

**Suzanne M. Leal, Ph.D.**

Professor

Director, Center for Statistical Genetics

Department of Molecular and Human Genetics

One Baylor Plaza, Suite 700D

Houston, Texas 77030-3498

TEL: (713) 798-4011

FAX: (713) 798-4373

E-mail: sleal@bcm.edu

November 6, 2014

Prof. Gertjan van Ommen  
Editor in Chief, European Journal of Human Genetics

Re: Manuscript

Dear Gert,

Thank you for the opportunity to resubmit our manuscript *Collapsed Haplotype Pattern Method for Linkage Analysis of Next-Generation Sequence Data*. We appreciate the reviewers’ comments, questions and suggestions. Some of our responses to the reviewers led to modifications of our manuscript and others, due to space limitations, led us to addition of a supplemental document.

For additional considerations, in this updated manuscript we added a table for sample size estimations when the disease causing variants in a large proportion of families are missing, to demonstrate the robustness of our proposed method to missing data which is commonplace in sequence data of human pedigrees. We believe our linkage method to analyze sequence data will be of practical interest to the readers, as such an important issue cannot be rectified by variant filtering strategies and classic two-point linkage analysis.

Below we provide detailed responses to the reviewers’ questions.

Sincerely,



Suzanne. M. Leal, Ph.D.

Professor

Department of Molecular and Human Genetics

**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 now use the term variant throughout the manuscript.

*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 as suggested 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. However, multipoint linkage is not as powerful as our method, and is not valid to use under some circumstances, e.g. when there is missing parental genotype data. Specifically, we investigated performance of multipoint linkage analysis using the Genehunter program. As the reviewer suggested we provide genetic map distances for each variant which in effect specifies a very small inter-marker recombination fraction. Power comparisons were performed for an autosomal recessive model suggests that although multipoint linkage is more powerful than single marker analysis, our CHP method is more powerful than multipoint linkage analysis. (Table S1). Regardless of statistical power, caution has to be used when performing multipoint linkage analysis on sequence data because when parental genotypes are missing and haplotype frequencies are inaccurately estimated or intermarker linkage disequilibrium is ignored type I error for multipoint linkage can be extremely inflated. We did not investigate the inflation of type I error when there is missing parental genotypes because this is already well documented in the literature (Huang *et al.*, 2004). The majority of available linkage programs , e.g. only allow for performing multipoint linkage analysis assuming inter-marker linkage equilibrium and therefore can have increased type I error. Even for programs that model inter-marker LD, e.g., LINKAGE/FASTLINK or Merlin, the haplotype frequency estimates involving rare variants can be inaccurate for studies with limited number of founders, leading to inflated type I error. Our method to estimate haplotype frequency using rare variants frequencies from public database to ensure that parametric linkage is properly performed and type I error is well controlled when some parental genotypes are missing from data. We have added description for type I error simulation and evaluations in the revised manuscript (lines xxx). Discussions on issues with multipoint linkage are added to the text in the following lines (lines xxx).

We agree with the reviewer and agree that missing data should not be “imputed” missing genotypes for individuals are only reconstructed if this can be done with a 100% certainty from the children. If there are Mendelian inconsistencies the variant sites is removed. Please find the revision on line XXX.

Reference:

Huang,Q. *et al.* (2004) Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am. J. Hum. Genet.*, **75**, 1106–1112.

**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 a figure (Figure 1 in the revised manuscript) as suggested, with details explaining the CHP method in the caption.

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

We now define what is meant by a rare variant in the introduction see line XX. Additionally recommendations for allele frequency cutoffs are now made in the Discussion see line XX.

*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 have to do so to make it possible to combine LOD scores from all families. We edited the text to make this point clear (lines xxx - xxx). We believe using the largest LOD will not create a bias and have now evaluated type I error using an extreme recombination scenario where for each family there are multiple recombination events within a region and the recombination breakpoint varies between families. We demonstrate that there is no increase in type I error please see Materials and Methods lines XXX and Results lines XXX. *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 have now performed extensive type I error evaluations please see the Materials and Methods lines XXX and the Results sections lines XXX.

*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 our methods such that the regional marker frequencies are now properly calculated (lines xxx - xxx). This change has also been implemented in our software SEQLinkage version 1.0.0. Additionally we have demonstrated in our type I error evaluation that when there are missing genotypes for parental samples and using our updated method to infer marker frequencies in linkage analysis, the type I error is well controlled even when variants within the region are in complete LD with each other.

*"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 frequencies to sum to 1 is essentially to compute the haplotype frequencies conditional on observations in given family, and naturally haplotype frequency for wildtype-haplotype will still be the largest. 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 agree that clarification is necessary. In the revised manuscript we have moved these arguments to “Discussion” and expanded our explanations as follows (line xxx):

*"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. The problem is that we will rarely run into the situation when all families have the same causal variants and this is why analyzing individual rare variants can be less powerful than analyzing genotyping arrays. In the manuscript we do not suggest that analyzing common variants are more powerful than analyzing causal variants. In the revised manuscript we do discuss analyzing common variants as well as variant sites which are likely to be functional see Discussion lines XXX.