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.

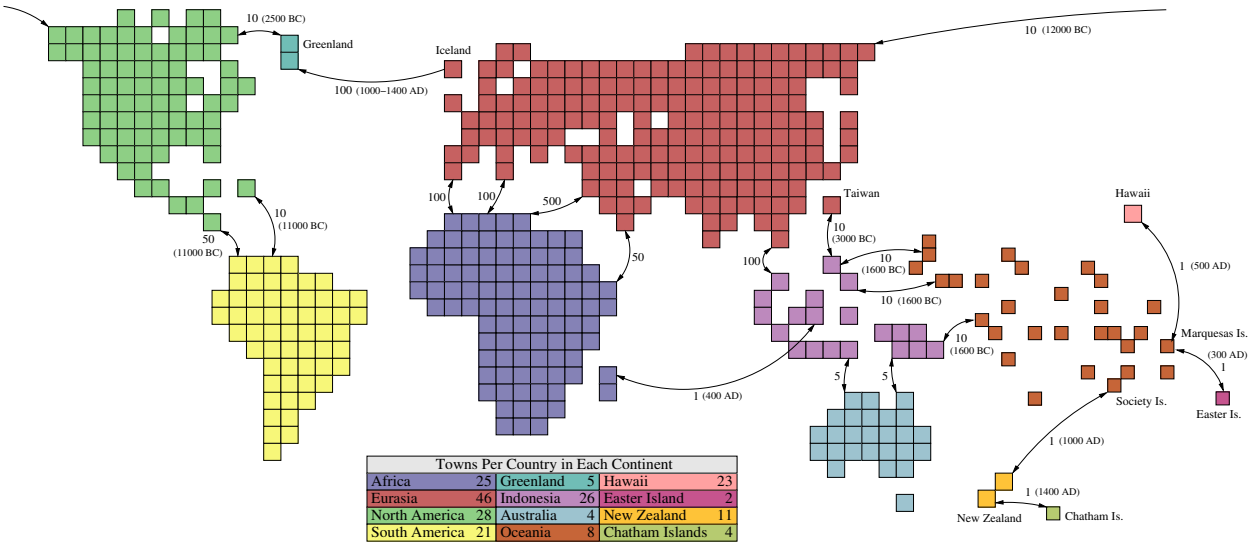
Not all humans confine themselves to a single location throughout their lives and a critical factor in the model is the rate at which people migrate to different places in the world. Although it seems likely that many people, and perhaps the vast majority historically, live out their lives close to where 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.