cutoffs.*

We will address this issue later in our discussion on analyzing rare vs. common variants in sequence data.

*What happens in an area where a recombination event is observed? (The text on lines 136-138 on this is not clear: "For regions with recombination events, the sub-unit that shows the strongest evidence of linkage among all sub-units created by recombination events is chosen to represent the entire region." Is this not some kind of maximizing the test statistic that is not corrected for?)*

Recombination events within genes, though very rare, can occur in data and have to be considered by the CHP method when creating regional markers. This is done for per-family data and we edited the text to make this point clear (lines xxx - xxx). We believe using the maximum LOD will not create a bias. Under the assumption of no linkage, all sub-units created by recombination events will result in LOD scores close to 0, with negligible contribution to the combined LOD score across. When there is linkage, all sub-units other than the one containing causal variant will have LOD score close to 0 and can safely be ignored because they do not carry any information on linkage. Instead recombination events may results in a power loss under compound recessive model when two causal variants are separated into different sub-units due to recombination.

To validate the point we simulated recombination event with in families and evaluated the type I error. We found that type I error is controlled in this setup. Details see Table 1.

*Critically, the paper shows comparative power estimates under the alternative hypothesis of linkage, but not under the null hypothesis of no linkage. To evaluate the merits of the approach, it is important that the behavior of the approach be examined under both hypotheses. Both null and alternative hypothesis behavior need to be examined before robust determination of utility of the approach can be made.*

We agree with the reviewer that type I error evaluation should be performed and we added an empirical type I error calculation in Table 1.

*The equation on line 81 seems to assume that rare variants occur independently from one another. Could it not be the case that, say, two rare variants occur together on the same haplotype? In this case, the equation would lead to an underestimation of the haplotype frequency, which would lead to a bias in favor of linkage in the analysis. More generally, it is appropriate to ignore linkage disequilibrium between rare variants?*

We agree with the reviewer that linkage disequilibrium should not be ignored and we revised the equation to properly calculate the regional marker frequency. This change has also been implemented in our software SEQLinkage version 1.0.0.

*"For collapsed haplotype patterns within a pedigree, the frequencies will be normalized such that they sum to 1": It seems to me that the haplotype frequencies will sum to 1 by design. Is this not the case?*

When MAF of variants estimated from the population are used to compute haplotype frequencies, the resulting haplotype frequencies are also population frequencies which will only sum to 1 if all possible haplotype configurations are considered. This is not the case for data in a family because there are only limited haplotype patterns observed per family each with very small population frequency. Normalizing haplotype frequency to sum to 1 is essentially to compute the haplotype frequencies conditional on observations in given family. We have edited the text to clarify the point (lines xxx - xxx).

*"For variants having high MAFs it is not advisable to include them in regional marker generation, as their genotypes may predominate the marker pattern. We therefore exclude variants above a specified MAF cutoff and these markers are analyzed individually.": All markers jointly reflect the transmission of chromosomes from parents to offspring. Do you advocate analyzing the same region twice, with inheritance inferred in a non-redundant manner from two different types of variants?*

We assume here that common variants do not cause Mendelian diseases, which is reasonable assumption under most circumstances. When a “binning” theme is applied (see Figure 1) the presence of common variants may mask out the contribution of rare variants and attenuate linkage signals (i.e. producing negative LOD scores). If a gene harbors a rare disease causing variant, it is certain that linkage signal from rare variants analysis via CHP method will be stronger than from common variants in the same gene when analyzed separately so there is no need to consider results from common variants. Our suggestion of analyzing common variants separately was to consider scenarios when there is no rare variants available in a gene, in which case the common variants may tag the disease causing rare variants. We have revised the paragraph in question (lines xxx - xxx) to clarify the point.

*"Although it has been shown that analyzing single nucleotide variants (SNVs) from WES data provides acceptable linkage results, due to the low heterozygosity of SNVs this approach can be less powerful than analysis of SNPs from genotyping arrays.": This statement should be qualified. It is true that rare variants are not very informative regarding transmission simply because most individuals are homozygous for the common allele. However, if a rare variant is, in fact, the functional site, then only the heterozygotes provide linkage information, and analyzing the variant by itself should then be much more powerful than analyzing common nearby variants.*

We agree with the reviewer that analyzing a causal rare variant by itself is more powerful than analyzing a nearby common variant. However when there is allelic heterogeneity (under compound recessive model, or the causal variants are different among families but are within the same gene), analyzing common variants across families may result in a larger combined LOD score. We have made a clarification in the text (lines xxx - xxx). Of course for such situation our proposed method would perform the best.