Genet. Res., Camb. (2007), **89**, *pp.* 425–426.
doi:10.1017/S0016672308009622 Printed in the United Kingdom

Estimating the recombination parameter: a commentary on 'Estimating the recombination parameter of a finite population model without selection' by Richard R. Hudson

B. S. WEIR*
Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA

In 1987, Hudson proposed an estimator for the scaled recombination parameter $C=4Nc$, where N is the population size and c is the recombination rate between the two most distant of a set of segregating sites. This work came shortly after Kreitman (1983) published the first set of population genetic data at the DNA sequence level. Kreitman had been able to sequence 2·7 kilobases of the *Drosophila melanogaster* genome in 11 samples. It was felt at that time that population genetics was entering a new era, although Hudson cautioned that sufficiently large data sets for his new estimator 'may require prohibitively large research efforts'.

Hudson's estimator is based on the variance of the number of site differences between pairs of haplotypes and an estimate of the scaled mutation rate $\theta=4N\mu$. The variance of the number of differences had already been shown by Brown *et al.* (1980) to be a convenient single-statistic summary of all the pairwise linkage disequilibria among a set of loci. The need for such a statistic continues as there is still doubt as to how well two-locus associations capture the full multilocus structure. Hudson provided an elegant derivation of the expected value of his statistic as a function of the unknown value C. His method of moments approach to estimation has the great virtue of simplicity although it would not be expected to behave as well as the maximum-likelihood methods that he (Hudson, 1993) and others (e.g. Kuhner *et al.*, 2000; Wall, 2000; Fearnhead and Donnelly, 2001) developed later. Likelihood methods exploit all the information in a data set rather than just the information in a summary statistic and will do well provided the underlying evolutionary model is appropriate for the data being addressed. Writing 10 years after Hudson, Wakeley kept the same moment approach but provided modifications to Hudson's method that improved its performance.

Since 1983 the human genome has been sequenced, as have the genomes of several other species. There is now a '1000 genomes' project (http://www.1000genomes.org) under way for humans, and new sequencing techniques will make it possible very soon for population geneticists to obtain large samples of DNA sequence data. In 1987, Hudson wished for more extensive DNA sequence data but he could not have foreseen the remarkable explosion of intermediate data – single-nucleotide polymorphisms (SNPs). Human geneticists are now generating 1 million SNP profiles for samples of thousands of individuals. By 2002, Hudson had produced a simulation procedure for SNP data (Hudson, 2002), and this has been used in studies such as Li and Stephens (2003) to detect recombination rate 'hotspots'.

Hudson's 1987 paper has the hallmarks of a classic paper. It introduced a new and simple method for estimating recombination rates from population samples rather than from pedigree data. More sophisticated methods have since been introduced, including composite-likelihood (Hudson, 2001) and others reviewed by Hellenthal and Stephens (2006), but the original method still has utility in evolutionary studies (e.g. Meikeljohn *et al.*, 2004).

* e-mail: bsweir@u.washington.edu

Acknowledgements

This work was supported in part by National Institutes of Health (NIH) grant GM 075091. The assistance of Dr T. Bhangale was very helpful.

References

Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* **96**, 523–536.

Fearnhead, P. & Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.

Hellenthal, G. & Stephens, M. (2006). Insights into recombination from population genetic variation. *Current Opinion in Genetics and Development* **16**, 565–572.

Hudson, R. R. (1993) The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution* (ed. N. Takahata and A. G. Clark), pp. 23–36. Sunderland, MA: Sinauer Associates.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401.

Li, N. & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.

Meikeljohn, C. D., Kim, Y., Hartl, D. L. & Parsch, J. (2004). Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics* **168**, 265–279.

Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genetical Research* **69**, 45–48.

Wall, J. D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution* **17**, 65–163.