# Simple haplotype analyses in *R*

**Benjamin French[1,*], Nandita Mitra[1], Thomas P Cappola[2], Thomas Lumley[3]**

1. Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA USA
2. Penn Cardiovascular Institute, Philadelphia, PA USA
3. Department of Statistics, University of Auckland, Auckland, New Zealand

*Contact author: bcfrench@upenn.edu

**Keywords:** Statistical genetics, regression models

Statistical methods of varying complexity have been proposed to efficiently estimate haplotype effects and haplotype-environment interactions in case-control and prospective studies. We have proposed an alternate approach that is based on a non-iterative, two-step estimation process: first, an expectation-maximization algorithm is used to compute posterior estimates of the probability of all potential haplotypes consistent with the observed genotype for each subject; second, the estimated probabilities are used as weights in a regression model for the disease outcome, possibly including environmental factors. Standard error estimates are based on a robust variance estimator. We have shown that the two-step process provides valid tests for genetic associations and reliable estimates of modest genetic effects of common haplotypes for case-control studies (French et al, 2006). The two-step process has also been applied to prospective studies with a survival outcome subject to censoring (Neuhausen et al, 2009). An advantage of the two-step process is its straightforward implementation in software, so that analyses combining genetic and environmental information can be conducted by researchers expert in that subject matter using standard software, rather than by statisticians using specialized software. We illustrate the use of the two-step process for case-control studies using our *R* package **haplo.ccs**, which implements weighted logistic regression, and for prospective studies with a survival outcome using our working *R* package **haplo.cph**, which implements weighted Cox regression. We illustrate our method and software using data from a study of chronic heart failure patients (Cappola et al, 2011) to estimate the effect of *CLCNKA* haplotypes on time to death or cardiac transplantation.

## References

Cappola TP, Matkovich SJ, Wang W, et al. (2011). Loss-of-function DNA sequence variant in the *CLCNKA* chloride channel implicates the cardio-renal axis in interindividual heart failure risk variation. *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1017494108

French B, Lumley T, Monks SA, et al. (2006). Simple estimate of haplotype relative risks for case-control data. *Genetic Epidemiology* 30, 485–494.

Neuhausen SL, Brummel S, Ding YC, et al. (2009). Genetic variation in insulin-like growth factor signaling genes and breast cancer risk among *BRAC1* and *BRAC2* carriers. *Breast Cancer Research* 11, R76.