# Pritchard, Stephens, and Donnelly on Population Structure

**John Novembre**[*,†,1]
*Department of Human Genetics and †Department of Ecology and Evolutionary Biology, University of Chicago, Illinois 60637
ORCID ID: 0000-0001-5345-0214 (J.N.)

**ORIGINAL CITATION**
Inference of population structure using multilocus genotype data
Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly
*GENETICS* June 1, 2000 **155**: 945–959

In essentially any species, genetic similarity among individuals is structured by the existence of subgroups and geographic isolation. For researchers, understanding this population structure can be of direct interest, a necessary waystation to further analyses, or a confounding nuisance. Regardless of the motivation, understanding population structure is an essential step for population genetic analysis. In 2000, Pritchard, Stephens, and Donnelly published one of the most widespread and important frameworks for addressing this task: the model-based clustering method known as *STRUCTURE* (Pritchard *et al.* 2000).

The birth of the method owes much to having the right expertise in a room. In September of 1998, Pritchard arrived for a postdoc in Oxford just as a workshop at the Newton Institute in Cambridge was starting. At the time, he was finishing up work on a test for cryptic population structure in disease association studies (Pritchard and Rosenberg 1999) and had become interested in clustering related individuals. At the workshop, he shared these interests with his new advisor Peter Donnelly and fellow postdoc, Matthew Stephens. Donnelly was deeply experienced with Bayesian models in genetics, including models for forensic samples of uncertain origins (*e.g.*, Balding and Donnelly 1995), and Matthew had written his PhD dissertation on Bayesian clustering using Markov chain Monte Carlo techniques (Stephens 2000). Bringing these backgrounds together, they hashed out the first model on a board within a couple of hours. Then, Pritchard implemented it with the help of Stephens over the next several days. The prototype worked well, and there were no real changes from what was written out in the first meeting to what was ultimately published (J. K. Pritchard and P. Donnelly, personal communication).

The novelty was in taking a Bayesian approach that assigns individuals to source populations or allows them to have proportional assignment of their ancestry to multiple populations (the "admixture model"). Their work followed that of those who had developed likelihood-based individual assignment to populations (Paetkau *et al.* 1995; Rannala and Mountain 1997), population mixture models (Smouse *et al.* 1990), and Bayesian models of cryptic population structure (Foreman *et al.* 1997; Roeder *et al.* 1998). The resulting *STRUCTURE* method had a tremendous impact in human genetics, evolutionary genetics, and molecular ecology, and went on to be highly cited and awarded (Noor 2013). The admixture model is also used widely in machine learning where it is known as latent Dirichlet allocation (Blei *et al.* 2003). In addition, in the late 1990s, coalescent-based approaches dominated population genetic methods, and in this milieu the impact of *STRUCTURE* reinforced that relatively simple models can have tremendous utility.

After the initial publication, Pritchard and colleagues extended the work to include addressing ancestry along chromosomes (Falush *et al.* 2003), dominant markers and null alleles (Falush *et al.* 2007), and prior group information (Hubisz *et al.* 2009). Others developed extensions that, for instance, carry out assignment to hybrid categories (Anderson and Thompson 2002) or assume a spatial distribution for populations (Guillot *et al.* 2005; François *et al.*


doi: 10.1534/genetics.116.195164
[1]Address for correspondence: Department of Ecology and Evolutionary Biology, University of Chicago, Chicago, IL 60637. E-mail: jnovembre@uchicago.edu

2006; Durand *et al.* 2009; François and Durand 2010). Most recently there have been improvements to computational speed. Likelihood-based approaches (Tang *et al.* 2005; Alexander *et al.* 2009) and variational approximations (Raj *et al.* 2014) now make it feasible to analyze thousands of individuals (Novembre 2014).

One persistent challenge with applying *STRUCTURE* is inferring the number of source populations ($K$). Increasing the value of $K$ adds parameters, which can lead to overfitting the data, and so model-choice procedures are necessary to estimate $K$. Numerous procedures have been proposed (Evanno *et al.* 2005; Huelsenbeck and Andolfatto 2007), though the conventional wisdom is that reproducible inference of $K$ is a difficult problem, with less stability than conventional parameter estimation (*e.g.*, Gilbert *et al.* 2012). Pritchard, Stephens, and Donnelly were prescient and acknowledged problems in the inference and interpretation of $K$. They have long advocated instead using *STRUCTURE* as an exploratory tool and inspecting results from a range of values of $K$.

Another challenge is that *STRUCTURE* has become, in some sense, a victim of its own success. It is applied by default in most studies without consideration of whether the underlying model is relevant. For example, if applied to a geographic continuum, the method will infer source populations that are vaguely spatial but have no real interpretation as source populations in an admixed sample (*e.g.*, Witherspoon *et al.* 2007). A recent paper captures the care needed with its colorful title: "A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots" (Falush *et al.* 2016). All this being said, the need for careful interpretation is ubiquitous in population genetics, and the extra attention on the *STRUCTURE* method is warranted because of its widespread use. Looking forward, we can expect new methods for studying population structure will continue to leverage the principles of individual-based analysis and proportional ancestry so nicely deployed by Pritchard, Stephens, and Donnelly.

## Literature Cited


Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Anderson, E. C., and E. A. Thompson, 2002 A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160: 1217–1229.

Balding, D. J., and P. Donnelly, 1995 Inferring identity from DNA profile evidence. Proc. Natl. Acad. Sci. USA 92: 11741–11745.

Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003 Latent Dirichlet allocation. J. Mach. Learn. Res. 3: 993–1022.

Durand, E., F. Jay, O. E. Gaggiotti, and O. François, 2009 Spatial inference of admixture proportions and secondary contact zones. Mol. Biol. Evol. 26: 1963–1973.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14: 2611–2620.

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.

Falush, D., M. Stephens, and J. K. Pritchard, 2007 Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol. Ecol. Notes 7: 574–578.

Falush D., Daniel F., L. van Dorp, L. Daniel, 2016 A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. bioRxiv DOI: http://dx.doi.org/10.1101/066431.

Foreman, L. A., A. F. M. Smith, and I. W. Evett, 1997 Bayesian analysis of DNA profiling data in forensic identification applications. J. R. Stat. Soc. Ser. A Stat. Soc. 160: 429–459.

François, O., and E. Durand, 2010 Spatially explicit Bayesian clustering models in population genetics. Mol. Ecol. Resour. 10: 773–784.

François, O., S. Ancelet, and G. Guillot, 2006 Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics 174: 805–816.

Gilbert, K. J., R. L. Andrew, D. G. Bock, M. T. Franklin, N. C. Kane *et al.*, 2012 Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. Mol. Ecol. 21: 4925–4930.

Guillot, G., A. Estoup, F. Mortier, and J. F. Cosson, 2005 A spatial statistical model for landscape genetics. Genetics 170: 1261–1280.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard, 2009 Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. 9: 1322–1332.

Huelsenbeck, J. P., and P. Andolfatto, 2007 Inference of population structure under a Dirichlet process model. Genetics 175: 1787–1802.

Noor, M. A. F., 2013 The 2013 Novitski prize: Jonathan Pritchard. Genetics 194: 15–17.

Novembre, J., 2014 Variations on a common STRUCTURE: new algorithms for a valuable model. Genetics 197: 809–811.

Paetkau, D., W. Calvert, I. Stirling, and C. Strobeck, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. 4: 347–354.

Pritchard, J. K., and N. A. Rosenberg, 1999 Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65: 220–228.

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.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197–9201.

Roeder, K., M. Escobar, J. B. Kadane, and I. Balazs, 1998 Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika 85: 269–287.

Smouse, P. E., R. S. Waples, and J. A. Tworek, 1990 A genetic mixture analysis for use with incomplete source population data. Can. J. Fish. Aquat. Sci. 47: 620–634.

Stephens, M., 2000 Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. Ann. Stat. 28: 40–74.

Tang, H., J. Peng, P. Wang, and N. J. Risch, 2005 Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28: 289–301.

Witherspoon, D. J., S. Wooding, A. R. Rogers, E. E. Marchani, W. S. Watkins *et al.*, 2007 Genetic similarities within and between human populations. Genetics 176: 351–359.

## Further reading in *GENETICS*


Novembre, J., 2014 Variations on a common STRUCTURE: new algorithms for a valuable model. Genetics 197: 809–811.

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


## Other *GENETICS* articles by J. K. Pritchard, M. Stephens, and P. Donnelly


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.

Nyswaner, K. M., M. A. Checkley, M. Yi, R. M. Stephens, and D. J. Garfinkel, 2008 Chromatin-associated genes protect the yeast genome from Ty1 insertional mutagenesis. Genetics 178: 197–214.

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: M. Turelli*