Na Li and Matthew Stephens on Modeling Linkage Disequilibrium

Yun S. Song[1]
Computer Science Division and Department of Statistics, University of California, Berkeley, California 94720, Department of Mathematics and Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

ORIGINAL CITATION

Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data
Na Li and Matthew Stephens
GENETICS December 1, 2003 **165:** 2213–2233

Probabilistic models have played an indispensable role in population genetics for close to a century. They provide a powerful lens through which one can investigate how various evolutionary forces interact and produce the intricate patterns of genetic variation in a population. Furthermore, through statistical inference, one can estimate evolutionary parameters from population genetic data and draw important biological conclusions. There is a formidable challenge in this effort, however. Although the models (*e.g.*, diffusion processes and coalescent theory) are relatively straightforward to describe, computing the relevant likelihoods for statistical inference is often intractable. This is especially true when recombination is taken into account. Indeed, a far-reaching consequence of recombination is that different loci can have different evolutionary histories, and this complication leads to an overwhelming explosion in the dimensionality of the model.

In their influential work, Li and Stephens (2003) proposed a simple and elegant approach to circumvent this computational challenge, leading to a paradigm shift in modeling genetic relatedness with large-scale data. Their key methodological insight was to construct an approximate probabilistic model that captures the essential features of a genealogical process with recombination but produces a dramatic computational speed-up. Specifically, they considered the problem of approximating the conditional sampling probability (CSP) of the next haplotype given the haplotypes that have already been observed. A useful approximation had been proposed by Stephens and Donnelly (2000) for the simpler case of completely linked loci. They suggested approximating the next haplotype h_k as an imperfect copy of one of the first $k-1$ haplotypes, $h_1, \ldots, h_{k-1}$, with copying errors corresponding to mutation. Fearnhead and Donnelly (2001) generalized this approach to incorporate recombination, assuming that haplotype h_k is generated by copying segments from $h_1, \ldots, h_{k-1}$, where recombination can change the haplotype from which copying is performed. The associated CSP can be computed efficiently using standard methods (dynamic programming applied to a hidden Markov model). Li and Stephens (2003) proposed a modification to the Fearnhead and Donnelly approximation to obtain a simpler generative model, thus providing a computational speed-up. More important, they showed how the CSPs could be combined to approximate the likelihood of the haplotype data under a model that allowed recombination rates to vary over short distances. This permitted effective inference of the fine structure of recombination rate variation from population genomic data. This clever approach opened up new avenues of statistical inference in population genetics.

The modeling framework proposed by Li and Stephens has had a profound impact. Their "copying" model has been extended and applied to a wide range of problems, including inference of gene conversion parameters (Gay *et al.*, 2007; Yin *et al.*, 2009), recombination rates in admixed populations (Hinch *et al.*, 2011; Wegmann *et al.*, 2011), human colonization history (Hellenthal *et al.*, 2008), fine population structure (Lawson *et al.*, 2012), and local ancestry in admixed populations (Sundquist *et al.*, 2008; Price *et al.*, 2009). The model has

doi: 10.1534/genetics.116.191817
[1]Address for correspondence: Department of Statistics, University of California, Berkeley 321 Evans Hall #3860, Berkeley, CA 94720-3860. E-mail: yss@berkeley.edu

also been used to phase genotype sequence data into haplotypes and impute missing data to improve the power of genome-wide association studies (Stephens and Scheet 2005; Marchini *et al.*, 2007; Howie *et al.*, 2009; Li *et al.*, 2010).

In addition to these impressive applications, the work of Li and Stephens has also fueled theoretical research on deriving improved approximations for CSPs directly from the underlying population genetic models (Griffiths *et al.*, 2008; Paul and Song 2010). This research has led to genealogically interpretable approximations (Paul *et al.*, 2011) that can be applied to more complex models, allowing inference of population demographic histories (Sheehan *et al.*, 2013; Steinrücken *et al.*, 2013, 2015), including variable population sizes, divergence times, and gene flow between populations. A recent extension (Rasmussen *et al.*, 2014) shows how related ideas can be used to develop a Monte Carlo algorithm to sample genealogical histories, which have applications in association mapping and detecting signatures of natural selection.

In summary, the paper by Li and Stephens contains groundbreaking work employing biologically motivated approximations to allow efficient statistical inference from genomic data using informative models. Their innovative modeling approach has facilitated the development of numerous useful analytical tools that scale up to whole genomes while capturing important features of realistic population genetic models. It is a must-read for anyone interested in developing inference methods in population genetics and computational biology.

Literature Cited

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics 159: 1299–1318.

Gay, J. C., S. Myers, and G. McVean, 2007 Estimating meiotic gene conversion rates from population genetic data. Genetics 177: 881–894.

Griffiths, R. C., P. A. Jenkins, and Y. S. Song, 2008 Importance sampling and the two-locus model with subdivided population structure. Adv. Appl. Probab. 40: 473–500.

Hellenthal, G., A. Auton, and D. Falush, 2008 Inferring human colonization history using a copying model. PLoS Genet. 4: e1000078.

Hinch, A. G., A. Tandon, N. Patterson, Y. Song, N. Rohland *et al.*, 2011 The landscape of recombination in African Americans. Nature 476(7359): 170–175.

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5(6): e1000529.

Lawson, D., G. Hellenthal, S. Myers, and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genet. 8(1): e1002453.

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

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34: 816–834.

Marchini, J., B. Howie, S. R. Myers, G. McVean, and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39(7): 906–913.

Paul, J. S., and Y. S. Song, 2010 A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. Genetics 186: 321–338.

Paul, J. S., M. Steinrücken, and Y. S. Song, 2011 An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187: 1115–1128.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5(6): e1000519.

Rasmussen, M. D., M. J. Hubisz, I. Gronau, and A. Siepel, 2014 Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10(5): e1004342.

Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194(3): 647–662.

Steinrücken, M., J. S. Paul, and Y. S. Song, 2013 A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theor. Popul. Biol. 87: 51–61.

Steinrücken, M., J. A., Kamm, and Y. S. Song, 2015 Inference of complex population histories using whole-genome sequences from multiple populations. bioRxiv. DOI: 10.1101/026591.

Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics. J. R. Stat. Soc. B 62: 605–655.

Stephens, M., and P. Scheet, 2005 Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. 76(3): 449–462.

Sundquist, A., E. Fratkin, C. B. Do, and S. Batzoglou, 2008 Effect of genetic divergence in identifying ancestral origin using hapaa. Genome Res. 18(4): 676–682.

Wegmann, D., D. E. Kessner, K. R. Veeramah, R. A. Mathias, D. L. Nicolae *et al.*, 2011 Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. 43: 847–853.

Yin, J., M. I. Jordan, and Y. S. Song, 2009 Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. Bioinformatics 25(12): i231–i239.

Other *GENETICS* Articles by N. Li and M. Stephens

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Gao, Z., D. Waggoner, M. Stephens, C. Ober, and M. Przeworski, 2015 An estimate of the average number of recessive lethal mutations carried by humans. Genetics 199: 1243–1254.

Hellenthal, G., J. K. Pritchard, and M. Stephens, 2006 The effects of genotype-dependent recombination, and transmission asymmetry, on linkage disequilibrium. Genetics 172: 2001–2005.

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

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Raj, A., M. Stephens, and J. K. Pritchard, 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197: 573–589.

Roychoudhury, A., and M. Stephens, 2007 Fast and accurate estimation of the population-scaled mutation rate, theta, from microsatellite genotype data. Genetics 176: 1363–1366.

Communicating editor: C. Gelling