

MR2785147 (Review) 62P10 (60J10 60J27 62D05 92D10)
Sainudiin, Raazesh (NZ-CANT-BM); **Thornton, Kevin** (1-CA3-EV);
Harlow, Jennifer (NZ-CANT-MS); **Booth, James [Booth, James G.]** (1-CRNL-BSC);
Stillman, Michael (1-CRNL); **Yoshida, Ruriko** (1-KY-S);
Griffiths, Robert [Griffiths, Robert C.] (4-OX-S); **McVean, Gil** (4-OX-S);
Donnelly, Peter (4-OX-S)

Experiments with the site frequency spectrum. (English summary)

Bull. Math. Biol. **73** (2011), no. 4, 829–872.

The paper is devoted to evaluating the likelihood function of parameters from a summary of the data such as the site-frequency-spectrum (SFS) or its linear combinations, at a non-recombining locus that is neutrally evolving under the infinitely-many-sites mutation model. First, the authors develop a Markov lumping of Kingman's n -coalescent to Kingman's unlabeled n -coalescent as suggested in [J. F. C. Kingman, *J. Appl. Probab.* **1982**, Special Vol. 19A, 27–43; [MR0633178 \(83d:92043\)](#)]. The latter is a Markov chain on a many-to-one map of the state space of the n -coalescent (or more specifically, the labeled n -coalescent) and it is sufficient and necessary to prescribe the family of measures for the sample space of the SFS. Second, they exactly evaluate the posterior density based on one or more linear combinations of the observed SFS. This is achieved by an elementary study of the algebraic geometry of such statistics using Markov bases [see P. W. Diaconis and B. Sturmfels, *Ann. Statist.* **26** (1998), no. 1, 363–397; [MR1608156 \(99j:62137\)](#)]. This reviewer thinks that this is an impressive piece of work.

Reviewed by *M. Iosifescu*

References

1. Bahlo, M., & Griffiths, R. (1996). Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* **57**, 79–95.
2. Barvinok, A. (1994). Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math. Oper. Res.* **19**, 769–779. [MR1304623 \(96c:52026\)](#)
3. Beaumont, M., Zhang, W., & Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.
4. Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a exible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* **19**, 2609–2625.
5. Birkner, M., & Blath, J. (2008). Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.* **57**, 435–465. [MR2411228 \(2009i:92046\)](#)
6. Cam, L.L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Stat.* **35**, 1419–1455. [MR0207093 \(34 #6909\)](#)
7. Casanellas, M., Garcia, L., & Sullivant, S. (2005). Catalog of small trees. In L. Pachter & B. Sturmfels (Eds.), *Algebraic statistics for computational biology* (pp. 291–304). Cambridge: Cambridge University Press. [MR2205880](#)

8. Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.* 26, 363–397. [MR1608156 \(99j:62137\)](#)
9. Duflo, M. (1997). *Random iterative models*. Berlin: Springer. [MR1485774 \(98m:62239\)](#)
10. Erdős, P., Guy, R., & Moon, J. (1975) On refining partitions. *J. Lond. Math. Soc. (2)* 9, 565–570. [MR0360302 \(50 #12752\)](#)
11. Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3, 87–112. [MR0325177 \(48 #3526\)](#)
12. Ewens, W. (1974). A note on the sampling theory of infinite alleles and infinite sites models. *Theor. Popul. Biol.* 6, 143–148.
13. Ewens, W. (2000). *Mathematical population genetics* (2nd edn.). Berlin: Springer. [MR0554616 \(81f:92019\)](#)
14. Fay, J., & Wu, C. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.
15. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.
16. Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* 23, 691–700.
17. Grayson, D., & Stillman, M. (2004). Macaulay 2, a software system for research in algebraic geometry. Available at www.math.uiuc.edu/Macaulay2.
18. Griffiths, R., & Tavaré, S. (1994). Ancestral inference in population genetics. *Stat. Sci.*, 9, 307–319. [MR1325431 \(96d:62213\)](#)
19. Griffiths, R., & Tavaré, S. (1996). Markov chain inference methods in population genetics. *Math. Comput. Modelling*, 23, 141–158. [MR1398007 \(97c:92007\)](#)
20. Griffiths, R., & Tavaré, S. (2003). The genealogy of a neutral mutation. In P. Green, N. Hjørt, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 393–412). London: Oxford University Press.
21. Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109. [MR2082417](#)
22. Hemmecke, R., Hemmecke, R., & Malkin, P. (2005). 4ti2 version 1.2—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de.
23. Hosten, S., Khetan, A., & Sturmfels, B. (2005). Solving the likelihood equations. *Found Comput. Math.* 5(4), 389–407. [MR2189544](#)
24. Hudson, R. (1993). The how and why of generating gene genealogies. In: Clark, A., Takahata, N. (Eds.) *Mechanisms of molecular evolution* (pp. 23–36). Sunderland: Sinauer. [MR1013322](#)
25. Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338.
26. Iorio, M., & Griffiths, R. (2004). Importance sampling on coalescent histories. I. *Adv. Appl. Probab.*, 36, 417–433. [MR2058143 \(2005b:60106\)](#)
27. Jones, G., & Hobert, J. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Stat. Sci.* 16(4), 312–334. [MR1888447](#)
28. Jukes, T., & Cantor, C. (1969). Evolution of protein molecules. In H. Munro (Ed.), *Mammalian protein metabolism* (pp. 21–32). San Diego: Academic Press.
29. Kemeny, Snell (1960). *Finite Markov chains*. Princeton: Van Nostrand. [MR0115196 \(22 #5998\)](#)

29. Kendall, D. (1975). Some problems in mathematical genealogy. In: Gani, J. (Ed.), *Perspectives in probability and statistics* (pp. 325–345). San Diego: Academic Press. [MR0420893 \(54 #8904\)](#)
30. Kingman, J. (1982a). The coalescent. *Stoch. Process. Their Appl.* 13, 235–248. [MR0671034 \(84a:60079\)](#)
31. Kingman, J. (1982b). On the genealogy of large populations. *J. Appl. Probab.* 19, 21–43.
32. Kolmogorov, A. (1942). Sur l'estimation statistique des paramètres de la loi de gauss. *Bull Acad. Sci. URSS Ser. Math.* 6, 3–32. [MR0007959 \(4,221e\)](#)
33. Loera, J. D., Haws, D., Hemmecke, R., Huggins, P., Tauzer, J., & Yoshida, R. (2004). Lattice Point Enumeration: LattE, software to count the number of lattice points inside a rational convex polytope via Barvinok's cone decomposition. Available at www.math.ucdavis.edu/latte.
34. Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* 100, 15, 324–15,328.
35. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092.
36. Mossel, E., & Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309, 2207–2209.
37. Mossel, E., & Vigoda, E. (2006). Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Probab.*, 16(4), 2215–2234. [MR2288719 \(2007k:60229\)](#)
38. Rosenblatt, M. (1974). *Random processes*. Berlin: Springer. [MR0346883 \(49 #11604\)](#)
39. Sainudiin, R., & Stadler, T. (2009) A unified multi-resolution coalescent: Markov lumpings of the Kingman-Tajima n -coalescent. UCDMS Research Report 2009/4, 5 April 2009 (submitted). Available at <http://www.math.canterbury.ac.nz/r.sainudiin/preprints/SixCoal.pdf>.
40. Sainudiin, R., & York, T. (2009). Auto-validating von Neumann rejection sampling from small phylogenetic tree spaces. *Algorithms Mol. Biol.* 4, 1.
41. Sainudiin, R., Clark, A., & Durrett, R. (2007). Simple models of genomic variation in human SNP density. *BMC Genomics* 8, 146.
42. Semple, C., & Steel, M. (2003). *Phylogenetics*. Oxford University Press, London. [MR2060009 \(2005g:92024\)](#)
43. Sisson, S., Fan, Y., & Tanaka, M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* 104, 1760–1765. [MR2301870 \(2008a:65009\)](#)
44. Slatkin, M. (2002). A vectorized method of importance sampling with applications to models of mutation and migration. *Theor. Popul. Biol.* 62, 339–348.
45. Stephens, M., & Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. B* 62, 605–655. [MR1796282 \(2001h:62047\)](#)
46. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
47. Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26, 119–164. [MR0770050 \(86f:92017\)](#)
48. Thornton, K., Jensen, J. D., Becquet, C., & Andolfatto, P. (2007). Progress and prospects in mapping recent selection in the genome. *Heredity* 98, 340–348.

49. Wakeley, J. (2007). *Coalescent theory: an introduction*. Greenwood Village: Roberts & Co.
50. Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7, 256–276. [MR0366430 \(51 #2677\)](#)
51. Weiss, G., & von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, 149, 1539–1546.
52. Yang, Z. (2000). Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Land. B Biol. Sci.* 267, 109–119.

Note: This list reflects references listed in the original paper as accurately as possible with no attempt to correct errors.

© Copyright American Mathematical Society 2011