Fumio Tajima and the Origin of Modern Population Genetics

Rasmus Nielsen
Departments of Integrative Biology and Statistics, University of California, Berkeley, California 94720 and Natural History Museum of Denmark, 1350 Copenhagen, Denmark
ORCID ID: 0000-0003-0513-6591 (R.N.)

ORIGINAL CITATION
Evolutionary Relationship of DNA Sequences in Finite Populations
Fumio Tajima
GENETICS October 1, 1983 **105:** 437–460

A fundamental concept in population genetics is the representation of the evolutionary history of a sample, of a single locus, as a tree. The theory describing such trees is called "coalescence theory," and the primary mathematical model used is called "the coalescent," or sometimes "Kingman's coalescent," as it was first discovered by the British mathematician Kingman (1982a,b). Coalescence theory is used to understand the statistical properties of a sample from a population and it underlies almost all the computational methods used for analysis of population-level DNA sequence data.

The discovery of Kingman's coalescent is arguably one of the most important theoretical discoveries in all of biology over the past 50 years. It was the culmination of decades of work on population genetic theory by Ewens (1972), Watterson (1975), Gladstein (1978), Griffiths (1980), and others. The basic idea of representing the history of a sample as a tree had been percolating for a while. For example, Gladstein (1978) described a process, akin to Kingman's coalescent, of loss of evolutionary lineages in the population over time. Griffiths (1980) derived mathematical properties of the tree structure of a sample, but used a more complicated, and less general, construction than the one eventually discovered by Kingman (1982a,b). Today, we celebrate the seminal contributions of Kingman in the development of the coalescent by appropriately naming the process after him. However, by the early 1980s the field had matured to such a degree that coalescence theory, in one form or another, was being developed independently by several researchers, including two graduate students: Hudson (1983a,b) working with John Gillespie at the University of California, Davis, and Tajima (1983) working with Masatoshi Nei at the University of Texas at Houston, both of whom would become central in the development of modern population genetic theory.

Tajima was trained in both phylogenetics and population genetics and was therefore well positioned to make inroads into problems regarding tree representations of the genealogical structure of a sample in a population. In his 1983 paper in *GENETICS* (Tajima 1983), he developed many of the most important results in coalescence theory, such as means and variances of the time to most recent common ancestor of the sample, and he illustrated how many classical population genetic results could be easily rederived using coalescence theory. He did so apparently independently of the work of Kingman, which he was unaware of at the time. In addition, he studied coalescence trees in models with two diverging populations and derived the probabilities of different tree topologies in this context. Probabilities of tree topologies in models with multiple populations (or species), were also an important part of the contemporaneous paper by Hudson (1983b). Together, these papers provided the first mathematical descriptions of tree structures caused by what will later become known as "incomplete lineage sorting" (ILS)—a very important concept in our understanding of phylogenetic

doi: 10.1534/genetics.116.195271
Image courtesy of Fumio Tajima and the University of Tokyo.
Address for correspondence: Department of Integrative Biology, UC Berkeley, 4098 Valley Life Sciences Bldg., Berkeley, CA 94720. E-mail: rasmus_nielsen@berkeley.edu

trees. They initiated decades of research on the interface between phylogenetics and population genetics and on understanding ILS and its consequences. However, this is not the main reason Tajima's (1983) paper became so highly cited. In a later section of the paper, Tajima provided the first derivation of the variance of the average number of pairwise differences (π) under the infinite sites model, and argued in favor of using π as an estimator of the mutation scaled effective population size (θ). This estimator came into common use and is now often referred to as "Tajima's estimator." It remains one of the standard statistical methods for analyzing population genetic data.

Tajima (1983) was one of the founding papers of modern population genetics and was arguably the first paper that truly demonstrated the tremendous power of the coalescent when deriving statistical properties of a sample of DNA sequences. It also introduced the problem of incomplete lineage sorting in biology. It remains one of the pillars of modern population genetics and should be required reading for any graduate student entering into the field of population genetics.

Literature Cited

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3: 87–112.

Gladstein, K., 1978 The characteristic values and vectors for a class of stochastic matrices arising in genetics. SIAM J. Appl. Math. 34: 630–642.

Griffiths, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Popul. Biol. 17: 37–50.

Hudson, R. R., 1983a Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23: 183–201.

Hudson, R. R., 1983b Testing the constant-rate neutral allele model with protein-sequence data. Evolution 37: 203–217.

Kingman, J. F. C., 1982a The coalescent. Stochastic Process. Appl. 13: 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. 19: 27–43.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Further reading in *GENETICS*

Kingman, J. F. C., 2000 Origins of the coalescent: 1974–1982. Genetics 156: 1461–1463.

Other *GENETICS* articles by F. Tajima

Innan, H., and F. Tajima, 1997 The amounts of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. Genetics 147: 1431–1444.

Innan, H., F. Tajima, R. Terauchi, and N. T. Miyashita, 1996 Intragenic recombination in the Adh locus of the wild plant *Arabidopsis thaliana*. Genetics 143: 1761–1770.

Innan, H., R. Terauchi, G. Kahl, and F. Tajima, 1999 A method for estimating nucleotide diversity from AFLP data. Genetics 151: 1157–1164.

Misawa, K., and F. Tajima, 1997 Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites. Genetics 147: 1959–1964.

Nei, M., and F. Tajima, 1981a DNA polymorphism detectable by restriction endonucleases. Genetics 97: 145–163.

Nei, M., and F. Tajima, 1981b Genetic drift and estimation of effective population size. Genetics 98: 625–640.

Nei, M., and F. Tajima, 1983 Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics 105: 207–217.

Noguchi, Y., K. Endo, F. Tajima, and R. Ueshima, 2000 The mitochondrial genome of the brachiopod *Laqueus rubellus*. Genetics 155: 245–259.

Tajima, F., 1989a DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 123: 229–240.

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.

Tajima, F., 1989c Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tajima, F., 1990a Relationship between DNA polymorphism and fixation time. Genetics 125: 447–454.

Tajima, F., 1990b Relationship between migration and DNA polymorphism in a local population. Genetics 126: 231–234.

Tajima, F., 1993 Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599–607.

Tajima, F., 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143: 1457–1465.

Tajima, F., and M. Nei, 1984 Note on genetic drift and estimation of effective population size. Genetics 106: 569–574.

Communicating editor: M. Turelli