www.sciencemag.org/cgi/content/full/336/6082/740/DC1

Supplementary Materials for

Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants

Alon Keinan* and Andrew G. Clark

*To whom correspondence should be addressed. E-mail: ak735@cornell.edu

Published 11 May 2012, *Science* **336**, 740 (2012)
DOI: 10.1126/science.1217283

This PDF file includes:

Supplementary Text
Fig. S1
References

Supplementary Text

Large sample size challenges population genetic assumptions

Consideration of very large sample size violates a basic assumption underlying the theory employed by most population genetic analysis tools. In particular, the Standard Coalescent theory assumes that the sample size is much smaller than the effective population size (*33-35*). This assumption leads to there being only two possibilities in each generation, either no coalescence events, or a single pair of lineages that coalesce, but no more. However, as the sample size approaches or exceeds the population size, the topology of the genealogy violates these assumptions.

At the tips of the genealogy, where the number of lineages is largest, the probability of many coalescence events is largest, including *multiple mergers*, which are the simultaneous coalescence of more than two lineages. On the other hand, for humans the effective population size has also been much larger recently, which reduces the probability of multiple mergers at the tips of the genealogy. However, since the growth of human populations has been so extreme and so recent, a large fraction of the many extant lineages will make it all the way back to when growth started. At that point in time, only about 400 generation ago, the effective population size is best estimated to be at most 10,000 (see also Table 1 in main text). Hence, multiple mergers present a potential problem both at the tips of the genealogy—due to the large number of lineages—and before the start of growth—due to the reduced population size at that time—and similarly in other epochs as well.

The theoretical foundation for overcoming this problem has been put forth in a generalized coalescent theory that considers the case of sample size larger than the effective population size (*14*). This generalization is based in turn on the Λ-coalescent that allows multiple mergers to occur (*15-17*).

Another issue is the fact that multiple mergers imply immediate coancestry in a pedigree sense since individuals must share a parent to have alleles that coalesce the previous

generation. The actual pedigree relationships of a sample are likely to preclude multiple mergers in the first generation. Aspects of the way that the sample pedigree constrains the coalescent were recently considered (*36*), showing that basically the times of coalescence are being pushed back a few generations.

Accounting for multiple mergers in large sample sizes also presents an interesting opportunity. Population genetics has been mostly concerned with the composite parameter $\theta = 4N_e\mu$, where μ is the per-site mutation rate. However, when the sample size is larger than the effective population size (N_e) and multiple mergers are accounted for, it is possible to decouple μ and N_e and estimate them separately (*14*).

Recent demographic modeling (*18*) has allowed for multiple mergers based on the Λ-coalescent using a population genetic model similar to those of Wakeley and Takahashi (*14*) and Boyko *et al.* (*37*). This study capitalized on this framework to estimate μ separately from N_e and reported mutation rate estimates in 'neutral' sites of $\mu = 5.1\text{x}10^{-8}$ and $4.9\text{x}10^{-8}$ per site for each of the two genes (*18*).

However, we note that while their modeling allowed for multiple mergers (*18*), it provided no evidence that a model lacking multiple mergers will fail to provide an equally good fit to observed data. Additionally, we are not aware of any evidence showing that the potential problem of multiple mergers will be an actual problem in studying *human* population genetics using a large sample size.

The skewed genealogical structure when the sample size exceeds the effective population size can also greatly affect the distribution of identical-by-descent blocks. Most notably, many coalescent events in very recent history, near the tips of the genealogy, leads to much extended identical-by-descent blocks. This phenomenon can be put to practical use for associating rare variants with complex disease in a large sample of whole-genome sequences using mapping methods that are based on identity-by-descent (*38-40*).

n=5 **n=50**

n=500 **n=5,000**

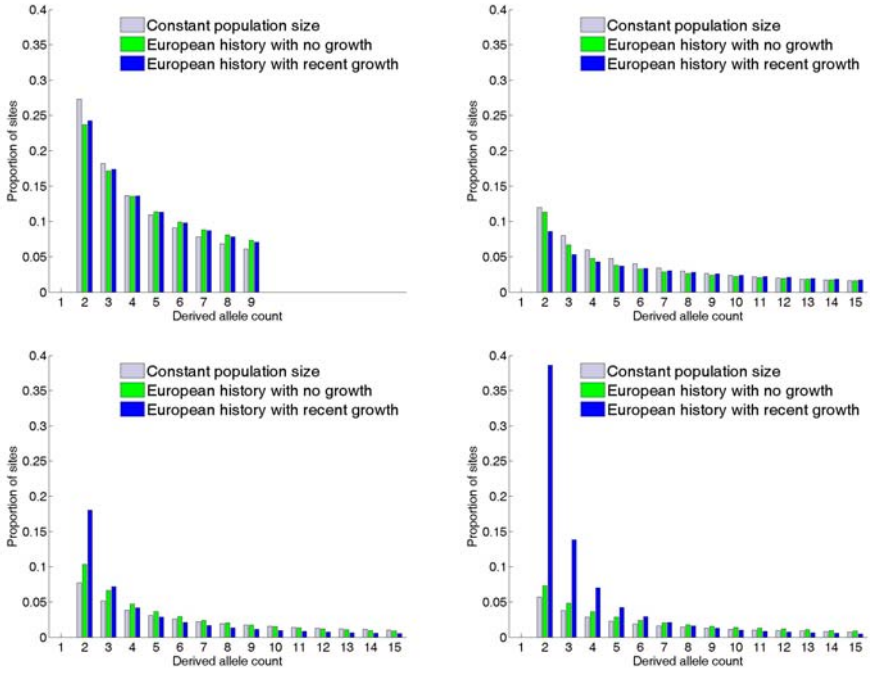

Fig. S1.
The expected site frequency spectrum of the derived allele when singletons (sites with derived allele count of 1) are excluded and the site frequency spectrum is renormalized to sum up to 1. Models, simulations, and presentation are identical to Fig. 2 in main text, except for the exclusion of singletons, which also results in different scale of the *y*-axis from that of Fig. 2.

References

1. J. E. Cohen, *How Many People Can the Earth Support?* (Norton, New York, ed. 1, 1995).

2. L. Roberts, 9 billion? *Science* **333**, 540 (2011). doi:10.1126/science.333.6042.540 Medline

3. United Nations Department of Economic and Social Affairs Population Division, 2011.

4. D. Hartl, A. Clark, *Principles of Population Genetics* (Sinauer, Sunderland, MA, 2007).

5. S. Gravel *et al*.; 1000 Genomes Project, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983 (2011). doi:10.1073/pnas.1019276108 Medline

6. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009). doi:10.1371/journal.pgen.1000695 Medline

7. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr., H. H. Hobbs, Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264 (2006). doi:10.1056/NEJMoa054013 Medline

8. K. A. Fawcett *et al*., Detailed investigation of the role of common and low-frequency WFS1 variants in type 2 diabetes risk. *Diabetes* **59**, 741 (2010). doi:10.2337/db09-0920 Medline

9. C. E. Glatt *et al*., Screening a large reference sample to identify very low frequency sequence variants: Comparisons between two genes. *Nat. Genet.* **27**, 435 (2001). doi:10.1038/86948 Medline

10. C. T. Johansen *et al*., Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684 (2010). doi:10.1038/ng.628 Medline

11. R. M. Durbin *et al*., A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline

12. J. Akey, Analysis of 2,440 human exomes highlights the evolution and functional impact of rare coding variation. *Genome Biol.* **12**, (Suppl 1), 1 (2011). doi:10.1186/1465-6906-12-S1-I1

13. See supplementary materials on *Science* Online.

14. J. Wakeley, T. Takahashi, Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.* **20**, 208 (2003). doi:10.1093/molbev/msg024 Medline

15. J. Pitman, Coalescents with multiple collisions. *Ann. Probab.* **27**, 1870 (1999). doi:10.1214/aop/1022677552

16. S. Sagitov, The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36**, 1116 (1999). doi:10.1239/jap/1032374759

17. J. Schweinsberg, Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5**, 1 (2000). doi:10.1214/EJP.v5-68

18. A. Coventry *et al*., Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010). doi:10.1038/ncomms1130 Medline

19. The ARIC Investigators, The Atherosclerosis Risk in Communities (ARIC) Study: Design and objectives. *Am. J. Epidemiol.* **129**, 687 (1989). Medline

20. J. Novembre *et al*., abstr. 6, presented at the 12th International Congress of Human Genetics, 12 October 2011, Montreal.

21. A. Keinan, J. C. Mullikin, N. Patterson, D. Reich, Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* **39**, 1251 (2007). doi:10.1038/ng2116 Medline

22. D. M. Altshuler *et al*.; International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010). doi:10.1038/nature09298 Medline

23. M. Lynch, Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 961 (2010). doi:10.1073/pnas.0912629107 Medline

24. T. A. Manolio *et al*., Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009). doi:10.1038/nature08494 Medline

25. K. A. Frazer, S. S. Murray, N. J. Schork, E. J. Topol, Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241 (2009). doi:10.1038/nrg2554 Medline

26. B. Maher, Personal genomes: The case of the missing heritability. *Nature* **456**, 18 (2008). doi:10.1038/456018a Medline

27. E. E. Eichler *et al*., Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446 (2010). doi:10.1038/nrg2809 Medline

28. J. A. Tennessen, T. D. O'Connor, M. J. Bamshad, J. M. Akey, The promise and limitations of population exomics for human evolution studies. *Genome Biol.* **12**, 127 (2011). doi:10.1186/gb-2011-12-9-127 Medline

29. S. F. Schaffner *et al*., Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576 (2005). doi:10.1101/gr.3709305 Medline

30. C. Haub, How many people have ever lived on Earth? *Popul. Today* **23**, 4 (1995). Medline

31. M. Kremer, Population growth and technological change: One million B.C. to 1990. *Q. J. Econ.* **108**, 681 (1993). doi:10.2307/2118405

32. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337 (2002). doi:10.1093/bioinformatics/18.2.337 Medline

33. J. Kingman, The coalescent. *Stochastic Process. Appl.* **13**, 235 (1982). doi:10.1016/0304-4149(82)90011-4

34. J. Wakeley, *Coalescent Theory: An Introduction* (Roberts, Greenwood Village, CO, 2005).

35. R. R. Hudson, *Oxford Surv. Evol. Biol.* **7**, 44 (1990).

36. J. Wakeley, L. King, B. S. Low, S. Ramachandran, Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* **190**, 1433 (2012). doi:10.1534/genetics.111.135574 Medline

37. A. R. Boyko *et al*., Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008). doi:10.1371/journal.pgen.1000083 Medline

38. S. Purcell *et al*., PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559 (2007). doi:10.1086/519795 Medline

39. A. Albrechtsen *et al*., Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.* **33**, 266 (2009). doi:10.1002/gepi.20378 Medline

40. S. R. Browning, B. L. Browning, High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* **86**, 526 (2010). doi:10.1016/j.ajhg.2010.02.021 Medline