Human evolution

Pedigrees for all humanity

Jotun Hein

Simulations based on a model of human population history and geography find that an individual that is the genealogical ancestor of all living humans existed just a few thousand years ago.

Writing on page 562 of this issue, Rohde, Olson and Chang[1] address a simple but fascinating question: how far back in time must we go to find an individual who was the ancestor of all present-day humans? After a little consideration, the existence of such an individual (the 'universal ancestor' or, as the authors put it, our 'most recent common ancestor') should not surprise: I have two parents, four grandparents, and the growth in the population of my ancestors is close to exponential as I trace them back in time. This is true for anybody's ancestors, and there must soon be an overlap between the ancestors of two or more randomly chosen individuals (Fig.1).

In simplified models, which assume random mating, the average number of generations back to a universal common ancestor has been estimated[2–4] to be around $\log_2 n$, where n is the population size. So if, for instance, the present-day population were to consist of 1,000 people, the average number of generations back to the universal ancestor would be $\log_2(1{,}000)$ — about 10 generations. For populations of size 10^6, or the present human population of size 6×10^9, it would be 20 or 33 generations, corresponding to 500 or a bit more than 800 years, respectively (assuming a generation time of 25 years). This is surprisingly recent.

And an even more surprising conclusion from such models is that, only a little farther back in time, a large fraction of the population will be the ancestors of everybody alive today. The remaining individuals back then will be the ancestors of no one. As Rohde *et al.*[1] describe it, "When genealogical ancestry is traced back beyond the [universal ancestor], more and more people in earlier generations become ancestors of the [whole] present-day population". At a certain point in history (the 'identical ancestors' point), people can be divided into two groups: either they are common ancestors of all present-day humans, or their lineages have died out. Being the ancestor of only some living humans is not an option. At this point, Rohde *et al.* say, "everyone alive now had exactly the same ancestors". In the simplest model, the fraction of 'ancestors-of-all' is about 80%, and in most estimates so far, the time back to the 'identical ancestors' point is a bit less than twice the number of generations back to the first universal ancestor.

These estimates are not only astonishing, however; they are also unrealistically low, because of the simplicity of the underlying models. Key missing factors are geography (which influences population structure) and history (which affects population growth), and these are the ingredients that Rohde *et al.* have taken seriously to arrive at more credible estimates of the time back to the universal and identical ancestors.

The authors carried out simulations based on several scenarios, incorporating different degrees of population growth and different degrees of isolation of subpopulations, with occasional migration linking these subpopulations. The authors' first model is relatively simple and includes up to ten large subpopulations, which exchange just one pair of migrants per generation. In one set of estimates based on this model, the mean time back to the universal ancestor is 2,300 years (76 generations, assuming a generation time of a bit less than 30 years) and to the identical ancestors it is 5,000 years (169 generations) — the time of Aristotle and the first pyramids, respectively. The latter date is especially startling: had you entered any village on Earth in around 3,000 BC, the first person you would have met would probably have been your ancestor! A considerably more detailed model, which describes population density within continents, the opening of ports and more, does not change these estimates much.

The main weakness in the models comes from migration. As the authors point out, if one region is totally isolated (something that they do not simulate), with no migrants connecting it to other subpopulations, then the universal ancestor must logically have lived before the period of isolation began. Only after that period ends would the dates for the universal ancestor become less distant. Because of the effects of isolation, had we been living in 1700, say, and tried to work out when our universal and identical ancestors lived, the answers would have been further back in time than the answers we obtain now. Tasmania, for instance, was conceivably completely isolated at the time, and probably had been for millennia; this would therefore have pushed back the dates for universal and identical ancestry. So uncertainties about population structure introduce uncertainty into the proposed dates.

The genealogical questions addressed by Rohde *et al.* are distinct from questions about the history of our genetic material. In models that trace genetic material back in time, any given nucleotide position in our genomes can eventually be found in a single individual and on a single chromosome. Thus, being in the *pedigree* of all of humanity

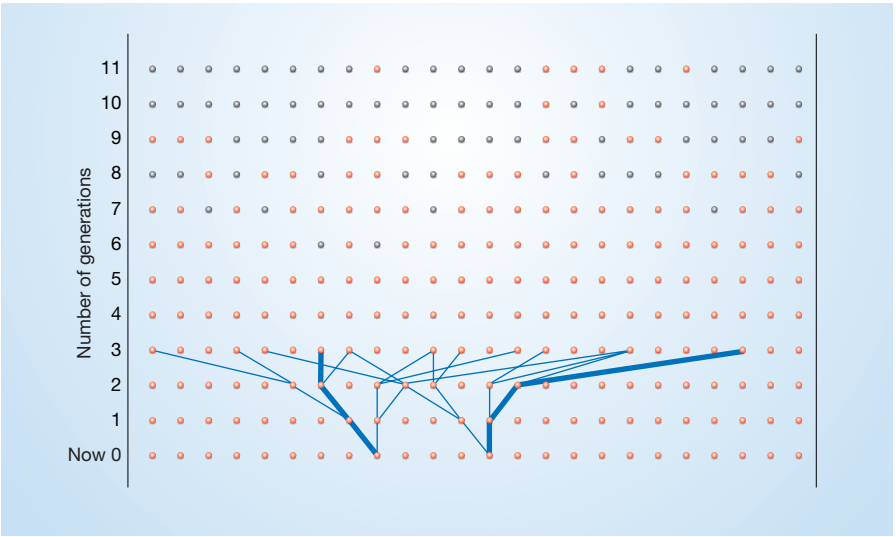


Figure 1 **Searching for our universal common ancestor. The figure shows how the number of ancestors of two people alive today builds up in a manner that is close to exponential. Because the human population has a finite size, however, we do not need to go back many generations before we find an ancestor that is common to both people. The same applies in searching for the ancestor of all living humans (universal ancestors are represented as grey balls). In simplified models, the expected time back to this universal ancestor is $\log_2 n$, where n is the population size. If we were to trace not both parents of each individual, but only one random parent for each (thick lines), we would in effect be tracing the history of gene variants (alleles). In standard models, the number of generations back to the common ancestor of a particular allele will be of the order $2n$, which is much longer ago. If we trace the history of genomes, not genes, recombination would complicate matters; this genetic 'shuffling' ensures that each child does not inherit exactly the same genomic information as its siblings, and means that the genealogical relationship of different genome segments can be different.**

Global change

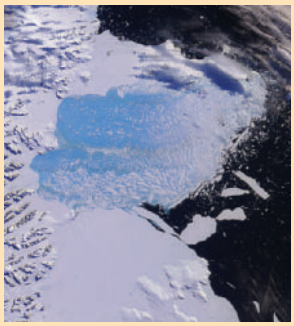
Glacial pace picks up

When a huge chunk of Antarctic ice shelf broke up in 2002, it provided dramatic pictures (see right) for the world's press and a control experiment for researchers. The ice shelf, Larsen B, is a floating extension of the ice of the Antarctic peninsula. The collapse of a substantial part of it — more than 3,000 km^2 — was attributed to increasing temperatures and released shoals of icebergs into the Weddell Sea. But a southerly remnant remained in place, enabling ideas to be tested about how ice shelves might affect glacier flow from the continental interior.

Two groups now report their results of satellite-tracking glacier behaviour in the region (E. Rignot *et al.* and T. A. Scambos *et al. Geophys. Res. Lett.* 10.1029/2004GL020697; 10.1029/2004GL020670). They found that five glaciers flowing into the area formerly buttressed by the ice shelf all accelerated at various times, whereas two farther south, which ran into the remnant ice shelf, did not. Speed of glacier flow is also reflected in their thickness: higher flow rates stretch and thin the ice, in these cases yielding estimated rates of thinning of tens of metres per year.

The main implication is that ice shelves act as a restraint on glacier flow. This conclusion was by no means obvious. Earlier, theoretical studies gave conflicting results; and there are also possible confounding factors, such as water, produced by seasonal melting of surface ice, acting as a lubricant at the glacier base.

A prospect for the future — and a worrying one as far as larger ice shelves and glaciers are concerned — is that a feedback system could kick in, accelerating glacier melting and producing significant rises in sea level. **Tim Lincoln**



NASA/GSFC/LARC/JPL, MISR TEAM

does not imply that an individual makes a significant *genetic* contribution to the present population. In fact, that individual might have contributed nothing. This distinction is also illustrated by 'mitochondrial Eve' — the woman who purportedly lived hundreds of thousands of years ago and carried mitochondrial genes that are ancestral to all present mitochondrial genes. In Fig. 1 you would reach this Eve by tracing only female lineages backwards (rather than both lineages).

Universal common ancestry (in the pedigree sense) and genetic common ancestry thus occur on different timescales. The former is proportional to $\log_2 n$, and if you were to double the current population size, the expected time back to the universal ancestor would move back by only one generation in the simple model. But the time back to the genetic common ancestor is typically proportional to the population size, and so doubling the population size would double the time back to that kind of ancestor. The fact that the number of ancestors in a pedigree increases exponentially, whereas the number of genetic ancestors increases much more slowly, has the consequence that not many generations ago (about six), members of our pedigree existed that did not contribute to us genetically. So being somebody's great-great-great-great grandparent is no guarantee of genetic relatedness. To properly understand genetic ancestry, we need the concept of the ancestral recombination graph[5,6] — a generalization of traditional phylogeny that traces genetic material back in time in the presence of genetic recombination.

The increased ease of obtaining genome-sequence data from individuals, and the number of large-scale projects cataloguing variation in the human population, will increase our ability to test hypotheses about human history. Combining pedigree and genetic ancestry will become more and more important, both for data analysis and in exploring properties of population models[7]. Many interesting questions lie ahead. For instance, how much genetic material (if any) did the universal ancestor pass on to the present population? What about that for a non-universal ancestor from the same time? In the idealized models, how far back would one have to go to find a single couple who are the lone ancestors of everybody? And how much could be known about humanity's pedigree if we knew the genome of everybody? ■

Jotun Hein is in the Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.
e-mail: hein@stats.ox.ac.uk

1. Rohde, D. L. T., Olson, S. & Chang, J. T. *Nature* **431,** 562–566 (2004).
2. Kammerle, K. *J. Appl. Prob.* **27,** 880–885 (1989).
3. Chang, J. *Adv. Appl. Prob.* **31,** 1002–1026 (1999).
4. Derrida, B., Manrubia, S. C. & Zanette, D. H. *J. Theor. Biol.* **203,** 303–315 (2000).
5. Griffiths, R. C. *Theor. Popul. Biol.* **19,** 169–186 (1981).
6. Hudson, R. R. *Theor. Popul. Biol.* **23,** 183–201 (1983).
7. Hein, J. J., Schierup, M. H. & Wiuf, C. H. *Gene Genealogies, Variation and Evolution* (Oxford Univ. Press, 2004).

Cosmology

What is dark energy?

Lawrence M. Krauss

It seems that the rate of expansion of the Universe is accelerating, driven by the so-called dark energy. Is Einstein's cosmological constant behind it? There might be a way to find out.

The nature of the 'dark energy' that is causing the apparent accelerated expansion of the Universe is, without doubt, the biggest mystery in physics and astronomy. Although it was astrophysical observations of the acceleration that led to the discovery of dark energy, there are precious few tests that can be performed to work out what dark energy is — whether it is simply the rebirth of Einstein's cosmological constant, or whether it might stem from something even weirder. All the evidence so far is consistent with the existence of a cosmological constant, which, in modern language, is understood to be the quantum-mechanical energy associated with otherwise empty space. In *Physical Review D*, Kunz *et al.*[1] suggest, however, that by comparing data on a range of astrophysical phenomena, it might be possible to rule out a cosmological constant as the origin of dark energy.

Dark energy is perplexing. Physical theory currently has no explanation of why the energy of empty space should be precisely zero (quantum-mechanical effects combined with relativity in fact predict quite the opposite). But it also gives no explanation of why that energy should not instead be so huge that it would dwarf all of the energy in anything else (making galaxy formation impossible). Yet arguments based on a host of different cosmological observations — even before the direct observation of the accelerated expansion — implied that the energy in empty space could not be more than three to four times greater than the energy contained in the matter and radiation of the Universe. To decide on what physics might be associated with dark energy, we have to rely on experiments and observations. No laboratory experiment we can imagine would be sensitive enough to do the job, so we are left with astrophysical probes. Which is where Kunz *et al.*[1] come in.

The ages for MJG-I, MJG-II and MJG-III are considerably older than previous age estimates of Palaeolithic sites in northern China[1] and indicate that humans might have reached northeast Asia earlier than previously thought. Along with estimated ages for the sites of Gongwangling (1.15 Myr)[14] and Xihoudu (1.27 Myr)[15] in the southern Loess Plateau and for Xiaochangliang (1.36 Myr)[1] and Donggutuo (1.1 Myr)[16] sites in the Nihewan basin, our new results imply an expansion and lengthy flourishing of human groups from northern to north-central China during the early Pleistocene.

The estimated age of 1.66 Myr for the MJG-III artefact layer predates the previous oldest age of unambiguous human presence at 40° N in East Asia by about 0.3 Myr. Our findings, particularly for the MJG-III layer, document the oldest coexistence of stone tools and man-made bone modifications in East Asia, indicating possible continuity with the oldest stone tools and artificial bone modifications reported in eastern Africa[17,18]. Archaeological evidence at MJG indicates the oldest known use of animal tissues, especially marrow processing, by early humans in Asia. The earliest archaeological level in the Nihewan basin is slightly younger than the 1.75 Myr estimated age for early humans at the Dmanisi site at 40° N latitude in western Eurasia[2,3]. Our estimated ages also fall within the 1.66–1.51-Myr range for the earliest known human fossils in southeast Asia[19,20]. The combined evidence suggests that, near the start of the Pleistocene, early human populations spread relatively rapidly across Asia, presumably from an African origin, and reached at least 40° N latitude. Our findings further establish that the earliest populations to reach northeast Asia were able to survive for at least 500 kyr before the mid-Pleistocene onset of high-amplitude climate oscillation[21–23]. □

Received 19 February; accepted 8 July 2004; doi:10.1038/nature02829.

1. Zhu, R. X. *et al.* Earliest presence of humans in northeast Asia. *Nature* **413,** 413–417 (2001).
2. Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: Taxonomy, geological setting, and age. *Science* **288,** 1019–1025 (2000).
3. Vekua, A. *et al.* A new skull of early *Homo* from Dmanisi, Georgia. *Science* **297,** 85–89 (2002).
4. Wei, Q. Banshan Paleolithic site from the lower Pleistocene in the Nihewan Basin in northern China. *Acta Anthropol. Sinica* **13,** 223–238 (1994).
5. HPICR, *Papers on Archaeology in Hebei Province* 30–45 (East, Beijing, 1998).
6. Potts, R. *Early Hominid Activities at Olduvai* (Aldine de Gruyter, New York, 1988).
7. Potts, R., Behrensmeyer, A. K. & Ditchfield, P. Paleolandscape variation and Early Pleistocene hominid activities: Members 1 and 7, Olorgesailie Formation, Kenya. *J. Hum. Evol.* **37,** 747–788 (1999).
8. Tang, Y. J., Li, Y. & Chen, W. Y. Mammalian fossils and the age of Xiaochangliang paleolithic site of Yangyuan, Hebei. *Vertebrata Palasiatica* **33,** 74–83 (1995).
9. Berggren, W. A., *et al.* in *Geochronology, Timescales, and Stratigraphic Correlation* (eds Berggren, W. A., Kent, D. V., Aubry, M. & Hardenbol, J.) 129–212 (SEPM Spec. Publ. 54, Tulsa, Oklahoma, 1995).
10. Wei, Q., *et al.* in *Evidence for Evolution—Essays in Honor of Prof. Chungchien Yong on the Hundredth Anniversary of His Birth* (ed. Tong, Y.) 193–207 (Ocean, Beijing, 1997).
11. Huang, W. P. & Fang, Q. R. *Wushan Hominid Site* 105–109 (Ocean, Beijing, 1991).
12. Qiu, Z. X. Nihewan fauna and Q/N boundary in China. *Quat. Sci.* **20,** 142–154 (2000).
13. Singer, B. S. *et al.* Dating transitionally magnetized lavas of the late Matuyama chron: Toward a new $^{40}Ar/^{39}Ar$ timescale of reversals and events. *J. Geophys. Res.* **104,** 679–693 (1999).
14. An, Z. S. & Ho, C. K. New magnetostratigraphic dates of Lantian *Homo erectus*. *Quat. Res.* **32,** 213–221 (1989).
15. Zhu, R., An, Z., Potts, R. & Hoffman, K. A. Magnetostratigraphic dating of early humans in China. *Earth Sci. Rev.* **61,** 341–359 (2003).
16. Quaternary Research Association of China, Li, H. M. & Wang, J. D. *Quaternary Geology and Environment of China* 33–38 (Ocean, Beijing, 1982).
17. Semaw, S. *et al.* 2.5-million-year-old stone tools from Gona, Ethiopia. *Nature* **385,** 333–336 (1995).
18. de Heinzelin, J. *et al.* Environment and behavior of 2.5-million-year-old Bouri hominids. *Science* **284,** 625–629 (1999).
19. Swisher, C. C. III *et al.* Age of the earliest known hominids in Java, Indonesia. *Science* **263,** 1118–1121 (1994).
20. Larick, R. *et al.* Early Pleistocene $^{40}Ar/^{39}Ar$ ages for Bapang Formation hominins, Central Java, Indonesia. *Proc. Natl Acad. Sci. USA* **98,** 4866–4871 (2001).
21. Potts, R. in *Human Roots: Africa and Asia in the Middle Pleistocene* (eds Barham, L. & Robson-Brown, K.) 5–21 (Western Academic & Specialist Press, Bristol, 2001).
22. Clark, P. U., Alley, R. B. & Pollard, D. Northern Hemisphere ice-sheet influences on global climate change. *Science* **286,** 1104–1111 (1999).
23. Tian, J., Wang, P., Cheng, X. & Li, Q. Astronomically tuned Plio-Pleistocene benthic $\delta^{18}O$ record from South China Sea and Atlantic–Pacific comparison. *Earth Planet. Sci. Lett.* **203,** 1015–1029 (2002).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank R. J. Enkin for providing palaeomagnetic software. This work was supported by the National Natural Science Foundation of China and Chinese Academy of Sciences. R.P. was supported by the US National Science Foundation and the Smithsonian Human Origins Program. K.A.H. also received support from the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.X.Z. (rxzhucn@yahoo.com and rxzhu@mail.igcas.ac.cn) or R.P. (potts.rick@nmnh.si.edu).

Modelling the recent common ancestry of all living humans

Douglas L. T. Rohde[1], Steve Olson[2] & Joseph T. Chang[3]

[1]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*7609 Sebago Road, Bethesda, Maryland 20817, USA*
[3]*Department of Statistics, Yale University, New Haven, Connecticut 06520, USA*

If a common ancestor of all living humans is defined as an individual who is a genealogical ancestor of all present-day people, the most recent common ancestor (MRCA) for a randomly mating population would have lived in the very recent past[1–3]. However, the random mating model ignores essential aspects of population substructure, such as the tendency of individuals to choose mates from the same social group, and the relative isolation of geographically separated groups. Here we show that recent common ancestors also emerge from two models incorporating substantial population substructure. One model, designed for simplicity and theoretical insight, yields explicit mathematical results through a probabilistic analysis. A more elaborate second model, designed to capture historical population dynamics in a more realistic way, is analysed computationally through Monte Carlo simulations. These analyses suggest that the genealogies of all living humans overlap in remarkable ways in the recent past. In particular, the MRCA of all present-day humans lived just a few thousand years ago in these models. Moreover, among all individuals living more than just a few thousand years earlier than the MRCA, each present-day human has exactly the same set of genealogical ancestors.

In investigations of the common ancestors of all living humans, much attention has focused on descent through either exclusively maternal or exclusively paternal lines, as occurs with mitochondrial DNA and most of the Y chromosome[4,5]. But according to the more common genealogical usage of the term 'ancestor', ancestry encompasses all lines of descent through both males and females, so that the ancestors of an individual include all of that person's parents, grandparents, and so on.

For a population of size n, assuming random mating (and so ignoring population substructure), probabilistic analysis[2] has proved that the number of generations back to the MRCA, T_n, has a distribution that is sharply concentrated around $\log_2 n$. We express this using the notation $T_n \sim \log_2 n$, meaning that the quotient $T_n/\log_2 n$ converges in probability to 1 as $n \to \infty$. In contrast, the mean time to the MRCA along exclusively matrilineal or patrilineal lines is approximately n generations[6], and the distribution is not sharply concentrated. For example, in a panmictic population of one million people, the genealogical MRCA would have lived about 20 generations ago, or around the year AD 1400, assuming a generation time of 30 years. The MRCA along

Box 1
Graph-theoretical definitions

The length of a path in a graph, G, is the number of edges in the path. For each pair of nodes i and j in G, the distance $d(i, j)$ is defined to be the length of a shortest path joining i and j. The radius of G is

$$R = \min_{i \in G}\{\max_{k \in G} d(i,k)\}$$

and a node i is called a centre of G if $\max_{k \in G} d(i, k) = R$. Assume $R \geq 1$; the case $R = 0$ (G has one node) was treated previously[2]. For each centre node i, let S_i be a set of minimal size that consists of neighbours of node i and satisfies $\min\{d(j, k): j \in \{i\} \cup S_i\} \leq R - 1$ for all $k \in G$. H_i is defined as the number of nodes in S_i, H is the minimum of H_i over all centres i, and $\Delta = 1 - \frac{1}{H}$. The diameter of G is $D = \max_{i,k \in G} d(i, k)$.

exclusively maternal lines would have lived something like 50,000 times earlier—in the order of one million generations ago.

As genealogical ancestry is traced back beyond the MRCA, a growing percentage of people in earlier generations are revealed to be common ancestors of the present-day population. Tracing further back in time, there was a threshold, let us say U_n generations ago, before which ancestry of the present-day population was an all or nothing affair. That is, each individual living at least U_n generations ago was either a common ancestor of all of today's humans or an ancestor of no human alive today. Thus, among all individuals living at least U_n generations ago, each present-day human has exactly the same set of ancestors. We refer to this point in time as the identical ancestors (IA) point. As with the MRCA point, the IA point is also quite recent in a randomly mating population: $U_n \sim 1.77 \log_2 n$ generations ago[2].

The major problem in applying these results to human populations is that mating is not random in the real world. Mating patterns are structured by geography, proximity, culture, language and social class. Nevertheless, even in populations with considerable internal structure, the time to the MRCA can be remarkably brief. To demonstrate this in a tractable mathematical model, consider a population of size n divided into randomly mating subpopulations that are linked by occasional migrants. The population is represented by a graph, G, with a node for each subpopulation. Edges indicate pairs of nodes that exchange a small number (for example, one pair) of migrants per generation. Let R denote the radius of G, and let Δ be a quantity ranging between 0 and 1 that depends on the structure of G (see Box 1). A probabilistic analysis (see Supplementary Information) shows that as $n \to \infty$, $T_n \sim (R + \Delta)\log_2 n$. Furthermore, if we let D denote the diameter of the graph, then the number of generations, U_n, since the IA point satisfies $U_n \sim (D + 1.77)\log_2 n$.

Computer simulations accord with these theoretical predictions. Tables 1 and 2 give distributions of T_n and U_n for small populations of varying sizes in graphs with one node, three connected nodes, five fully connected nodes and for a ten-node graph loosely based on world geography as shown in Fig. 1. In these simulations, neighbouring subpopulations exchange one pair of migrants per generation. Each mean is calculated from 100 model runs. Although guaranteed to be accurate only for sufficiently large n, the theoretical predictions describe the simulations quite well even for models with just a few thousand individuals. Whenever n is doubled, T_n is expected to increase by $R + \Delta$, and U_n is expected to increase by $D + 1.77$. These predicted increases, which are listed in the last columns of Tables 1 and 2, agree closely with the simulation results.

To hazard a rough first guess about human recent common ancestors, we could extrapolate the results for the graph of Fig. 1 to a growing population with a final size of 250 million. When applying this model to a growing population, the fixed population size that provides the best approximation is the size at the time that the MRCA lived. We take this effective population size to be 250 million, which is approximately the global population in the year AD 1. Starting from $n = 16{,}000$, a population of 250 million is reached by doubling 14 times. Approximating the increases in T_n and U_n beyond the values seen in Tables 1 and 2 by their theoretical predictions for each doubling of n, we arrive at $T_n \approx 34 + 14 \times 3 = 76$ generations (about 2,300 years) and $U_n \approx 74 + 14 \times 6.77 = 169$ generations (about 5,000 years). These estimates would suggest, with the exchange of just one pair of migrants per generation between large panmictic populations of realistic size, that the MRCA appears in about the year 300 BC, and all modern individuals have identical ancestors by about 3,000 BC. Such estimates are extremely tentative, and the model contains several obvious sources of error, as it was motivated more by considerations of theoretical insight and tractability than by realism. Its main message is that substantial forms of population subdivision can still be compatible with very recent common ancestors.

The dynamics of human subpopulations are much more complex than those in the simple graph model discussed above. Although these complexities make theoretical analysis difficult, a computer model incorporating more complicated forms of population substructure and migration allows the demographic history of human populations to be simulated. The Supplementary Information contains more details on the model and computations; here we briefly outline some of the main points.

This model is based on a simplified projection of the world's

Table 1 **Simulations of T_n**

Graph	$n = 1{,}000$	$n = 2{,}000$	$n = 4{,}000$	$n = 8{,}000$	$n = 16{,}000$	$R + \Delta$
One node	10.8 (0.4)	11.8 (0.4)	12.8 (0.4)	13.9 (0.3)	14.8 (0.4)	1.00
Three fully connected nodes	14.0 (0.7)	15.6 (0.7)	17.1 (0.9)	18.9 (0.8)	20.3 (1.0)	1.50
Five fully connected nodes	14.0 (0.5)	15.8 (0.5)	17.8 (0.5)	19.6 (0.5)	21.5 (0.6)	1.75
Ten-node graph shown in Fig. 1	21.1 (1.3)	24.3 (1.5)	27.6 (1.5)	30.5 (1.5)	33.8 (1.7)	3.00

Means (standard deviations in parentheses) of T_n (the number of generations back to the MRCA) for graph-structured populations exchanging a single pair of migrants per edge per generation. The last column shows $R + \Delta$, the expected asymptotic increase in T_n per doubling of n.

Table 2 **Simulations of U_n**

Graph	$n = 1{,}000$	$n = 2{,}000$	$n = 4{,}000$	$n = 8{,}000$	$n = 16{,}000$	$D + 1.77$
One node	20.8 (1.6)	22.6 (1.5)	24.6 (1.5)	26.5 (1.6)	28.3 (1.4)	1.77
Three fully connected nodes	27.4 (1.5)	30.3 (1.4)	33.4 (1.5)	36.2 (1.7)	38.9 (1.5)	2.77
Five fully connected nodes	25.9 (1.3)	28.9 (1.4)	32.1 (1.7)	35.3 (1.5)	37.9 (1.4)	2.77
Ten-node graph shown in Fig. 1	46.3 (2.7)	53.0 (2.7)	59.8 (2.7)	66.8 (2.9)	73.6 (2.7)	6.77

Means (standard deviations in parentheses) of U_n (the number of generations back to the IA point) for graph-structured populations exchanging a single pair of migrants per edge per generation. The last column shows $D + 1.77$, the expected asymptotic increase in U_n per doubling of n.

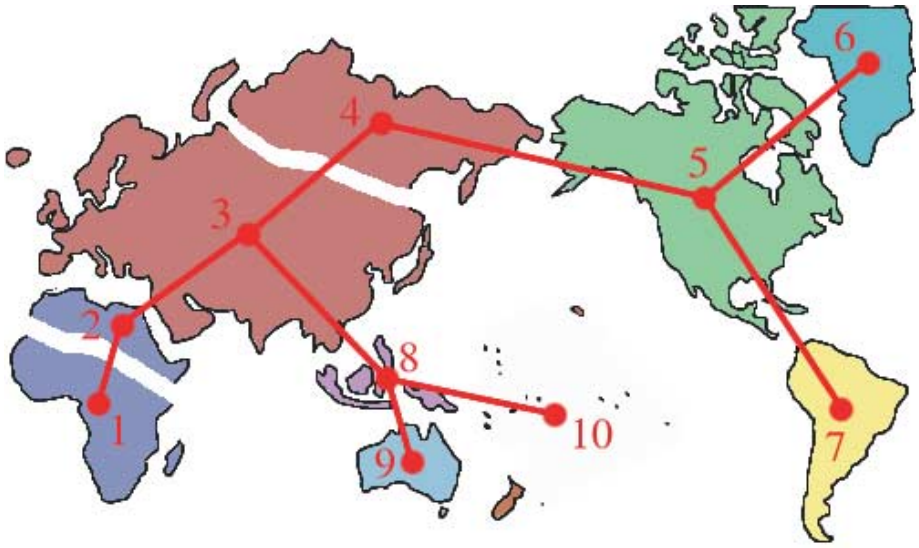


Figure 1 World map viewed as a ten-node graph. This graph has radius 3 and diameter 5.

actual inhabited land masses and has three levels of substructure: continents, 'countries' and 'towns.' Figure 2 depicts the model's geography and migration routes used before AD 1500, with the countries shown as squares and the number of towns per country differing from continent to continent. Towns and countries represent both the local geographical areas and the relevant social and ethnic groups from which most people find mates.

The model uses a simplified migration system in which each person has a single opportunity to migrate from his or her town of birth. The probabilities of leaving a town or a country are set at various levels to reflect different migration patterns. Migrants who move between towns can travel to any other town within the country. A migrant who leaves a country for another country within the same continent chooses the destination with a probability that diminishes as the inverse square of the geographical distance.

Each continent has a number of port countries from which migrants can travel to another continent. A fixed, large percentage (for example, 95% in some simulations) of the migrants through a port come from the country in which the port is located, with the remainder drawn from other countries in the continent in proportion to their inverse squared distance. The value next to a port in Fig. 2 is its migration rate, in people per generation, and the date in parentheses indicates when the port opens, if it is more recent than the start of the simulation in 20,000 BC. When a port opens, there is usually a single generation of migration at a higher rate than the steady-state rate shown in the figure. After the year AD 1500, additional large ports, which are not shown, begin to open to simulate colonization of the Americas, Australia and elsewhere. Immediately before this, the native population of the Americas is markedly reduced to simulate the effects of European-introduced diseases[7].

Generations overlap in this model and we explicitly simulated the lifespan and the times at which mating and reproduction events occur for each individual[8,9], as described in more detail in Supplementary Information. The birth rate of each continent or island was individually adjusted so that the populations match historical

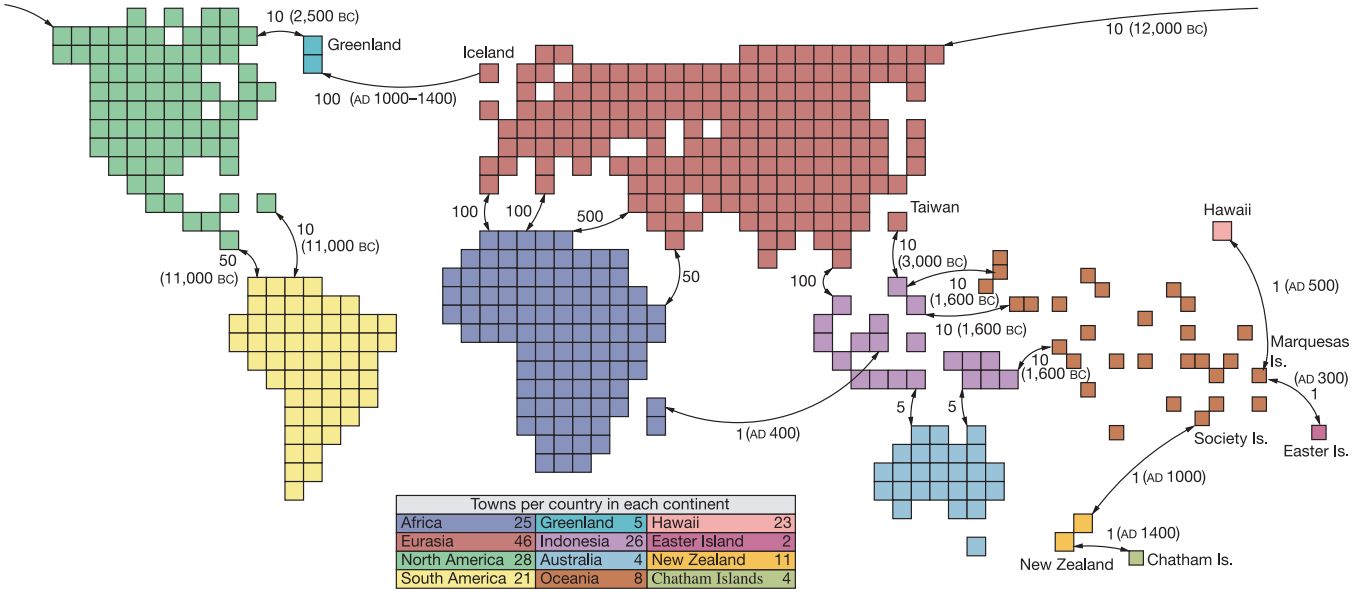


Figure 2 Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.

estimates, and growth rates were higher in under-populated areas. Full-sized populations were used until the world population reached 50 million in 1,000 BC. Subsequently, birth rates were reduced to achieve a worldwide level of 55 million, carried out in such a way that sparsely populated areas were less affected. This limit was a computational necessity, but simulations show that population growth has little effect, especially if it occurs after the MRCA has died.

With 5% of individuals migrating out of their home town, 0.05% migrating out of their home country, and 95% of port users born in the country from which the port emanates, the simulations produce a mean MRCA date of 1,415 BC and a mean IA date of 5,353 BC. Interestingly, the MRCAs are nearly always found in eastern Asia. This is due to the proximity of this region to both Eurasia and either the remote Pacific islands or the Americas, allowing the MRCA's descendants to reach a few major world regions in a relatively short time.

Arguably, this simulation is far too conservative, especially given its prediction that, even in densely populated Eurasia, only 55.3 people will leave each country per generation in AD 1500. If the migration rate among towns is increased to 20%, the local port users are reduced to 80%, and the migration rates between countries and continents are scaled up by factors of 5 and 10, respectively, the mean MRCA date is as recent as AD 55 and the mean IA date is 2,158 BC. The predictions of the simple ten-node graph model sketched earlier fall somewhere between these dates and those of the more conservative computational model.

The model also can be used to calculate the percentage of ancestry that current individuals receive from different parts of the world. In generations sufficiently far removed from the present, some ancestors appear much more often than do others on any current individual's family tree, and can therefore be expected to contribute proportionately more to his or her genetic inheritance[1,10,11]. For example, a present-day Norwegian generally owes the majority of his or her ancestry to people living in northern Europe at the IA point, and a very small portion to people living throughout the rest of the world. Furthermore, because DNA is inherited in relatively large segments from ancestors, an individual will receive little or no actual genetic inheritance from the vast majority of the ancestors living at the IA point[12].

Several factors could cause the time to the true MRCA or IA point to depart from the predictions of our model. If a group of humans were completely isolated, then no mixing could occur between that group and others, and the MRCA would have to have lived before the start of the isolation. A more recent MRCA would not arise until the groups were once again well integrated. In the case of Tasmania, which may have been completely isolated from mainland Australia between the flooding of the Bass Strait, 9,000–12,000 years ago, and the European colonization of the island, starting in 1803 (ref. 13), the IA date for all living humans must fall before the start of isolation. However, the MRCA date would be unaffected, because today there are no remaining native Tasmanians without some European or mainland Australian ancestry.

No large group is known to have maintained complete reproductive isolation for extended periods. The populations on either side of the Bering Strait appear to have exchanged mates throughout the period documented in the archaeological record[14]. Religious isolates such as the Samaritans occasionally have absorbed migrants from outside the group[15]. Even populations on isolated Pacific islands have experienced occasional infusions of newcomers[16]. Even if rates of migration between some adjoining populations are very low, the time to the MRCA tends not to change substantially. For example, with a migration rate across the Bering Strait of just one person in each direction every ten generations, rather than the ten per generation in the more conservative simulation described earlier, T_n only increases from 3,415 years to 3,668 years.

Conversely, other factors could reduce the time to the MRCA from that predicted by the model. Examples of such factors include the existence of more diverse intercontinental migration routes, the large-scale movement and mixing of populations documented in the historical record[17], marked individual differences in fertility[18], and the population increase of the past two millennia, which would result in more migrants.

Actual migration rates among populations are very poorly known and undoubtedly have varied considerably in different times and places. Studies of hunter-gatherer groups and subsistence agricultural communities have found that anywhere from 1% (ref. 19) to as much as 30% (ref. 20) of mates are from outside the group. The tendency of most human groups to marry out with surrounding groups, at least to a limited extent, links networks of ancestry within specific regions (see http://www.compapp.dcu.ie/~humphrys/FamTree/Royal/Famous.descents.html).

Given the remaining uncertainties about migration rates and real-world mating patterns, the date of the MRCA for everyone living today cannot be identified with great precision. Nevertheless, our results suggest that the most recent common ancestor for the world's current population lived in the relatively recent past—perhaps within the last few thousand years. And a few thousand years before that, although we have received genetic material in markedly different proportions from the people alive at the time, the ancestors of everyone on the Earth today were exactly the same.

Further work is needed to determine the effect of this common ancestry on patterns of genetic variation in structured populations[21–24]. But to the extent that ancestry is considered in genealogical rather than genetic terms, our findings suggest a remarkable proposition: no matter the languages we speak or the colour of our skin, we share ancestors who planted rice on the banks of the Yangtze, who first domesticated horses on the steppes of the Ukraine, who hunted giant sloths in the forests of North and South America, and who laboured to build the Great Pyramid of Khufu. □

Received 30 December 2003; accepted 14 July 2004; doi:10.1038/nature02842.

1. Wachter, K. W. in *Genealogical Demography* (eds Dyke, B. & Morrill, W. T.) 85–93 (Academic, New York, 1980).
2. Chang, J. T. Recent common ancestors of all present-day individuals. *Adv. Appl. Probab.* **31,** 1002–1026, 1027–1038 (1999).
3. Derrida, B., Manrubia, S. C. & Zanette, D. H. On the genealogy of a population of biparental individuals. *J. Theor. Biol.* **203,** 303–315 (2000).
4. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408,** 708–713 (2000).
5. Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl Acad. Sci. USA* **97,** 7360–7365 (2000).
6. Hudson, R. R. in *Oxford Surveys of Evolutionary Biology* (eds Harvey, P. H. & Partridge, L.) 1–44 (Oxford Univ. Press, New York, 1990).
7. Stannard, D. E. *American Holocaust: Columbus and the Conquest of the New World* (Oxford Univ. Press, New York, 1992).
8. US National Office of Vital Statistics, *Death Rates by Age, Race, and Sex, United States, 1900–1953, Vital Statistics—Special Reports* **Vol. 43** (US Government Printing Office, Washington DC, 1956).
9. Pletcher, S. D. Model fitting and hypothesis testing for age-specific mortality data. *J. Evol. Biol.* **12,** 430–439 (1999).
10. Ohno, S. The Malthusian parameter of ascents: What prevents the exponential increase of one's ancestors? *Proc. Natl Acad. Sci. USA* **93,** 15276–15278 (1996).
11. Derrida, B., Manrubia, S. C. & Zanette, D. H. Distribution of repetitions of ancestors in genealogical trees. *Physica A* **281,** 1–16 (2000).
12. Wiuf, C. & Hein, J. On the number of ancestors to a DNA sequence. *Genetics* **147,** 1459–1468 (1997).
13. Jones, R. Tasmanian archaeology: Establishing the sequences. *Ann. Rev. Anthropol.* **24,** 423–446 (1995).
14. Fitzhugh, W. W. & Chausonnet, V. (eds) *Crossroads of Continents: Cultures of Siberia and Alaska* (Smithsonian Institution Press, Washington DC, 1988).
15. Bonné-Tamir, B. *et al.* Maternal and paternal lineages of the Samaritan isolate: Mutation rates and time to most recent common male ancestor. *Ann. Hum. Genet.* **67,** 153–164 (2003).
16. Morton, N. E., Harris, D. E., Yee, S. & Lew, R. Pingelap and Mokil atolls: Migration. *Am. J. Hum. Genet.* **23,** 339–349 (1971).
17. Hoerder, D. *Cultures in Contact: World Migrations in the Second Millennium* (Duke Univ. Press, Durham, North Carolina, 2002).
18. Zerjal, T. *et al.* The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72,** 717–721 (2003).
19. Weiss, K. M. & Maruyama, T. Archeology, population genetics and studies of human racial ancestry. *Am. J. Phys. Anthropol.* **44,** 31–50 (1976).
20. Ward, R. H. & Neel, J. V. Gene frequencies and microdifferentiation among the Makiritare indians. IV. A comparison of a genetic network with ethnohistory and migration matrices; a new index of genetic isolation. *Am. J. Hum. Genet.* **22,** 538–561 (1970).

21. Jorde, L. B. in *Current Developments in Anthropological Genetics* (eds Mielke, J. H. & Crawford, M. H.) 135–208 (Plenum, New York, 1980).
22. Notohara, M. The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29,** 59–75 (1990).
23. Wilkinson-Herbots, H. M. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37,** 535–585 (1998).
24. Hey, J. & Machado, C. A. The study of structured populations—new hope for a difficult and divided science. *Nature Rev. Genet.* **4,** 535–543 (2003).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements The research of D.L.T.R. was supported by the National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.L.T.R. (dr@tedlab.mit.edu).

Phenotypic consequences of 1,000 generations of selection at elevated CO_2 in a green alga

Sinéad Collins & Graham Bell

Biology Department, McGill University, Montreal, Quebec H3A 1B1, Canada

Estimates of the effect of increasing atmospheric CO_2 concentrations on future global plant production rely on the physiological response of individual plants or plant communities when exposed to high CO_2 (refs 1–6). Plant populations may adapt to the changing atmosphere, however, such that the evolved plant communities of the next century are likely to be genetically different from contemporary communities[7–12]. The properties of these future communities are unknown, introducing a bias of unknown sign and magnitude into projections of global carbon pool dynamics. Here we report a long-term selection experiment to investigate the phenotypic consequences of selection for growth at elevated CO_2 concentrations. After about 1,000 generations, selection lines of the unicellular green alga *Chlamydomonas* failed to evolve specific adaptation to a CO_2 concentration of 1,050 parts per million. Some lines, however, evolved a syndrome involving high rates of photosynthesis and respiration, combined with higher chlorophyll content and reduced cell size. These lines also grew poorly at ambient concentrations of CO_2. We tentatively attribute this outcome to the accumulation of conditionally neutral mutations in genes affecting the carbon concentration mechanism.

Plant growth depends on CO_2 concentration[1,2], which is expected to rise from current levels of about 400 parts per million (p.p.m.) to between 700 and 1,000 p.p.m. during the next century[3]. In response, global plant productivity in forests[4], grasslands[5], agroecosystems[6] and other ecosystems is expected to increase. Projections of future net primary productivity are complicated by synchronous changes in temperature and other factors, but most models predict increases in the land–atmosphere and ocean–atmosphere fluxes from current values of >-2 petagrams (Pg) C per year to about -5 Pg C per year[3]. This process is likely to be complicated by shifts in the species composition of plant communities[7], and more fundamentally by evolutionary changes within plant populations. In the very long term, this may involve the extinction of some groups and the radiation of others[8], but within a few hundred generations most plant populations may adapt to the increased supply of inorganic carbon. Selection experiments with plants have demonstrated a variety of responses[9–12], but have been limited to fewer than ten generations. The long-term response to selection and the properties of populations adapted to elevated CO_2 remain unknown, and constitute an important limit on our ability to predict future plant productivity.

We used a microbial model system in which large population size and short generation time make it possible to evaluate evolutionary change caused by the spread of novel mutations over hundreds of generations. *Chlamydomonas reinhardtii* is a unicellular green alga that has been extensively used to study the physiology and genetics of photosynthesis[13]. It possesses a carbon-concentrating mechanism (CCM), which increases the concentration of CO_2 near the active site of ribulose 1,5-bisphosphate carboxylase–oxygenase (Rubisco), in common with most other eukaryotic microalgae that have been studied[14]. We set up ten isogenic selection lines from each of two ancestral genotypes, half being grown at ambient CO_2 (ambient lines) and half at a concentration that increased from ambient to 1,050 p.p.m. over about 600 generations and was then maintained at this level for a further 400 generations (high lines). At least 10^5 cells per line were transferred for 125 transfers in a buffered, nutrient-rich medium. The history of these lines thus emulates the conditions that photosynthetic organisms are likely to experience during the next century or so, with respect to CO_2 levels alone.

The physiological effect of elevated CO_2 concentration is expected to be an increase in photosynthesis, causing an increase in growth. Net photosynthesis in the ambient lines increased by about 30% when they were grown at high CO_2 (Fig. 1a). The ambient lines diverged through time so that by the end of the experiment they varied significantly in the rate of photosynthesis (one-way analysis of variance (ANOVA): $F_{9,18} = 9.0$, $P < 0.001$) when grown at ambient CO_2 concentrations. The high lines had normal rates of photosynthesis at ambient CO_2, which increased by more than 50% as an average over all lines at high CO_2. However, this effect was very inconsistent: one group of high lines had low rates whereas a second group had very high rates of photosynthesis at high CO_2 concentration (Fig. 1a). This distinction was not related to the identity of the ancestor, and represented significantly more divergence in photosynthetic rates than was seen in the ambient lines ($F_{1,16} = 10.5$, $P = 0.005$).

The growth rate of cultures grown at elevated CO_2 was correlated with their photosynthetic rate among the ambient lines, but not among the high lines (Fig. 1b). The physiological effect of CO_2 on photosynthesis was reflected by growth in pure culture, where the maximal rate of increase (Fig. 1c) and the limiting density (Fig. 1d) of both the ambient and the high lines are enhanced substantially by high CO_2. However, there was no indication of a parallel evolutionary response: by the end of the selection experiment, the high lines had not become specifically adapted to growth at high CO_2; their growth at high CO_2 being no greater than, and perhaps even less than, the growth of the ambient lines. There was nevertheless an indirect response: the growth of some high lines was markedly impaired at ambient CO_2 concentrations where two of the lines could scarcely be propagated. This result was supported by the outcome of competition assays in which the selection lines were mixed with standard genetically marked strains and the change in frequency during growth in culture recorded (Table 1). The high lines had considerably lower competitive ability at ambient CO_2, where three of them (including the two with strongly reduced growth in pure culture) were such weak competitors that they were consistently eliminated by the tester strains within 10–15 generations. They were, however, no more successful than the ambient lines at high CO_2. In short, 1,000 generations of selection at high CO_2 concentrations had caused no increase in growth at high CO_2, whereas growth at ambient CO_2 was often considerably reduced.

Photosynthesis is linearly related to respiration in the dark among lines at ambient CO_2; this relationship is the same for ambient and high lines (Fig. 2a). It has been shown in *Chlamydomonas* that post-illumination rates of O_2 consumption provide a

Modeling the recent common ancestry of all living humans
Supplementary Methods A:
Further Explanation and Derivations of Mathematical Results

Douglas L. T. Rohde, Steve Olson, Joseph T. Chang

The purpose of this note is to explain the derivations of the mathematical results in the paper. Much of the reasoning follows similar lines and uses similar techniques to those presented in full detail in the earlier paper [1]. Here we will draw freely upon results and arguments from that paper, and some of those arguments will be sketched in a less rigorous way here.

One of the principal conceptual messages of the results discussed here about the simple graph models is that a seemingly rather severe form of population subdivision can still be compatible with recent common ancestors. In these simple models, the population is divided up into subpopulations that exchange migrants very infrequently. We assume some small fixed number of migrant individuals per generation; for example, that number could be just one migrant per generation, of even smaller.

The model begins with a connected graph $\mathbb{G}$ consisting of G nodes, which we will refer to here as "islands," with a constant population size of n/G on each island. This is a discrete-time model with time measured in generations. We could choose to call an arbitrary generation $t = 0$, and then t increases by 1 whenever time proceeds forward by one generation. Each individual lives on a particular island (the individual's "home island") in a particular generation. We will use the notation $I(t,i,m)$ to refer to individual number m on island i in generation t.

Each individual has two parents in the previous generation. The two parents are chosen independently, both according to the following probabilistic process. There is a "migration probability" which we will denote by μ_n. With probability μ_n, an individual's parent is chosen from a different island from the home island of the individual, in which case the parent's island is equally likely to be any of the neighbors of the individual's home island (where neighbors are determined by the edges in the graph $\mathbb{G}$). With probability $1-\mu_n$ an individual's parent is chosen from the same island as the individual. In either case – whether chosen from the individual's home island or a neighboring island – the parent is taken to be uniformly distributed, that is, equally likely to be any of the n/G individuals on the chosen island in the previous generation. In other words, to choose a parent of individual $I(t+1,i,m)$, we would first choose a random index m' uniformly distributed over the set $\{1,2,\ldots,\frac{n}{G}\}$ and then with probability $1-\mu_n$ take the parent to be $I(t,i,m')$, and with probability μ_n, choose an island j randomly from among the neighbors of i and take the parent to be $I(t,j,m')$. For each individual, two parents are chosen in this way.

We call an individual a *migrant* if at least one of that individual's parents is from an island other than the individual's home island. (It might be more natural to call the parent the migrant, but we will retain this terminology here as it has been convenient.)

Here we take the migration probability μ_n to be of order $1/n$ by letting c be a constant and taking $\mu_n = c/n$. With the idea of modeling a strongly subdivided population, we are letting the migration rate have a very small order of magnitude. For example, we could choose c so that the expected number of migrants in the whole population is just 1 per generation. Or we could choose c one tenth as large, which

would model a situation in which in a span of ten generations just one migrant is expected in the whole population.

As defined in the paper, a *common ancestor* of a given set of individuals is an individual who is an ancestor of everyone in the set. For example we will speak of a common ancestor of everyone on a particular island at a particular time, or the whole population at a particular time. We use "CA" as an abbreviation for common ancestor. T_n is the number of generations back to the most recent common ancestor (MRCA) of the population. U_n is the number of generations back to the "IA point," the most recent generation in which all current individuals have identical ancestors.

We will use standard notation related to orders of magnitude and asymptotic behavior as $n \to \infty$:

- $f(n) = o\big(g(n)\big)$ means $\dfrac{f(n)}{g(n)} \to 0$ as $n \to \infty$,
- $f(n) = O\big(g(n)\big)$ means $\dfrac{f(n)}{g(n)}$ is bounded,
- $f(n) \asymp g(n)$ means that both $f(n) = O\big(g(n)\big)$ and $g(n) = O\big(f(n)\big)$ hold, and
- $f(n) \sim g(n)$ means $\dfrac{f(n)}{g(n)} \to 1$ as $n \to \infty$.

For notational convenience we will omit writing the obvious "greatest-integer" type functions that are needed in order to round real numbers into integers, such as in the phrase "in generation $t = (1-\varepsilon)\big(D + 1 + \zeta\big)\log_2 n$."

Statements of Results

We are given a connected graph $\mathbb{G}$ and define the distance $d(i,j)$ to be the number of edges in a shortest path joining i and j. This definition is extended to sets of nodes by considering shortest paths joining some node of one set to some node of the other set. That is, for sets of nodes A and B we define $d(A,B) = \min\{d(i,j) : i \in A, j \in B\}$, and as a special case, for a set of nodes A, we define $d(A,j) = \min\big\{d(i,j) : i \in A\big\}$.

The *radius* of $\mathbb{G}$ is $R = \min\limits_{i\in\mathbb{G}} \{\max\limits_{k\in\mathbb{G}} d(i,k)\}$, and a node i is called a *center* of $\mathbb{G}$ if $\max\limits_{k\in\mathbb{G}} d(i,k) = R$. Let $C(\mathbb{G})$ denote the collection of all centers of $\mathbb{G}$. The *diameter* of $\mathbb{G}$ is $D = \max\limits_{i\in\mathbb{G}} \{\max\limits_{k\in\mathbb{G}} d(i,k)\}$.

We assume throughout that $R > 0$; the case $R = 0$ (that is, $\mathbb{G}$ has just one node) was treated in [1].

For $i \in C(\mathbb{G})$, let S_i be a set of minimal size that consists of neighbors of node i and satisfies $\max\limits_{j\in\mathbb{G}} d(\{i\} \cup S_i, j) = R - 1$. Define H_i to be the number of nodes in S_i, $\Delta_i = (H_i - 1)/H_i$, and $\Delta = \min\limits_{i\in C(\mathbb{G})} \Delta_i$. Note $0 \le \Delta < 1$.

The results are asymptotic, with the number of islands and the graph fixed, and n tending to infinity. We use "$\lg$" to denote the base-2 logarithm.

Theorem 1: $\frac{T_n}{(R+\Delta)\lg n}$ converges in probability to 1 as $n \to \infty$.

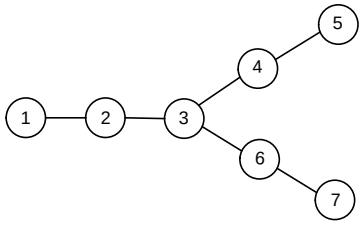
Theorem 2: Let $\zeta \approx 0.7698$ be as defined in Theorem 2 of [1]. $\frac{U_n}{(D+1+\zeta)\lg n}$ converges in probability to 1 as $n \to \infty$.

To attempt a quick and very rough explanation of the theorems in a nutshell: the main idea of Theorem 1 is that every $\lg n$ generations, the set of nodes occupied by descendants of any given individual expands to include all of its neighbors. One can imagine this as a set that expands like clockwork, with a clock that ticks once every $\lg n$ generations. At time 0 the set starts out including only one node. With each tick of the clock, the set expands to encompass all neighbors of nodes currently in the set. Applying this idea to the individuals on a center node of a graph of radius *R* gives the essence of the result: in roughly *R* ticks of the clock, or $R \lg n$ generations, this set of nodes expands to include the whole graph. Similar comments apply to give a rough explanation of Theorem 2 – at least why the diameter appears. The longest path in the graph is of length D. So after the clock ticks about D times, everyone who is destined to become a CA of the full population should have done so.

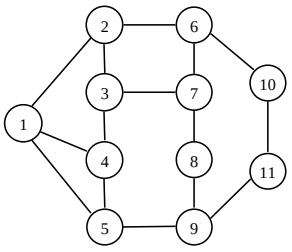
A few more remarks about the results are in order here. First, the reason that the process behaves much like a regularly ticking clock is related to the fact that the distribution of T_n is concentrated around $\lg n$, with little variability. Second, as in the case where $\mathbb{G}$ has one node [1], at the IA point about 80 percent of the population in the structured model consists of common ancestors of everyone in the population today and the lineages of the remaining 20 percent have gone extinct. The derivation will also show that as *n* increases MRCAs are increasingly likely to be found in center nodes of the graph -- in particular, in nodes i that minimize H_i.

Examples

In the graph shown at right, $R = 2$, and the center node 3 has $H = 3$ neighbors 2, 4, and 6, such that the set $S_3 = \{3,2,4,6\}$ lies within $R - 1 = 1$ of each node, that is, $\max_{j \in G} d(S_3, j) = R - 1 = 1$. So here, $\Delta = (H-1)/H = 2/3$, and $T_n \sim (2 + \frac{2}{3})\lg n$. Since the diameter of this graph is 4, we have $U_n \sim (5+\zeta)\lg n$.



For another example, consider the 11-node graph at right. For this graph, the radius is $R = 3$ and the centers are nodes 1, 2, 5, 6, 7, 8, and 9. It turns out that each of the centers i has $H_i = 2$. For example, node 1 has $S_1 = \{2,5\}$; in fact every node is within $R - 1 = 2$ of the set $\{2,5\}$. Node 6 has $S_6 = \{2,7\}$; every node is within 2 of the set $\{2,6,7\}$. For this graph, $H = 2$, $\Delta = \frac{1}{2}$, and $T_n \sim (3 + \frac{1}{2})\lg n$. The diameter of this graph is 4; for example, $d(3,11) = 4$. Therefore, $U_n \sim (5+\zeta)\lg n$.



In a complete graph having $G > 1$ nodes, each pair of nodes is joined by an edge. Such a graph has radius $R = 1$ and diameter $D = 1$. Also, for each i the set S_i is simply $\{1,\ldots,G\} - \{i\}$, so that $H_i = G - 1$, $\Delta_i = (G-2)/(G-1)$, and $\Delta = (G-2)/(G-1)$. Thus, $T_n \sim (1 + (G-2)/(G-1))\lg n$; for example, for $G = 2$ we have $T_n \sim \lg n$, for $G = 3$ we have $T_n \sim (1 + \frac{1}{2})\lg n$, and for $G = 5$ we have $T_n \sim (1 + \frac{3}{4})\lg n$. The result for U_n is $U_n \sim (2 + \zeta)\lg n$.

More on Simulations and Approximations

The above theorems give the main term of asymptotic results of the form $T_n \sim (R + \Delta)\lg(n)$ as $n \to \infty$, for example, which do not distinguish among various possible explicit forms for the lower order effects. For example, the statements $T_n \sim (R + \Delta)\lg(n) + 2.7$ and $T_n \sim (R + \Delta)\lg(n) - 8.3$, which differ by having different additive offsets, would both be consistent with the theorem. However, we can find some simple formulas that agree with the asymptotic results in the theorems, make intuitive sense, and provide rather good approximations to the simulation results. Defining G as above to be the number of nodes in the graph, if we use the approximations $T_n \sim (R + \Delta)\lg(n/G)$ and $U_n \sim (D + 1.77)\lg(n/G)$, we get the following predicted values for the simulation results in Table 1 and Table 2.

Predictions for Table 1 using the formula $T_n \sim (R + \Delta)\lg(n/G)$:

	n=1000	*2000*	*4000*	*8000*	*16000*
One node	10.0	11.0	12.0	13.0	14.0
Three fully-connected nodes	13.4	14.9	16.4	17.9	19.4
Five fully-connected nodes	14.7	16.4	18.2	19.9	21.7
Ten-node graph in Figure 1	19.9	22.9	25.9	28.9	31.9

Predictions for Table 2 using the formula $U_n \sim (D + 1.77)\lg(n/G)$:

	n=1000	*2000*	*4000*	*8000*	*16000*
One node	17.6	19.4	21.2	22.9	24.7
Three fully-connected nodes	24.8	27.6	30.4	33.1	35.9
Five fully-connected nodes	23.2	26.0	28.8	31.5	34.3
Ten-node graph in Figure 1	45.0	51.7	58.5	65.3	72.1

Comparing these predictions to the actual simulation results in Table 1 and Table 2 of the paper, we see that the agreement is quite good, with the predictions capturing the main features of the simulation results.

More Terminology, Head Starts and the Idea of Delta

To introduce convenient terminology for some ideas that were introduced in [1] and are helpful here, let us say that a given individual $I(t_1, i_1, m_1)$ is *established* on a given island i_2 in a given generation t_2 if the number of descendants of $I(t_1, i_1, m_1)$ on island i_2 in generation t_2 is greater than $\lg^2(n)$. We say $I(t_1, i_1, m_1)$ is established in generation t_2 (without reference to a particular island i_2) if $I(t_1, i_1, m_1)$ is established on some island in generation t_2. We say an individual is *in jeopardy* in a given generation if that individual is not established on any island in that generation and is also not extinct in that generation.

The idea is that an individual who is *not* in jeopardy at a given time is either already extinct, which means he has no surviving descendants at that time, or established, which means that his number of descendants is large enough to assure that it is very unlikely that the individual will become extinct in the future. That is, individuals who are not in jeopardy are either extinct or are very likely to become CA's. We say an individual is *destined* to become established, a CA, and so on, if these events will occur for that individual in the future.

Let us say that a "*head start*" has been completed for a center island i as soon as some individual on island i has migrant descendants who have become established on each of the islands in some set S_i consisting of neighbors of node i and satisfying $\max_{j\in\mathbb{G}} d(\{i\}\cup S_i, j) = R-1$. We will express the time at which some individual from island i becomes a CA of the whole population as the sum of the time required to complete a head start for i and the additional time required to complete the process of becoming a CA after completing the head start.

Proposition H: For a center island i, the time required to complete a head start is $\sim \Delta_i \lg n$. That is, for $0<\alpha<1$, as $n\to\infty$, the probability that at least one individual from island i has completed a head start by time $\alpha \lg n$ converges to 0 if $\alpha < \Delta_i$ and converges to 1 if $\alpha > \Delta_i$.

To verify the proposition, consider the individuals on island i at time 0. Those individuals fall into a number of categories, defined in terms of the number of descendants those individuals have at time $\alpha \lg n$. As shown in [1], a fraction of nearly $1-\rho \approx 0.8$ of those individuals will have become established by time $\alpha \lg n$, with their descendants having grown geometrically to reach a size of $\asymp 2^{\alpha \lg n} = n^{\alpha}$. Most of the remaining individuals will have become extinct, and a few will be in between (including that small set of people who might remain in jeopardy, depending on the value of α). Consider one of these established individuals who has a number of descendants whose order of magnitude is n^{α}. Each of these descendants has probability $\asymp 1/n$ of being a migrant to a neighboring island. So the probability that the individual has migrant descendants on each of the H_i neighboring islands in the "head start" set S_i is $\asymp (n^{\alpha}/n)^{H_i} = n^{(\alpha-1)H_i}$. Consequently, the expected number of individuals who have migrant descendants on each of the H_i neighboring islands in S_i by time $\alpha \lg n$ is $\asymp n^{1+(\alpha-1)H_i}$. Therefore, if $1+(\alpha-1)H_i < 0$, that is, $\alpha < \Delta_i = (H_i-1)/H_i$, then the expected number of individuals having completed a head start by time $\alpha \lg n$ converges to 0, so that the probability that any of the n individuals on island i has completed a head start by time $\alpha \lg n$ converges to 0 as $n\to\infty$. On the other hand, if $1+(\alpha-1)H_i > 0$, that is, $\alpha > \Delta_i$, then the expected number of individuals who have completed a head start by time $\alpha \lg n$ grows to infinity as $n\to\infty$. From this, together with a demonstration of asymptotic pairwise independence among the events that different individuals complete a head start by time $\alpha \lg n$ (along lines similar to the proofs of Lemmas 19 and 20 of [1]), it follows that the probability that at least one individual has completed a head start by time $\alpha \lg n$ converges to 1.

Lower bound in Theorem 1

Recall we call a given individual a *migrant* if either parent of that individual lived on an island other than the given individual's home island. Note that since all of our results concern time spans that are only order $\lg n$, there are only order $\lg n$ migrants in total.

Proposition L: Let s be a positive integer and let $\varepsilon > 0$. The probability that there exists an individual at a given time t_0 who has any descendant on an island at distance s steps away from that individual's home island within $(1-\varepsilon)(s-1)\lg n$ generations approaches 0 as $n \to \infty$.

To see this, note that if there is such an individual $I(t_0, i_0, m_0)$, then there must be a path of islands $i_0 \to i_1 \to \cdots \to i_s$ and a chain $I(t_1, i_1, m_1), I(t_2, i_2, m_2), \ldots, I(t_s, i_s, m_s)$ of migrant descendants of $I(t_0, i_0, m_0)$, with each $I(t_k, i_k, m_k)$ being a descendant of $I(t_{k-1}, i_{k-1}, m_{k-1})$, and $I(t_k, i_k, m_k)$ having a parent from island i_{k-1}. We know nothing particular about t_1 except that $t_1 - t_0 > 0$ (for example, it is likely that we could find individuals at time t_0 who have a migrant child, so that we could have $t_1 - t_0 = 1$). However, for each $k \geq 2$, we claim that with probability approaching 1, none of the differences $t_k - t_{k-1}$ will be less than $(1-\varepsilon)\lg n$. In fact, with probability approaching 1, no migrant up to a time of order $\lg n$ can have any further migrant descendants within $(1-\varepsilon)\lg n$ generations. That is, the probability that there exist times $t < t' < s \lg n$ satisfying $t' - t \leq (1-\varepsilon)\lg n$ and there exist migrant individuals $I(t, i, m)$ and $I(t', i', m')$ with $I(t', i', m')$ being a descendant of $I(t, i, m)$ approaches 0 as $n \to \infty$. The reason for this is that for any given individual, the probability of having a migrant descendant within $(1-\varepsilon)\lg n$ generations is of order $n^{-\varepsilon}$, so that since there are only order $\lg n$ migrants in total up to time $s \lg n$, the probability of *some* migrant having a migrant descendant within $(1-\varepsilon)\lg n$ generations is $O(n^{-\varepsilon} \lg n)$, which converges to 0.

Lower bound: For $\varepsilon > 0$, we have $P\{T_n \geq (1-\varepsilon)(R+\Delta)\lg n\} \to 1$ as $n \to \infty$.

To prove this, fix a node i_0. We want to show that the probability that some individual on i_0 becomes a CA within $(1-\varepsilon)(R+\Delta)\lg n$ generations approaches 0. We will assume i_0 is a center of the graph. This is the more involved case; we omit the similar but easier proof for a node that is not a center of the graph.

Recall we say that a head start has been completed when some individual has descendants who are established on each island in a set of islands that is within a distance of $R-1$ from every node of the graph. Let τ denote the time required for some individual living on i_0 at time 0 to have completed a head start. That is, before time τ, no individual on i_0 has completed a head start.

Let $t_0 = (1-\varepsilon)\Delta \lg n$. We have

$$P\{T_n \leq (1-\varepsilon)(R+\Delta)\lg n\} \leq P\{\tau \leq t_0\} + P\left[\{\tau > t_0\} \cap \{T_n \leq (1-\varepsilon)(R+\Delta)\lg n\}\right].$$

By Proposition H, $P\{\tau \leq t_0\}$ approaches 0 as $n \to \infty$.

Next we want to show that the probability of the event $\{\tau > t_0\} \cap \{T_n \leq (1-\varepsilon)(R+\Delta)\lg n\}$ converges to 0. For convenience, let $I(0, i_0, 1)$ denote whichever of the individuals on island i_0 at time 0 becomes a CA the fastest (if there is a tie, choose one arbitrarily). Let *A* denote the set of islands reached by the descendants of $I(0, i_0, 1)$ by time t_0. Let j be an island such that $d(A, j) = R$; we know we can find such a j whenever $\tau > t_0$. We observe that for both events $\{\tau > t_0 = (1-\varepsilon)\Delta \lg n\}$ and

$\{T_n \le (1-\varepsilon)(R+\Delta)\lg n\}$ to occur it must be the case that at least one of the following two events occurs:

(1) There exist $i \in A$ and m such that individual $I(t_0, i, m)$ has a descendant on island j within $(1-\varepsilon)(R-1)\lg n$ generations

(2) The remaining time, after the first migrant descendant of some individual on island i_0 reaches island j, required for some individual on i_0 to become a CA of j, is at most $(1-\varepsilon)\lg n$ generations. In other words, some set of migrants to island j "collectively" become a common ancestor of island j within $(1-\varepsilon)\lg n$ generations, in the sense that there is a set of migrants $\{I(u_1, j, m_1), I(u_2, j, m_2), \ldots, I(u_K, j, m_K)\}$ with $u_1 \le u_2 \le \cdots \le u_K$ such that the union of their sets of descendants at time $u_1 + (1-\varepsilon)\lg n$ is everyone on j. That is,

$$\bigcup_{k=1}^{K} \{\text{descendants of } I(u_k, j, m_k) \text{ on island } j \text{ at time } u_1 + (1-\varepsilon)\lg n\}$$
$$= \{I(u_1 + (1-\varepsilon)\lg n, j, m) : 1 \le m \le n/G\}.$$

By Proposition L and because $d(A, j) = R$, the probability of the event in (1) approaches 0 as $n \to \infty$.

For the event in (2) to occur, clearly at least one migrant to island j must have at least order $n/(\lg n)$ descendants within $(1-\varepsilon)\lg n$ generations, since there are a total of only order $\lg n$ migrants to island j within the time span relevant to the desired result. But the proof of Proposition 15 of [1] shows that the probability that a migrant can have order $n/(\lg n)$ descendants within $(1-\varepsilon)\lg n$ generations is very small – in fact, it is of order $o(n^{-p})$ for all p. So taking the union over the order $\lg n$ migrants still gives a probability of $o(n^{-p}\lg n)$, which approaches 0.

Upper bound in Theorem 1

Proposition U: Suppose an individual $I(0, i, m)$ from island i is established on island j in a given generation t, and let $\varepsilon > 0$. Then, with probability that approaches 1 as $n \to \infty$, by generation $t + (1+\varepsilon)\lg n$, individual $I(0, i, m)$ will be a CA of island j and will be established on each island neighboring island j.

Arguments from the upper bound in Theorem 1 of [1] applied to this setting show that with probability approaching 1, within $(1+\varepsilon/2)\lg n$ generations, $I(0, i, m)$ will become a CA of island j. And within an additional $(\varepsilon/2)\lg n$ generations [in fact in $o(\lg n)$ generations], among the descendants of $I(0, i, m)$ will also be migrants who have become established on each island neighboring j. In fact, there are $\asymp 1$ migrants each generation, and each migrant is destined to become established with a probability that approaches $1 - \rho \approx 0.8$, so that it takes only $O(1)$ generations for migrants who are destined to become established to reach each island neighboring j. Furthermore, with probability approaching 1, each migrant who is destined to become established will in fact do so within a time that is only $O(\lg\lg n) = o(\lg n)$.

Upper bound: For each $\varepsilon > 0$, we have $P\{T_n \le (1+\varepsilon)(R+\Delta)\lg n\} \to 1$ as $n \to \infty$.

Let i^* be an island achieving the minimum in the definition $\Delta = \min_{i \in C(\mathbb{G})} \Delta_i$ and let S^* denote S_{i^*}. By Proposition H, within $(1+\varepsilon)\Delta \lg n$ generations, some individual $I(0, i^*, m)$ on island i^* will become established on each island in the set $\{i^*\} \cup S^*$. Next, by induction, applying Proposition U repeatedly, we see that within an additional $k(1+\varepsilon)\lg n$ generations, individual $I(0, i^*, m)$ will have become a CA of all islands whose distance from $\{i^*\} \cup S^*$ is less than *k*, and will be established on all islands whose minimum distance to $\{i^*\} \cup S^*$ is equal to *k*. In particular, since all islands are within a distance of $R-1$ from $\{i^*\} \cup S^*$, it follows that by generation $(1+\varepsilon)\Delta \lg n + (R-1)(1+\varepsilon)\lg n$, individual $I(0, i^*, m)$ will have become established on all islands in the graph. From here, an additional $(1+\varepsilon)\lg n$ generations suffices to complete the process, making $I(0, i^*, m)$ a CA of all islands.

Lower bound in Theorem 2

Let $\zeta \approx 0.7698$ be as defined in Theorem 2 of [1].

Lower bound: For each $\varepsilon > 0$, $P\{U_n \ge (1-\varepsilon)(D+1+\zeta)\log_2 n\} \to 1$ as $n \to \infty$.

Let islands i_0 and i_1 be separated by a distance of *D* from each other, and define $t_1 = (1-\varepsilon)(D+1+\zeta)\log_2 n$. We claim that with probability approaching 1, there are individuals $I(0, i_0, m_0)$ and $I(t_1, i_1, m_1)$ such that $I(0, i_0, m_0)$ is not extinct at time t_1 and $I(0, i_0, m_0)$ is not an ancestor of $I(t_1, i_1, m_1)$. In other words, the claim is that with probability approaching 1, at time t_1, there is an individual $I(0, i_0, m_0)$ on island i_0 who is not extinct but is not yet a CA of island i_1. By [1], we know that there are many (i.e. a number that approaches infinity as $n \to \infty$) individuals living on island i_0 at time 0 who, at time $(1-\varepsilon)\zeta \lg n$, are destined to become CA's but are not yet established and also have no descendants on any island other than i_0. Let $I(0, i_0, m_0)$ be one of these individuals. Now we just need to show that with probability approaching 1, it will take more than $(1-\varepsilon)(D+1)\lg n$ generations for $I(0, i_0, m_0)$ to become a CA of island i_1. This follows from Proposition L, from the same reasoning as used earlier to establish the lower bound in Theorem 1.

Upper bound in Theorem 2

Proposition U2: The probability that an individual who is established on a given island fails to become a CA of that island within $(1+\varepsilon)\lg n$ additional generations is $o(1/n)$.

This follows from minor modifications of the analogous result in [1]. The same arguments work because this statement concerns the number of descendants of a given individual on just a single island. The only issue to check here is that the result is not changed by those few individuals per generation who may have

a child on a different island; it turns out that this makes no important change in the behavior of the process of counts of descendants of a given individual on a single island.

Upper bound: For $\varepsilon > 0$, $P\{U_n \le (1+\varepsilon)(D+1+\zeta)\log_2 n\} \to 1$ as $n \to \infty$.

Let $0 < \delta < \varepsilon$.

Establishment Stage: From [1], we know that within $(1+\delta)\zeta \lg n$ generations, everyone (that is, all individuals $I(0,i,m)$ on all islands at time 0) is out of jeopardy – either extinct or established.

Let us call the individuals $I(0,i,m)$ who become established the "original established individuals." The remaining individuals from time 0 are all extinct at the end of the Establishment Stage.

To complete the proof we show that with probability approaching 1, the original established individuals will all become CA's of the full population within $(1+\delta)(D+1)\lg n + o(\lg n)$ additional generations. We have $D+1$ additional stretches of $(1+\delta)\lg n$ generations to work with.

Growth Stage 0: We wait until all established individuals have become CA's of their own islands.

Proposition U2 implies that with probability approaching 1, Growth Stage 0 requires less than $(1+\delta)\lg n$ generations. After Growth Stage 0 is completed, we begin Migration Stage 1.

Migration Stage 1: We monitor migrants and wait until we have seen, for each edge of the graph and for both directions along that edge, a migrant along that edge in that direction, with that migrant being established on the destination island.

Migration Stage 1 is readily seen to take only $o(\lg n)$ generations. In fact, the time required to collect a full set of migrants who are destined to become established on the destination island (but have not yet become established) has a distribution that is easily upper bounded by a geometric decay, so this time contributes just $O(1)$ to Migration Stage 1. Then, the additional time required for those migrants who are destined to become established actually to become established is just $O(\lg\lg n) = o(\lg n)$.

At the end of Migration Stage 1, each of the original established individuals is a CA of his home island and has become established on each island within a distance of 1 from his home island.

Growth Stage 1: We wait until the established migrants found during Migration Stage 1 all have become CA's of their islands.

Just as with Growth Stage 0, with probability approaching 1, Growth Stage 1 takes less than $(1+\delta)\lg n$ generations. At the end of Growth Stage 1, each of the original established individuals has become CA of each island within a distance of 1 from his home island. We continue to define Migration and Growth Stages in the same way, inductively.

Migration Stage k: After the end of Growth Stage $k-1$, we begin monitoring migrants and wait until we have seen, for each edge of the graph and for both directions along that edge, a migrant along that edge in that direction, with that migrant being established on the destination island.

Growth Stage k: We wait until the established migrants found during Migration Stage k have all become CA's of their islands.

As above, with probability approaching 1, Migration Stage k takes $o(\lg n)$ generations and Growth Stage k takes less than $(1+\delta)\lg n$ generations. At the end of Growth Stage k, each of the original established individuals has become CA of each island within a distance of k from his home island.

Since each island is within a distance of D from every other island, it follows that at the end of Growth Stage D, each of the original established individuals is a CA of the full population. With probability approaching 1, the total time taken for this to occur is less than $(1+\delta)\zeta \lg n$ for the Establishment Stage, plus $(1+\delta)\lg n$ for Growth Stage 0, plus $D\left((1+\delta)\lg n + o(\lg n)\right)$ for Migration Stage 1, Growth Stage 1, …, Migration Stage D, and Growth Stage D. That is, as n approaches infinity, $P\{U_n \le (1+\delta)\zeta \lg n + (1+\delta)\lg n + D\left((1+\delta)\lg n + o(\lg n)\right)\} \to 1$. Since $\delta < \varepsilon$, this implies $P\{U_n \le (1+\varepsilon)(D+1+\zeta)\lg n\} \to 1$.

Reference

[1] Chang, J. T. Recent common ancestors of all present-day individuals. *Adv. Appl. Probab.* **31**, 1002-1026, with invited discussion and author's reply, 1027-1038 (1999)

Modeling the recent common ancestry of all living humans
Supplementary Methods B: Further Details of the Computational Model

Douglas L. T. Rohde, Steve Olson, Joseph T. Chang

This supplement provides additional details about the implementation and analysis of the computational model of human mating and migration introduced in the main paper. The model simulates the lives of individual people, known as *sims*, over the course of thousands of years, including such events as the sims' birth and death, possible migrations, mate choices, and offspring production. As the model runs, it records this information in a series of large computer files. A second program, discussed in Section 6, traces ancestral lines through this data to identify the common ancestors.

1 Lifespan

The model does not assume discrete, uniform generations. Each sim is born in a certain year and has a particular life span. The maximum age of any sim is 100, as it was presumably quite rare, prior to modern medicine, for someone to live, let alone father children, beyond that age. The age of sexual maturity is taken to be 16 years for both men and women. Anyone who would have died before that age could not have produced offspring and is thus not a factor for the purposes of this study. Therefore, only the lives of those destined to at least reach adulthood are actually simulated. As a result, the population sizes discussed throughout this paper are effectively somewhat larger than stated because they do not include any children.

Otherwise, the probability that an individual dies at age s, conditional on not having died before age s, is assumed to follow a discrete Gompertz-Makeham form (Pletcher, 1999):

$$p(s) = \alpha + (1 - \alpha) \exp\{(s - 100)/\beta\}$$

In this equation, β is the *death rate*. A higher death rate results in shorter life spans on average, although the effect is not linear. The α parameter is the *accident rate*, which can be adjusted to reflect the probability that an individual of any age dies of unnatural causes. With an accident rate of 0.01 and a death rate of 10.5, this formula quite closely models the life span data for the U.S. between 1900 and 1930 (U.S. National Office of Vital Statistics, 1956). To account for historically shorter life spans due to poor nutrition, medicine, and so forth, the death rate, β, was raised to 12.5 for the purposes of the model. This produces an average life span of 51.8 for those who reach maturity.

2 Mating

Another important component of the model is the system by which mates are chosen and children are produced. In this respect, the model was implemented from the perspective of the mother. It first determines the years in which the mother will give birth, and then a father is chosen for each child. The assumption is made that women give birth between the ages of 16 and 40, inclusive, with an equal probability of producing a child in each of these years. Of course, some women may produce many children and others will produce none, and some may die before age 40. After taking this latter factor into account, we can control population growth by adjusting the average number of children (who reach adulthood) per woman. A value of exactly 2 children per woman results in a stable population size.

Once it has been determined that a woman will give birth in a certain year, the father is chosen. If possible, the father is always selected from the town in which the mother lives. It sometimes happens, especially early in the simulation when populations are low or when a new area is first colonized, that there are no suitable fathers living in the same town as a woman who is to have a child. In this case, fathers are sought in the other towns within the same country.

The father of a woman's first child is selected at random from the men who are at least as old as the woman. The prohibition against younger husbands was primarily for computational reasons, but it seems to be a fairly reasonable, if not entirely valid, assumption. There is an additional bias such that men are twice as likely to be chosen if they are not already married, in the sense that they have already produced a child with another woman. This tends to make mate selection more equitable. After the first child, there is an 80% chance that the father of the previous child will also father the next one, thus simulating marriage. There is a fundamental asymmetry in the sexes, in that a woman can only be "married" to one man, although a man could be married to more than one wife, or at least fathering children by more than one woman; but

there is a bias towards monogamous relationships. Also note that women cannot bear children past the age of 40, while men can father children throughout their adult lives.

As a result of these assumptions, the distribution of children per woman is essentially binomial, with 19% producing no (adult) children and only 2.8% producing more than 5 children. The distribution for men has greater variance, with nearly 36% of men producing no children and 8.6% producing more than 5 children. Thus, there is a higher percentage of men than women that have no children or many children, but relatively fewer men with a moderate number of children. The average age of a parent when a child is born is approximately 30 years, so this will be taken as the length of a generation.

3 Migration Overview

The model is organized into three structural levels: continents, countries, and towns. The continents, depicted in Supplementary Figure 1, represent physically separated land masses that are likely to have very low rates of intermigration, which we will carefully control. The models' 12 continents are divided into *countries*, arranged in a grid. These reflect major tribal, ethnic, or language groups, with both geographic and cultural barriers to intermarriage. Each country represents approximately 119,000 square miles, with the exception of Oceania, in which the countries are intended to resemble the major island groups and are typically much smaller in terms of both area and population. The distances shown between the continents in Supplementary Figure 1 are arbitrary, the only important factor being the number and migration rate of the ports connecting them, which we will discuss shortly.

The countries are divided into towns. These do not necessarily represent towns per se, but the relevant social unit from within which most people find mates. Thus, a town may actually reflect a clan, a rural county, or even a particular social class within a larger group. The towns within each country are assumed to be in relatively frequent contact with one another and are not in any particular geographic arrangement.

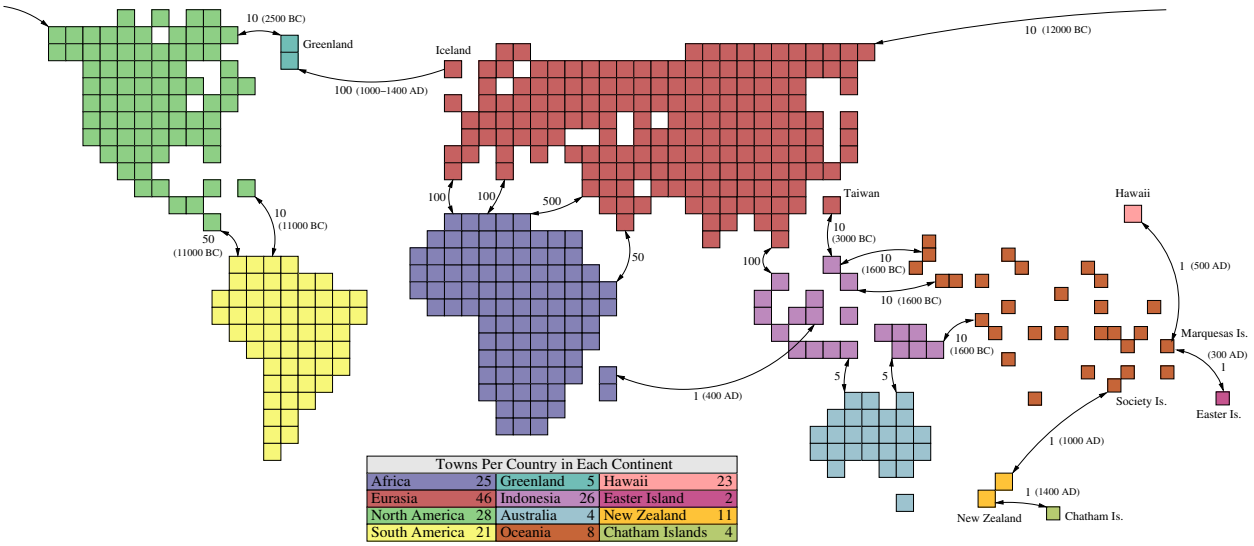
Not all humans confine themselves to a single location throughout their lives and a critical factor in the model is the rate at which people migrate to different places in the world. Although it seems likely that many people, and perhaps the vast majority historically, live out their lives close to where they were born, various forms of migration lead to the gradual spread of ancestral lineages over long distances. When men and women from different groups marry, one of them, often the wife but sometimes the husband, moves to the other's community. Merchants, soldiers, and bureaucrats, who are typically male, sometimes travel widely, potentially fathering children far from their place of birth. And, occasionally, large groups have conquered or colonized new areas.

The model uses a simplified migration system, in which each person can move only once in his or her life. A sim is born in the town in which his or her parents, or at least mother, lives, but then has a chance to migrate to a different continent, country, or town prior to adulthood. Henceforth, that person can produce children only with other inhabitants of his or her new town, provided it contains some potential mates.

As is the case in human mating patterns (Fix, 1979), the rate of exogamy decreases substantially with larger group size in the model. Adams and Kasakoff (1976) found that, across a variety of human societies, there was a recognizable threshold in group size at around a 20% exogamy rate, although the sizes of these groups differed as a function of population density. This "natural" group size is taken here to be that of the town. The *ChangeTownProb* parameter controls the percentage of sims who leave the town of their birth for another town within the same country. It typically ranges from 20% down to 1%.

There is a much lower chance that a sim will leave his or her home country for another country on the same continent. The probability that this occurs is governed in the model by the *ChangeCountryProb* parameter, which ranges from 0.1% to 0.001% (1 in 100,000), and is therefore a fixed fraction of the individuals who reach adulthood. The countries within a continent are arranged in a grid and locality plays a role in inter-country migration. The probability of reaching any other country in the continent is proportional to the inverse square of the Euclidean distance to the new country. Thus, the probability of traveling a distance of 2 countries is 1/4 that of traveling to a neighboring country, and the probability of traveling from a country at the northern tip of South America to one at the southern tip is less than 1% that of traveling to a neighboring country.

It is important to keep in mind that migration between countries is quite rare in the model. In the year 1500 AD, there will be about 191,000 people in each country in Eurasia, which translates to 111,000 born every generation. If the *ChangeCountryProb* is set to 0.05%, as in the first (more conservative) simulation reported in the paper, we can expect only 55.3 sims to leave each country per generation, or 1.8 each year. Because most of these migrants will go to neighboring countries, truly long-distance migrations only occur a few times per century. In other continents and during earlier time periods, population density, and therefore the number of inter-country migrants, is even lower. In the same year, Africa and Oceania have about 30.0 migrants per generation leaving each country, while South America has 22.1, North America has 17.6, and Australia has only 0.98. Thus, even the



Supplementary Figure 1: Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.

more liberal model reported in the paper, which has five times this rate of inter-country migration, is still quite conservative in this respect.

Intercontinental migration takes place through *ports*. Ports lead from a source country in one continent to a destination country in another. The rate of migration through a port can be regulated and monitored and is expressed in terms of migrants per generation. In most of the simulations, the majority of the sims using a port are born locally, in its source country, while a proportion of port users, governed by the *NonLocalPortProb* parameter, are drawn from random countries within the continent, including the source country. These typically account for 5% to 20% of the port users.

It is often the case in modern times, and presumably throughout history, that immigrants to a new continent will gravitate towards a sub-community of fellow immigrants who share the same cultural or linguistic background. The result is a delay in the exchange of lineages between the immigrants and hosts. This is simulated in the model by having new immigrants initially choose from one of five towns, out of up to 46, in the destination country. This set of towns is dependent on the source country from which the migrant came. As a result, immigrants with similar origin will tend to cluster together, though they will not be entirely segregated.

The migration choices of individual sims in the model are independent. However, there is a problem when a port opens to a previously uninhabited continent. Pioneers to this new territory cannot organize a sustainable colony in advance, and, because the rate of migration to new countries is typically very low, individual migrants will usually find themselves isolated and unable to reproduce. Therefore, the pioneers would tend to die off and it could take quite some time for them to gain a foothold. The result is that the earliest migrants into the Americas and Oceania would not spread out evenly but would tend to cluster around the port countries, only advancing once the population there reached sufficient density. It may take centuries before a sustainable population could take hold on a remote island.

Therefore, in order to avoid this problem, any sim who reaches an uninhabited town is essentially cloned and five more sims, of random sex, are created to join him or her. These new sims are given the same parents so the rate of lineage spread is minimally affected. This may be a reasonable assumption, given that most organized pioneering groups were probably quite closely related. With any luck, this new colony will be a sustainable, albeit incestuous, breeding population. Additionally, newly colonized countries will usually have considerably higher than average population growth rates, as discussed in Section 5

4 Migration Details

The simulations typically begin in the year 20000 BC, at which point the populated areas only include Africa, Eurasia, Indonesia (including New Guinea), and Australia. Some of the inter-continental ports are already open at the start of the model and remain at a fixed migration

rate, in terms of the expected number of sims per generation traveling in each direction. The ports are shown as arrows in Supplementary Figure 1, labeled with these migration rates. Between Africa and Eurasia, there are ports connecting modern-day Morocco and Spain (100 sims/generation), Tunisia and Italy (100 s/g), Egypt and Israel (500 s/g), and between Ethiopia and Yemen (50 s/g), providing several points of contact. Other static ports include a pair between Thailand and Malaysia (100 s/g), and from the tip of Indonesia (Timor) to Arnhem Land and from New Guinea to Cape York, both with rates of just 5 s/g. Aside from those already mentioned, the remainder of the ports in the model only open at particular points in time, indicated in Supplementary Figure 1 by the dates in parentheses.

The migration rates used in this model are not based on firm historical data, because such information is, for the most part, unknown (Jorde, 1980). They are based almost entirely on estimates, loosely taking into account proximity, population density, and available seafaring technology. Without a firm basis in fact, an attempt was made to err on the side of conservatism. Some of the migration rates may be considerably smaller than they should be, and many routes are undoubtedly missing. Some readers will disagree with particular details of the timing, location, and migration rate of these routes. Greater accuracy will certainly improve the quality of the results generated by the model and our confidence in them. However, experience suggests that its results are quite stable and insensitive to all but the most significant changes.

The port between the eastern tip of Siberia (Chukotka) and Alaska opens in the year 12000 BC. There continues to be scientific debate over the date of the first human arrival in North America, but this seems to fall at about the median of suggested dates. As with most other new ports, this one begins at a higher rate to create an initial wave of migrants. In the first generation, there are about 100 migrants from Chukotka to Alaska, with 10 in the reverse direction. Subsequently, the port rate remains at 10 s/g in both directions. A continuous, low rate of contact between Siberia and Alaska following the close of the Bering land bridge is supported by the available archaeological evidence. "It would appear…that Bering Strait was never a hindrance to the passage of materials and ideas among local populations living along both its shores," (Arutiunov & Fitzhugh, 1988, pg. 129). It seems reasonable to assume that this exchange of technology and culture was accompanied by, and perhaps driven by, a gradual exchange of people between the two continents.

One thousand years after the first migrants enter North America, ports open between Panama and Columbia (50 s/g) and between the Caribbean islands and Venezuela (10 s/g). These do not have an initial migration burst, as it is assumed that the earliest inhabitants would have gradually diffused throughout North America and into South America over the span of one or two thousand years. Much later, in 2500 BC, an additional port opens between Baffin Island and Greenland, to simulate the advance of Pre-Dorset or Independence I Inuit, whose earliest northern Greenland sites have been dated to 2400 BC (Arutiunov & Fitzhugh, 1988; Grønnow & Pind, 1996).

The Polynesian colonization of the Pacific islands is believed to have had its source in the expansion of the Ta-p'en-k'eng culture from Taiwan into the Philippines and later into Indonesia. This was followed, around 1600 BC, by the fairly rapid spread of the Lapita culture to Micronesia and Melasia and then eastward throughout Polynesia (Diamond, 1997; Cavalli-Sforza, Menozzi, & Piazza, 1994). This is simulated in the model by the opening of a direct port between Taiwan and the Philippines in 3000 BC, with an initial burst of 1000 migrants, settling to an exchange of 10 s/g. In 1600 BC, three more ports open, from the Philippines to the Mariana islands and Micronesia, and from New Guinea to the Solomons.

Most of the other inhabitable Pacific islands are then colonized via the standard inter-country migration mechanism. At this early stage, assuming a *ChangeCountryProb* of 0.05%, the most populous of the islands produce about 3 emigrants per generation, most of whom settle in neighboring islands. At this rate, it takes about 600 years for the majority of the island groups to be reached. Note that the inter-country migration mechanism does not only support the initial population spread but also the continuous exchange of people between neighboring islands. This is consistent with the recent view that early Polynesian societies were not entirely isolated (Terrell, Hunt, & Gosden, 1997), and yet the rate of long-distance migration is so low that it would not seem to contradict the views of critics who argue that such contacts were probably very rare.

Some of the more remote islands are not colonized until much later, including Easter Island (Rapa Nui), Hawaii, New Zealand, and the Chatham Islands, which are treated in the model as separate continents. Easter Island is reached from the Marquesas Islands in 300 AD, with an initial wave of 50 migrants followed by a steady exchange of just 1 per generation. Hawaii is reached from the Marquesas in 500 AD, with an initial wave of 200 migrants, although there is some question as to whether the first colonizers might have come from Tahiti or the Cook Islands. Meanwhile, in 400 AD, migrants begin traveling from Borneo to Madagascar, with an initial wave of 100. Although there is also some question about the source and date of the first inhabitation of New Zealand, it is settled in the model from the Society Islands in 1000 AD with an initial wave of 200 migrants. The last place to be populated is the Chatham Islands, reached from New Zealand in 1400 AD by a wave of 100 migrants.

Southern Greenland is known to have been colonized

by Vikings from Iceland in 985 AD. They were visited regularly for several hundred years and are thought to have died out or been assimilated by the Inuit sometime before 1500. In the model, a port opens from Iceland to Greenland in the year 1000, with 1000 initial inhabitants followed by 100 more per generation until 1400. There is no migration in the reverse direction because of the likelihood that no Inuit reached Iceland or other parts of Europe during the time period in question.

After 1500 AD, several additional large ports, not shown in Supplementary Figure 1, are opened to simulate colonization of the Americas and elsewhere. These include migration routes between Spain and Peru, Mexico, and the Caribbean, and between Portugal and Brazil. In 1600, ports open from England to the eastern U.S., from France to eastern Canada, from Spain, France and west Africa to the southern U.S., and from west Africa to the Caribbean and Brazil. In 1700, a port opens from Denmark to Greenland and in 1800 many more ports open, including various ones from Europe and China to the U.S., from England to South Africa, Australia, India, and New Zealand, and from the western U.S., China, and Japan to Hawaii. Most of these ports are quite substantial, with rates between 1,000 and 5,000 immigrants per generation in the primary direction of colonization, with 100 to 200 in the opposite direction. As discussed in Section 5, the first European migrations to North and South America are coincident with a significant decline in the size of the native populations due to disease.

In order to model generally increased mobility, the *NonLocalPortProb* was gradually increased towards the end of the simulation. A higher *NonLocalPortProb* permits more sims from outside of the source country to use a port, increasing the overall frequency of long-distance migration. In most of the simulations, this parameter starts at 5%, but increases to 20% in the year 1500 AD, 50% in 1600, 75% in 1700, 85% in 1800, and 90% in 1900. Smaller increases are used for the more conservative models. The *ChangeTownProb* also increases in recent centuries from an initial value of 5% to 10% in 1700 and 20% in 1900, with greater increases for the simulations with a baseline of 10%. The *ChangeCountryProb* likewise increases to simulate greater mobility, doubling in the years 1500, 1750, and 1900.

5 Population

Human population density differs throughout the world. Historically, this can be attributed to such factors as climate, disease, and the methods and success of food production. These differing densities are likely to have a significant impact on the distribution of common ancestry. Lineage will tend to spread faster, as a function of distance, with higher density populations because of the greater number of migrants. It is important, therefore, that the model take into account differing population density throughout the world.

The roman numbers in Supplementary Table 1 give the population estimates in each of the modeled "continents" at various points in time. These numbers are based primarily on Table 2.1.2 of Cavalli-Sforza, Menozzi, and Piazza (1994), which was itself adopted from Biraben (1980), as well as on other estimated populations found throughout their book. Other values were taken from various sources or were interpolated or extrapolated as necessary. The earliest values were set to achieve the desired overall world population with a gradually increasing proportion of inhabitants in Eurasia relative to Africa.[1]

Due to computational constraints, it was not possible to simulate world populations much larger than 60 million sims. Therefore, natural-size populations were used until the population reached 50 million, which occurs around the year 1000 BC. Reduced populations were used thereafter to achieve a maximum world population of 55 million. If the population is reduced after the death of the MRCA, it should have little effect on the results because this growth will not necessarily alter the percentage of the population descending from that ancestor, which is the primary determinant of the rate of spread of his or her lineage. If anything, smaller populations may result in less recent MRCAs because of the reduced intra-continental migration. So it is hoped that the population cap in this model will not lead to overly recent estimates.

A straightforward approach to limiting the world population would be to scale the population in every continent by the same factor. In the year 1970, this would require scaling the population by a factor of 1/68, from 3.75 billion to 55 million. However, this may have a serious impact on the small continents. The population of the average Greenland town would be reduced from 5,600 to 82, while the population of the Chatham Islands would be reduced from 1,000 to 15. These changes would force such populations below the lowest sustainable level of a few hundred sims and would have a serious impact on the effective migration rates out of the small countries. With a *ChangeCountryProb* of 0.01%, a country of 200,000 people can expect a sim to emigrate every 2.6 years. If the population is reduced by a factor of 10, the expected delay between sims would increase to 26 years, a significant but not necessarily detrimental change. However, if a country's population is scaled from 20,000 to 2,000, the expected delay between emigrants would increase from 26

[1] The final numbers in Supplementary Table 1 are based on data from 1970. However, in the model, these were used to determine the year 2000 population targets. The approximate doubling of the world population between 1970 and 2000 should have little or no effect on the outcome.

Supplementary Table 1: Continental populations, in thousands, at various points in time. The roman numbers are estimates of the true populations. The italic numbers below them are the rescaled values used in the simulations to achieve a maximum world population of 55 million.

Continent	20K BC	15K BC	10K BC	5K BC	2K BC	1K BC	500 BC	1 AD	500 AD	1000	1250	1500	1750	1970
Eurasia	1230	2030	2850	3350	18700	38800	125000	217000	158000	193000	323000	320000	629000	2722000
	1230	*2030*	*2850*	*3350*	*18700*	*38800*	*43979*	*44288*	*40251*	*38814*	*38513*	*34170*	*41655*	*37307*
Africa	670	870	950	1100	3220	5290	17000	26000	31000	39000	58000	87000	104000	353000
	670	*870*	*950*	*1100*	*3220*	*5290*	*6735*	*6371*	*8474*	*8434*	*7737*	*9474*	*7880*	*6192*
S. America	0	0	50	200	1500	3000	4000	5000	8000	12000	23000	40000	15000	283000
	0	*0*	*50.0*	*200*	*1500*	*3000*	*1882*	*1679*	*2556*	*2925*	*3271*	*4435*	*1876*	*4234*
N. America	0	0	50	200	1000	1500	2000	3000	5000	10000	20000	35000	5000	228000
	0	*0*	*50.0*	*200*	*1000*	*1500*	*1348*	*1581*	*2293*	*3195*	*3755*	*4862*	*1733*	*4639*
Indonesia	50	50	50	100	500	1000	1000	2000	3000	5000	8000	12000	16000	119000
	50.0	*50.0*	*50.0*	*100*	*500*	*1000*	*545*	*689*	*995*	*1227*	*1215*	*1462*	*1340*	*1788*
Australia	50	50	50	50	70	100	100	100	100	100	200	250	250	20000
	50.0	*50.0*	*50.0*	*50.0*	*70.0*	*100*	*66.1*	*59.5*	*61.6*	*59.2*	*81.2*	*88.2*	*83.0*	*317*
Oceania	0	0	0	0	0	300	1000	1000	1000	1000	2000	3000	3000	19000
	0	*0*	*0*	*0*	*0*	*300*	*439*	*329*	*364*	*324*	*381*	*449*	*366*	*430*
New Zeal.	0	0	0	0	0	0	0	0	0	2	50	100	150	3000
	0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1.9*	*18.6*	*24.9*	*26.3*	*53.8*
Hawaii	0	0	0	0	0	0	0	0	0	20	50	100	200	800
	0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*12.3*	*19.1*	*25.5*	*30.3*	*30.7*
Greenland	0	0	0	0	10	10	10	10	10	15	15	20	25	56
	0	*0*	*0*	*0*	*10.0*	*10.0*	*6.5*	*5.9*	*6.1*	*7.5*	*6.9*	*7.8*	*8.1*	*9.0*
Chatham Is.	0	0	0	0	0	0	0	0	0	0	0	2	2	1
	0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1.4*	*1.4*	*0.8*
Easter Is.	0	0	0	0	0	0	0	0	2	5	10	10	2	0
	0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1.2*	*2.0*	*2.5*	*2.4*	*1.1*	*0*
Total	2000	3000	4000	5000	25000	50000	150110	254110	206112	260142	434325	497482	772629	3747860
	2000	*3000*	*4000*	*5000*	*25000*	*50000*	*55000*	*55002*	*55001*	*55001*	*55000*	*55002*	*55000*	*55001*

to 260 years. This is likely to have a much more profound effect on the resulting rate of lineage spread.

Thus, a uniform scaling of population sizes will tend to have a greater impact on the small towns, countries, and continents. To avoid this problem, the estimated continental population sizes were scaled in the model in such a way that more of the impact falls on the more densely populated continents. The actual scaling was done with the following formula:

$$S_n = P_n \frac{K\frac{P_n}{T_n} + 1000}{\frac{P_n}{T_n} + 1000}$$

P_n is the full estimated population of continent n, S_n is its scaled down population, and T_n is the number of towns in the continent. Therefore, $\frac{P_n}{T_n}$ is the average town population, a measure of population density. K is the scaling factor, which is adjusted until the overall scaled population of the world reaches the desired level of 55 million. The italicized values in Supplementary Table 1 give the scaled populations that were actually used in the model. As a result of this formula, the year 1970 population of Eurasia is scaled by a factor of 73, from 2.7 billion to 37.3 million. The population of the smaller continents are reduced to a lesser extent: North America by a factor of 49 and Hawaii by a factor of 26, while the Chatham islands are only scaled down from 1000 to 800 sims.

The scaled population values cannot be strictly enforced in the model, but merely serve as targets, which the simulator attempts to achieve by making small adjustments to the birth rate in each continent. However, the growth rate of the population is not always the same throughout the continent. Diamond (1997) has noted that colonists to virgin lands are likely to experience higher than average population growth rates, presumably due to a lack of competition for resources. This is simulated in the model using a population balancing mechanism by which smaller towns will have higher than average growth rates. The formula for the average number of children per woman, C_c, in country c is:

$$C_c = \frac{C_n}{2}\left(1 + \frac{\overline{P}_{Cn}}{P_c}\right)$$

C_n is the desired number of children per woman for the continent as a whole, which is determined by the population growth targets. $\overline{P}_{Cn}$ is the average current population per inhabited country in the continent, while P_c is the population of country c. As a result of this formula, the overall weighted average number of children per woman is still equal to C_n, but the birth rate will be higher in the less densely populated countries, up to a maximum bound of 4 children per woman.

In order to simulate the dramatic reduction in native American populations as a result of European-introduced diseases (Stannard, 1992), the populations of these continents were scaled back starting in the year 1400. The population targets shown for North and South America under the year 1500 in Supplementary Table 1 were actually the targets used for 1400. At that point, the birth rate was reduced, causing the loss of much of the native population. The rate of this decline reached its peak around the year 1500, as Europeans began to arrive. The net effect of this was somewhat greater than intended, resulting in the loss of 97% of North Americans and 93% of South Americans before the populations began to recover in 1570. Diamond estimates that the actual decline may have been as large as 95%. It is unlikely that the more severe decline in North America will have a noticeable effect on the results of the simulation.

Because the population density varies between continents, the number of towns per country was adjusted to produce towns of reasonable average size. These counts are given in Supplementary Figure 1. In the year 1500 AD, the primarily agricultural continents have approximately 4,000 inhabitants per town. The primarily non-agricultural continents, including North America, Australia, Greenland, and Easter Island had approximately 2,000 inhabitants per town, while the Chatham Islands had 500. Overall, the model contains 497 countries and 15,059 towns.

5.1 Initialization

There is one remaining aspect of the model to be described, which is its method of initialization. Some initial sims are needed in order to get things going. A simple approach might be to create all of the initial sims in the same year. However, in that case, their children would form a baby boom and it would take some time for the age distribution within the population to stabilize. Unless that stable age distribution is known in advance, there will always be some instability introduced by the creation of the initial people.

Therefore, the simulation actually begins 100 years before the desired start date. An initial set of sims is generated, each in a random town and each born at a random time within a 40-year window. The model is then run as usual, with the initial sims starting to produce offspring. Although the population does not have a natural age profile initially, as there are no old people, it quickly settles into a near-normal distribution within the first 100 years. The population will roughly double during these first 100 years as fewer people die of old age than are born. Thus, the size of the initial population is adjusted to achieve the desired level at the end of the 100-year period.

6 Finding common ancestors

A simulation with a maximum population of 50 million sims will involve a total of approximately 1.2 billion sims over its course. As the model runs, it generates files containing the vital statistics of each sim, including his or her parents, sex, birth and death years, and place of birth, typically totaling about 60 gigabytes of compressed data per trial. Although running the simulation is relatively easy, analyzing this genealogical data to identify the common ancestors presents a significant computational problem.

Let us refer to all of the sims alive in the year 2000, when the simulations end, as *living sims*. A true common ancestor (CA) is someone who is an ancestor of all living sims. A straightforward search for common ancestors would start with the living sims and work backwards in time, tracking for every other sim which of the living sims are his or her descendants. These descendants are the union of all descendants of his or her children. Tracking these descendants would be fairly simple, except that it requires memory proportional to the square of the number of living sims. With a maximum population of 50 million, this would involve the computation and storage of over 300 terabytes of information.

Therefore, finding the common ancestors is not tractable using a straightforward approach. However, a method was developed to zero in on the common ancestors using an initial approximation followed by a series of refinements. This process begins by tracking the ancestry not of all living sims, but of a small, randomly selected subset of them. Depending on the available computer memory, there are typically between 192 and 512 of these individuals, who are known as *tracers*. By working backwards through the records, the ancestry of these tracers is determined. This is done by computing, for every other sim, a bit vector in which the ith bit is turned on if that sim is an ancestor of the ith tracer. Aside from the fact that the ith tracer automatically has his or her own bit turned on, a parent's bit vector will be the bit-wise disjunction of his or her children's vectors. These bit vectors still present a heavy memory burden, but can be handled more efficiently by storing only the unique vectors.

If a sim is not an ancestor of every one of the tracers, that sim could not possibly be a common ancestor (CA). However, if a sim is a common ancestor of all of the tracers, there is a high probability that the sim is an ancestor of a large proportion of the living sims. Such ancestors are referred to as *potential common ancestors* (PCAs). Unfortunately, it is generally the case that the most recent PCAs that are found in this first backward phase are not actually true CAs. Therefore, this superset of the CAs must be refined.

The next step is to start with a set of the most recent PCAs and trace their lineage forward through time.

This is done in much the same way that descendancy was traced in the backward phase—a sim's ancestors are the disjunction of his or her parents' ancestors. In this case, we eventually determine which of the most recent PCAs is an ancestor of each of the living sims. If one of the PCAs was an ancestor of all of the living sims, then we are guaranteed to have found the true MRCA. Otherwise, a new set of tracers is chosen and a second backward pass is performed to refine the set of PCAs.

Selecting the new set of tracers randomly would help a little bit, but not much. A more effective approach is to try to find the sims who are difficult to reach, meaning that they descend from the fewest number of the PCAs. We also need to find a diverse set of tracers. If they are all difficult to reach because they live in the same place, the use of more than one as a tracer would be redundant. In order to satisfy these constraints, the tracers are selected sequentially, with the next tracer chosen being the living sim with the highest score, defined as follows:

$$\text{score}_i = \sum_{p \in P} 2^{-\left(x_{p,i} \sum_{t \in T} x_{p,t}\right)}$$

In this equation, i is the sim being considered as a possible tracer. P is the set of PCAs whose descendants were tracked. The indicator variable $x_{p,i}$ is 1 if sim i is *not* a descendant of PCA p, and 0 otherwise. T is the set of tracers that have been selected thus far. This method essentially balances the number of new tracers that are not descended from each of the PCAs, thus increasing the diversity of the new tracers.

Once these tracers have been chosen, their ancestors are found as in the first step. In this case, sims are only identified as PCAs if they are ancestors of all of the new tracers and all of the original tracers. For this purpose, the prior PCA-status of every sim is stored using a compressed run-length encoding. The most recent PCAs are once again selected and their lineages traced forward through time. It is usually the case that one of these new PCAs is actually a CA, which means we have found the true MRCA. Occasionally, an additional set of difficult tracers is required, with one more backward and forward phase.

Working backwards in time from the date of the MRCA, the proportion of CAs in the population increases gradually until, eventually, everyone is either a CA of all of the living sims or is the ancestor of none of them, and is therefore *extinct*. Thus, a point will be reached at which 100% of the non-extinct sims are CAs. In other words, everyone living at the end of the similation will share the same set of ancestors who lived at that point. This is what we refer to as the *identical ancestors*, or IA, point. Although this successive refinement approach does find the true MRCA, it does not necessarily find the true IA point, only the point at which everyone is a potential CA. However, the IA point that appears in the same backward phase in which the MRCA is found is nearly always the correct one, or quite close to it. This can be verified with additional refinement steps, which generally lead to no further change in the IA point.

The models were simulated and analyzed on 2.7 GHz Pentium 4 workstations with 1 to 2 GB of RAM. Actually running the simulation requires about three hours, while the process of finding the common ancestors requires five to ten hours.

References

Adams, J. W., & Kasakoff, A. B. (1976). Factors underlying endogamous group size. In C. A. Smith (Ed.), *Regional analysis, vol. 2, social systems* (pp. 149–173). New York: Academic.

Arutiunov, S. A., & Fitzhugh, W. W. (1988). Prehistory of Siberia and the Bering Sea. In W. W. Fitzhugh & A. Crowell (Eds.), *Crossroads of continents: Cultures of Siberia and Alaska* (pp. 117–129). Washington, D.C.: Smithsonian Institution Press.

Biraben, J.-N. (1980). *An essay concerning mankind's evolution, population.* Selected papers.

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The history and geography of human genes (abridged).* Princeton, New Jersey: Princeton University Press.

Diamond, J. (1997). *Guns, germs, and steel: The fates of human societies.* New York: W. W. Norton & Company.

Fix, A. G. (1979). Anthropological genetics of small populations. *Annual Review of Anthropology*, *8*, 207–230.

Grønnow, B., & Pind, J. (1996). *The Paleo-Eskimo cultures of Greenland: New perspectives in Greenlandic archaeology, Papers from a symposium at the Institute of Archaeology and Ethnology, University of Copenhagen, 21-24 may, 1992.* Danish Polar Center Publications No. 1.

Jorde, L. B. (1980). The genetic structure of subdivided populations: A review. In J. H. Mielke & M. H. Crawford (Eds.), *Current developments in anthropological genetics: Vol. 1* (pp. 135–208). New York: Plenum Press.

Pletcher, S. (1999). Model fitting and hypothesis testing for age-specific mortality data. *Journal of Evolutionary Biology*, *12*, 430–439.

Stannard, D. E. (1992). *American holocaust: Columbus and the conquest of the new world.* New York: Oxford University Press.

Terrell, J. E., Hunt, T. L., & Gosden, C. (1997). The dimensions of social life in the pacific: Human diversity and the myth of the primitive isolate. *Current Anthropology*, *38*, 155–195.

U.S. National Office of Vital Statistics. (1956). *Death rates by age, race, and sex, United States, 1900–1953, Vital Statistics—Special reports vol 43, no 1.* Washington, D.C.: U.S. Government Printing Office.