

Computer Reconstructed Haplotypes in Pedigrees May Not Be Trusted

Wentian Li and Peter Gregersen

Center for Genomics and Human Genetics,
North Shore LIJ Research Institute

email: wli@nslj-genetics.org, peterg@nshs.org
web: www.nslj-genetics.org, www.naracdata.org

ABSTRACT

Genetic association analysis has become increasingly more important in mapping genes for human complex diseases, partly due to the availability of more densely spaced single-nucleotide-polymorphism (**SNP**) markers. Using single marker for association analysis may not be powerful if the marker is not very informative. It is currently a common practice to combine several marker for **constructing haplotypes** by a computer program. In our work with the North American Rheumatoid Arthritis Consortium (**NARAC**), we have observed that different computer programs may lead to different reconstructed haplotypes. This occurs more often if parents in a pedigree are not typed. We studied the relationship between the accuracy in the reconstructed haplotype with the percentage of the untyped persons in a pedigree. This study points to a practical implication that typing parents are important for reconstructing haplotypes.

Why Construct Haplotypes

- Current genotyping technology is unable to tell whether an allele is paternally derived or maternally derived (“phase” of an allele).
- Knowing the parental origin of alleles of several neighboring markers (“haplotypes”) may make the genetic association analysis more transparent, more effective.

How To Construct Haplotypes

- By computer programs

Two Different Situations Where Computer Programs Are Used To Construct Haplotypes

- 1. Unrelated Samples:** Each genotype is obtained from an unrelated person in a population.
- 2. Related Samples:** Each genotype is obtained from a person in a pedigree.

In situation #1, haplotype reconstruction is probability-based. Consequently, it is good in estimating haplotype frequencies, whereas not as good in reconstructing individual haplotypes.

In situation #2, because relatives' genotypes impose a constraint on the possible haplotype one may have, the individual haplotype reconstruction is more reliable.

We have observed that even for the pedigree data (situation #2), individual haplotype reconstruction by computer programs may not be reliable if there are enough untyped persons in the pedigree.

DATA and METHOD

- **Pedigrees:** 469 pedigrees collected by the North American Rheumatoid Arthritis Consortium (NARAC), most of them two-generation pedigrees with two or more affected children.
- **Markers:** 54 markers located on MHC/HLA region on chromosome 6.
- **Computer programs:**
 - (1) GENEHUNTER: one maximum-likelihood haplotype reconstruction is provided;
 - (2) SIMWALK2: the most likely haplotype reconstruction is provided, but it is based on a Monte-Carlo simulation. Different random seed in principle may lead to different outputs.
- **Reliability of a haplotype reconstruction** is based on the comparison of the haplotype constructed by the two computer programs. For a reliable reconstruction, two programs should lead to the same haplotype output.

comparison between GH and SW

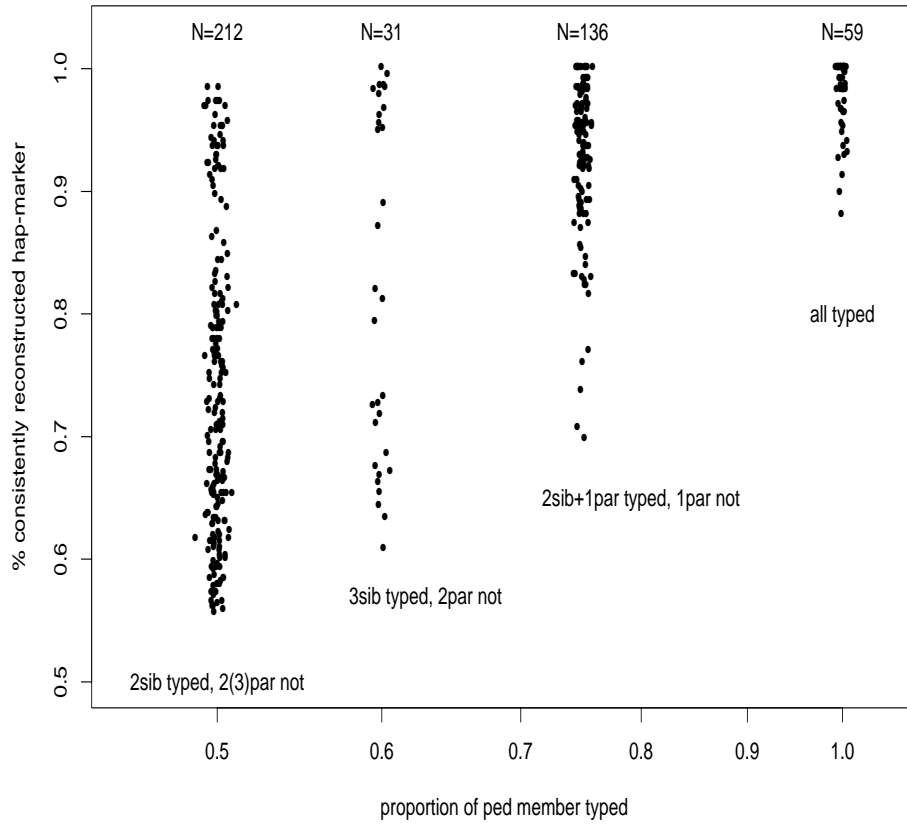


Fig.1

Fig.1 Comparison of haplotypes reconstructed by two computer programs, SIMWALK2 and GENEHUNTER, on the level of marker allele. **Each point represents a pedigree** and the height of the point (y value) is the **fraction of matched alleles in the two reconstructed haplotypes, averaged over all persons in the pedigree**. Types of pedigree are indicated by the **proportion of persons genotyped** (x value). For example, if $x=0.5$, half of the persons in the pedigree are typed, with most of them being two-sib-two-parent pedigrees and the two siblings are typed whereas the two parents are not. Small noise is added to the x value to separate overlapping dots. The number of pedigrees in each type is labeled by the N value on the top.

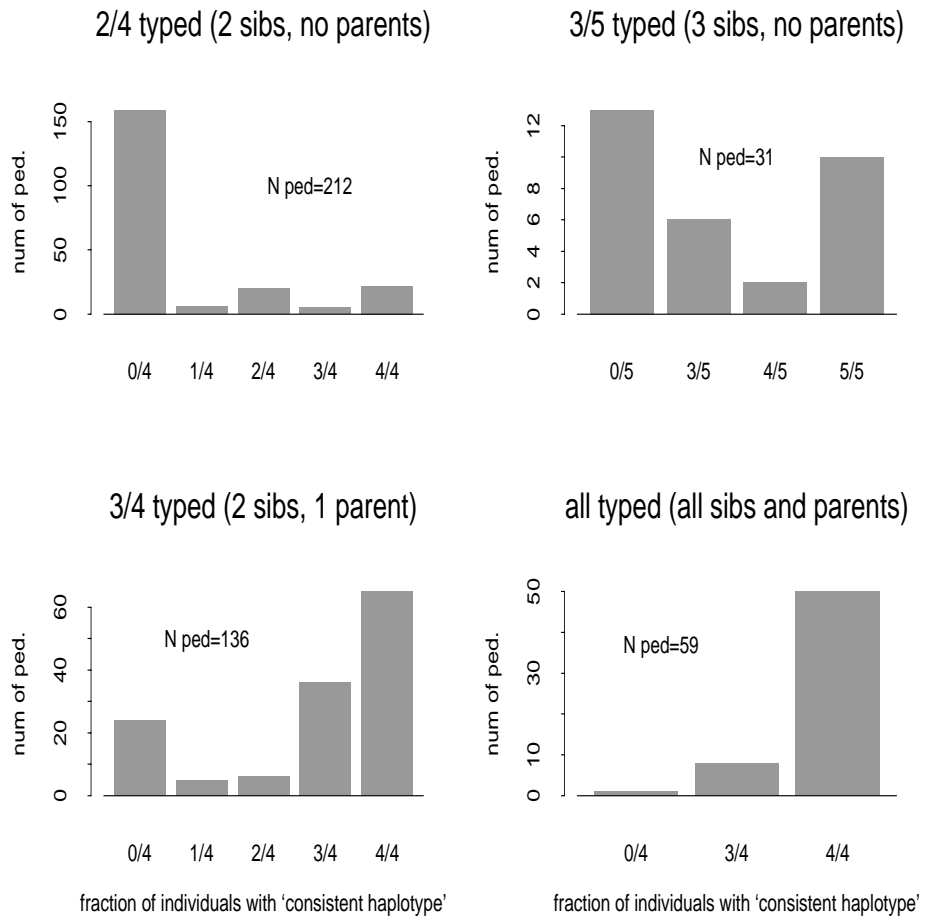


Fig.2

Fig.2 Comparison of haplotypes reconstructed by two computer programs, SIMWALK2 and GENEHUNTER, **on an individual level**. Four types of pedigree genotyping status are in four panels (2 out of 4 persons in the family typed, 3 out of 5, 3 out of 4, and all typed). For each type of pedigree, **the number of pedigrees that have 0, 1, 2... persons with consistently reconstructed haplotypes** is displayed in a bar plot, where “consistently reconstructed” means 49 out of 54 marker alleles being identical.

Conclusion and practical implications

Using a large number of pedigree data, we have shown that haplotype reconstruction is less reliable if there are “enough” untyped persons in the data. Little attention was paid to this observation in the genetic analysis community. People may take for granted the reconstructed haplotype generated by a computer program.

We have also shown more specifically that, if two children are typed but two parents are not, the haplotype reconstruction by computer programs is not reliable. In this situation, (1) it is important to get one parent typed; (2) or, if parents' DNAs are not available, skip this pedigree; (3) or, switch to genotype-based association analyses.

References

W Li, PG Gregersen, “Reconstructing haplotypes in pedigrees: the importance of parental information”, submitted to *Am. J. Med. Genet.* [**same as this poster**]

D Jawaheer, W Li, RR Graham, W Chen, A Damle, X Xiao, J Monteiro, H Khalili, A Lee, R Lundsten, A Begovich, T Bugawan, H Erlich, JT Elder, LA Criswell, MF Seldin, CI Amos, TW Behrens, PK Gregersen (2002), “Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis”, *American Journal of Human Genetics*, 71:565-574. [**genetic association analysis of these 54 markers**]