Fair use notice

On the Common Ancestors of All Living Humans

Douglas L. T. Rohde
Massachusetts Institute of Technology

November 11, 2003

Abstract

Questions concerning the common ancestors of all present-day humans have received considerable attention of late in both the scientific and lay communities. Principally, this attention has focused on 'Mitochondrial Eve,' defined to be the woman who lies at the confluence of our maternal ancestry lines, and who is believed to have lived 100,000–200,000 years ago. More recent attention has been given to our common paternal ancestor, 'Y Chromosome Adam,' who may have lived 35,000–89,000 years ago. However, if we consider not just our all-female and all-male lines, but our ancestors along all parental lines, it turns out that everyone on earth may share a common ancestor who is remarkably recent.

This study introduces a large-scale, detailed computer model of recent human history which suggests that the common ancestor of everyone alive today very likely lived between 2,000 and 5,000 years ago. Furthermore, the model indicates that nearly everyone living a few thousand years prior to that time is either the ancestor of no one or of all living humans.

1 Introduction

Advances in genetics have sparked interest in our common ancestors, the individuals from which all present-day humans descend. Initial interest focused on the topic of 'Mitochondrial Eve,' who is defined to be the most recent female ancestor from whom all individuals descend along strictly maternal lines (Cann, Stoneking, & Wilson, 1987; Vigilant, Stoneking, Harpending, Hawkes, & Wilson, 1991). An approximate date for Mitochondrial Eve of 100,000 to 200,000 years ago has been estimated based on the successive mutations in mitochondrial DNA, which are passed down from mother to child. A similar analysis can be performed on the strictly paternal lines of succession, using the Y chromosome, which is passed down from father to son, to determine the approximate date of 'Y Chromosome Adam.' This date was originally estimated very loosely to fall between 27,000 and 270,000 years ago (Dorit, Akashi, & Gilbert, 1995), a range that was more recently narrowed to 35,000–89,000 years ago (Ke et al., 2001).

Nevertheless, an individual's strictly maternal and strictly paternal lines are just two of a vast number of possible paths back through his or her ancestors. What if we adopt a more common-sense notion of ancestry that includes ancestors reachable along any path of succession, using both mothers and fathers? It seems likely that our most recent common ancestor (MRCA) under this broader definition will be much more recent than either Mitochondrial Eve or Y Chromosome Adam. Unfortunately, the age of our MRCA cannot as easily be estimated on the basis of genetic information because the relevant genes are not passed from parent to child with only occasional mutations but are, rather, the product of recombination. As a result of recombination, a given gene may not pass from parent to child. In fact, an individual's DNA may retain none of the genes specific to a particular ancestor who lived many generations in the past. These additional complications make accurately dating the MRCA or reconstructing other details of population history on the basis of our genes extremely difficult, if not impossible (Hey & Machado, 2003).

However, alternative methods may be able to answer this question. Some researchers have produced estimates of the age of our MRCA by means of the theoretical analysis of mathematical models. Building on work by Kämmerle (1991) and Möhle (1994), Chang (1999) analyzed a model that assumes a fixed-size population, with discrete and non-overlapping generations, and random mating. That is, each child is the product of two parents, randomly selected from all members of the previous generation. Chang showed that, in this model, the mixing of genes occurs quite rapidly. In fact, the number of generations back to the MRCA is expected to be about $\log_2$ of the population size. With a population of 6 billion people, this model predicts that the MRCA is likely to occur in just over 32 generations, or 800–975 years. This suggests that our all-paths MRCA may be exceptionally recent.

But Chang was well aware of the limitations of the simple model he analyzed. "What are the significance

*Work in progress. Do not cite.

of these results? An application to the world population of humans would be an obvious misuse... An important source of inapplicability of the model to this situation is the obvious non-random nature of mating in the history of mankind." (pg. 1005) There are many factors that limit the randomness of human mating. First of all, clearly, are sex differences, but Chang did address this, noting that adding distinct sexes to the model would not cause a substantial change in the estimate. Another factor is marriage. Once a couple has one child, they are likely to remain together as they produce more children. Moreover, broader sociological and geographic factors may have still more profound effects. In short, although we are becoming increasingly panmictic, humans groups have tended towards a high rate of endogamy, finding mates almost exclusively within the local population and social class, only occasionally transcending barriers of geography, language, race, and culture.

It seems likely that these restrictions on the randomness of human mating may dramatically decrease the rate of ancestral mixing in the model. As a result, the true date of the MRCA could be thousands or tens of thousands of years ago, rather than just hundreds. Thus, an obvious next step is to test this possibility by expanding the model to include some or all of these constraints. Unfortunately, conducting a theoretical analysis of a more complicated mathematical model would be very difficult. An alternative approach is to implement a computer simulation. The principal advantage of a computer simulation is that it can be arbitrarily complex. However, even given the speed of today's computer, efficiently simulating the ancestral history of a population whose size is even close to the scale of humanity is non-trivial. Furthermore, because a non-random model will necessarily involve numerous parameters that cannot be adequately constrained by available data, the simulation must typically be run many times to explore the consequences of various parameter settings.

This study involves the implementation and analysis of several large-scale computer models of recent human history. The models simulate individual human lives, including life span, birth rate, choice of mates, and migration, and the data they produce is analyzed to obtain more accurate estimates of the date of our most recent common ancestor. Given what seem to be reasonable parameter choices, the final, most detailed model presented here predicts that our most recent common ancestor probably lived between 2000 and 5000 years ago and that nearly everyone alive prior to a few thousand years before that are the ancestors of either no one or of everyone alive today.

1.1 Modeling human genealogy

Mathematical models of human genealogy must be quite simple if their analysis is to be possible. Most, like Chang's, are some variant of a Wright-Fisher model, with discrete generations and parents selected at random from the preceding generation (Nordborg, 2001). Because computer simulations are tested empirically rather than through theoretical analysis, they are not subject to such constraints.

However, there are practical limits to the complexity of a computer simulation. One is the matter of computational efficiency. A model cannot be so complex that running it requires an unreasonable amount of time or space. A more significant limitation, from a scientific perspective, must be placed on the number of free parameters in the model. Ideally, for the results of a model to be reliable, any free parameters should be constrained by historical data, such as statistics on actual birth or migration rates. However, much of the relevant data for the current models concern events occurring thousands of years ago and cannot be obtained with any accuracy. In this case, the parameters must be varied within the range of plausible values to obtain bounds on the model's predictions. A model with too many free parameters, especially ones unconstrained by empirical data, will have reduced power and will be difficult to study. Therefore, a good model must be complex enough to include relevant factors, but not overly burdened by irrelevant ones.

This study explores a progression of three models. The first extends Chang's results to a world consisting of five more or less panmictic islands, or continents, with only occasional migrants between any pair of continents. The second model, discussed in Section 3 arranges the islands in a graph that roughly reflects the topology of the major world continents. The final model, discussed in Section 4, is a more detailed simulation of the actual world, with migration routes and dates based on historical data or prehistoric estimates.

2 Model A: Fully-connected continents

The first model, A, is quite abstract but incorporates several levels of detail beyond those found in most Wright-Fisher models. The model typically starts between 5000 and 20000 BC and runs to the present day, which is taken to be the year 2000 AD. As the model runs, it simulates important details in the lives of individual people, known as *sims*, including their lifespans, possible migrations, choice of mate, and production of offspring. As the model runs, it records this information in a series of large computer files. A second program, discussed in Section 2.2, traces ancestral lines through this data to find the common ancestors.

2.1 Details of the model

2.1.1 Life span

The present models do 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 was set to 100, as it seems highly unlikely that anyone would live, let alone father children, beyond that age. The age of sexual maturity was 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 were 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):

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

In this equation, β is the *death rate*. A higher death rate results in shorter life spans on average, although the result 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.

The percentage of males born into the population was set at 50%. It is true that the actual percentage of males and females reaching adulthood may differ somewhat due to infanticide coupled with the fact that a slightly higher percentage of newborns are male than female (Davis, Gottlieb, & Stampnitzky, 1998). But this probably does not have much bearing on the results of the model. And while it is true that women tend to live longer, the life span of women past child-bearing age is also not relevant to the outcome of the model. Therefore, for simplicity, the life spans of males and females were generated using the same distribution.

2.1.2 Migration

The models are organized into three structural levels: continents, countries, and towns. The continents represent physically separated land masses that are likely to have very low rates of inter-migration. Europe and Asia are contiguous, with no substantial geographical barrier to migration, so they will be considered a single continent, along with Africa, North America, and South America. Indonesia, Australia, and Oceania are, taken together, somewhat more difficult to model, as there is clearly substantial internal structure. For the purposes of the first two models, they will be considered a single continent, but will be dealt with more appropriately in the third model.

The models' 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. Countries are, in turn, 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 of people have conquered or colonized new areas.

In terms of realism, it would certainly be desirable to distinguish between these and other specific types of migration in the model. However, doing so would introduce many new parameters, for which we are unlikely to find sufficient data. Therefore, the model uses a simplified migration system, in which each person can move only once in his or her life. Each 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 potential mates.

As is the case in human mating patterns (Fix, 1979), the rate of exogamy decreases substantially with larger group size in the models. 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

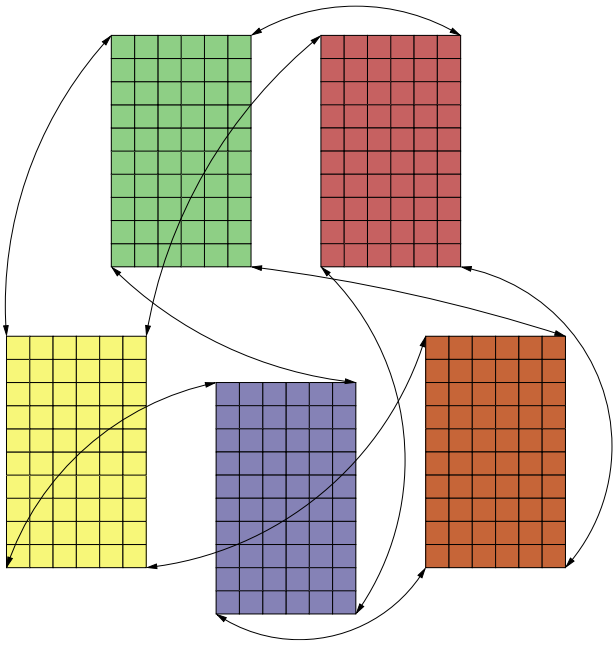

Figure 1: Model A: A fully connected structured model. Each continent consists of 60 countries with 80 towns per country.

leave the town of their birth for another town within the same country. In the current models it 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). The countries within a continent are arranged in a grid and locality also plays a role in inter-country migration. In Model A, shown in Figure 1, all continents contain 60 countries in a 6 by 10 rectangle. In the first two models, inter-country migration involves a two-tiered system. The majority of the sims leaving a country travel to a neighboring country (including diagonal neighbors). The remainder travel to a randomly chosen country within the continent, including the neighboring countries. The fraction of sims who choose randomly is governed by the *NonLocalCountryProb* parameter. A value of 0 means that all inter-country travel is to neighboring countries. A value of 100% means that travel to all countries in the continent is equally likely.

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. It is expressed in terms of migrants per generation, where a generation is taken to be 30 years, as that is the approximate average parental age. 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 long-distance migrants might, for example, be merchants. A *NonLocalPortProb* of 0 means that all migrants are born locally. A value of 100% confers no special advantage to the source country.

Migrants using a port arrive in a random town within the destination country. However, they then have the usual small chance of migrating to a new country within that continent. A sim can use at most one port in his or her lifetime.

Model A, depicted in Figure 1, consists of five continents arranged in a fully-connected graph. The five continents are each composed of 60 countries, with 80 towns per country, for a total of 24,000 towns. Each continent is connected to every other one, with ports lying at the corners. This is not meant to be an accurate depiction of the world by any means, but is an extension of the panmictic model to a simple form of structured model similar to those that have been studied previously (Nordborg, 2001). Our primary goal in studying such a model is to gain a better understanding of its sensitivity to the various parameters and to provide a baseline against which to compare models with more realistic geography.

2.1.3 Mating

Along with migration, the rate of ancestral mixing is also dependent on how mates are chosen and on the age distribution of the parents when children are born. In this respect, the model was implemented from the perspective of the mother. The program 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 the age of 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 2.0 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

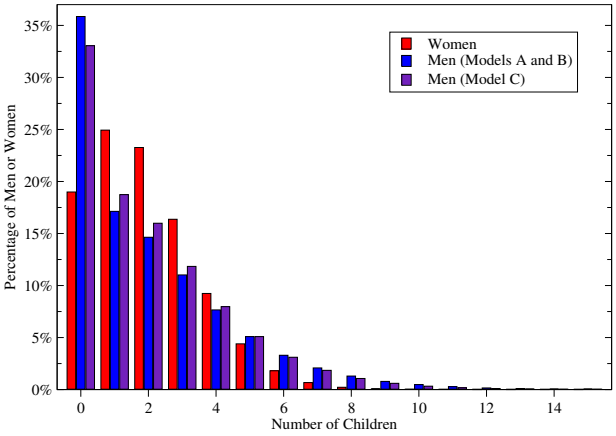

Figure 2: Distributions of the number of children per woman or man. Only children who reach adulthood are counted. Model C produces a slightly different distribution for the men.

same country.

The father of a woman's first child is chosen 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. 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.

Figure 2 shows the distribution of children per woman and man in the models. The distribution for women is essentially binomial, with 19% producing no children and only 2.8% producing more than 5 children. The middle bar shows the distribution for men in the first two models. It has greater variance than that for the women. Nearly 36% of men produce no adult children, while 8.6% produce more than 5 children. Thus, there is a higher percentage of men than women that produce no children or many children, but relatively fewer men who produce a moderate number of children.

For lack of a better term, we will refer to sims who have no living descendants as *extinct*. In other words, these are sims whose lineage has died out. Clearly, anyone who produces no children is extinct upon their death. But even those who produce some children may become extinct if all of those lines die out, which rarely occurs beyond the first few generations. In the model empirically, we find that 33.31% of females and 45.42% of males actually become extinct. These values are just slightly lower than we would predict based on the distributions in Figure 2. Thus, females have the reproductive advantage in terms of the likelihood their genes will survive for all future generations. In many cultures, sons are much preferred over daughters. However, unless one's sons are powerful enough to procure multiple wives, it is actually more advantageous, for the purpose of creating an enduring lineage, to produce daughters.

Table 1: World population estimates, in millions.

Year	Population
20000 BC	2
15000 BC	3
10000 BC	4
5000 BC	5
2000 BC	25
1000 BC	50

2.1.4 Population Growth

Because life span was not manipulated, the growth rate of the population was controlled by adjusting the average number of children per woman. As Chang's results suggest, the size of the population may be an important determiner of the date of the MRCA. Given fixed sizes, a larger population will tend to have a less recent MRCA. But a larger population will also tend to have a greater number of migrants, thus potentially leading to a more recent MRCA. The net effect of population size is, therefore, difficult to predict, but will be tested empirically.

Ideally, the model should be capable of simulating a full-size world population. However, due to the available disk space for recording the necessary data, the models were limited to a maximum population of 60 million sims at any one time. A natural population growth was simulated up to a point and then the population was capped. Table 1 shows the worldwide populations used in the first two models. These data are based on the estimates of McEvedy and Jones (1978), with the most ancient numbers extrapolated. Population size was regulated by adjusting the birth rate to achieve geometric growth between the given targets. In most of the model A and B simulations, the maximum population was 25 million, reached in the year 2000 B.C, and then maintained thereafter. Other simulations continued to a maximum level of 50 million.

The migration rates between towns and countries are expressed as a percentage of births. Therefore, as the pop-

ulation increases, the total number of migrants increases proportionally. In Models A and B, the rate of the ports was fixed to achieve a particular number of migrants per generation once the population had reached maximum size. However, prior to that point, proportionately fewer migrants would be using the ports.

2.1.5 Initialization

There is one remaining aspect of the model to be described, which is the method of initialization. Although some simulations were started in the year 20000 BC, others were started as recently as 5000 BC if a more recent start date would not interfere with the results. In order to get things going, we need some initial sims. 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.

2.2 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 the 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 the ith 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 in order, 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. This will be referred to as the *all common ancestors*, or ACA, point. Although this successive refinement approach does find the true MRCA, it does not necessarily find the true ACA point, only the point at which everyone is a potential CA. However, the ACA point that appears in the same backward phase that 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.

2.3 Results

A number of simulations were conducted with Model A under various parameter settings. Of principal interest is the date of appearance, working backwards, of the most recent common ancestor (MRCA date) and the date at which all of the non-extinct sims are common ancestors (ACA date). These dates will be measured in years before present (BP), where the present is taken to be 2000 AD. When we refer to the *MRCA time*, it is the length of time between the present and when the MRCA was last living. Therefore, a *longer MRCA time* means that the MRCA lived less recently.

For each simulation, at least three trials were performed and the results averaged. In general, the trials were quite consistent, with the MRCA and ACA dates having a standard deviation less than 10% of the mean, and often under 2%. The variance is larger for the ACA point and for the simulations with earlier dates.

The first simulation, referred to as A1, used fairly liberal parameters, as shown in the top row of Figure 3. The maximum population was 25 million, reached in the year 4000 BP. The *ChangeTownProb* was 20%, meaning that 80% of the sims marry within their birth town. The *ChangeCountryProb* was set to 0.1%, so about 1 in 1000 sims leave their home country, which may seem somewhat high. But to put this in perspective, with a population of 25 million, there are about 48,000 people born in each country per generation. So, on average, a *ChangeCountryProb* of 0.1% will result in 48 people leaving a country every 30 years, which certainly does not seem excessive.

More liberal is the fact that, in simulation A1, there are no locality constraints on inter-country migration or the use of ports. Migrants have an equal chance of traveling to any country within the continent and can use a port from anywhere within the continent. The *PortRate* was set to 10 migrants per generation, in each direction, which is about 1 migrant every three years.

The bars on the right half of Figure 3 depict the common ancestry timelines. The dates are in years before present, with the present located on the right. In the white region, there are not yet any common ancestors. Moving backwards in time, the MRCA is found at the border between the white and gray bars, in 1720 BP in this case. In the gray region there is an increasing number of common ancestors until we reach the ACA point, at 2880 BP. In the black region, all of the sims are either extinct or common ancestors of the living sims. Thus, in this case, there is a fairly rapid transition between the appearance of the first CA and the point at which everyone alive today shares the same set of ancestors.

The actual rate of this transition for one of the A1 simulations is shown in the red line in Figure 4. The small red marker at the bottom right of the figure denotes where the MRCA occurred, at 1800 BP. From that point on, the percentage of CAs in the population grows slowly at first, reaching 1% in 1940 BP, and then very rapidly, reaching 50% in 2160 BP and 99% in 2400 BP. Then there is a relatively long period during which most, but not all, of the sims are either CAs or extinct. It takes another 570 years to reach the true ACA point, denoted by the red marker at the top of the figure.

It is likely that the notion of a relatively recent ACA point may lead to some confusion. If we consider only ancestors who lived prior to the ACA point, a Japanese and a Norwegian today share the exact same set of ancestors. At first glance this seems patently ridiculous. Certainly the Japanese and Norwegian have quite different genotypes due to very different ancestry. The confusing fact is that

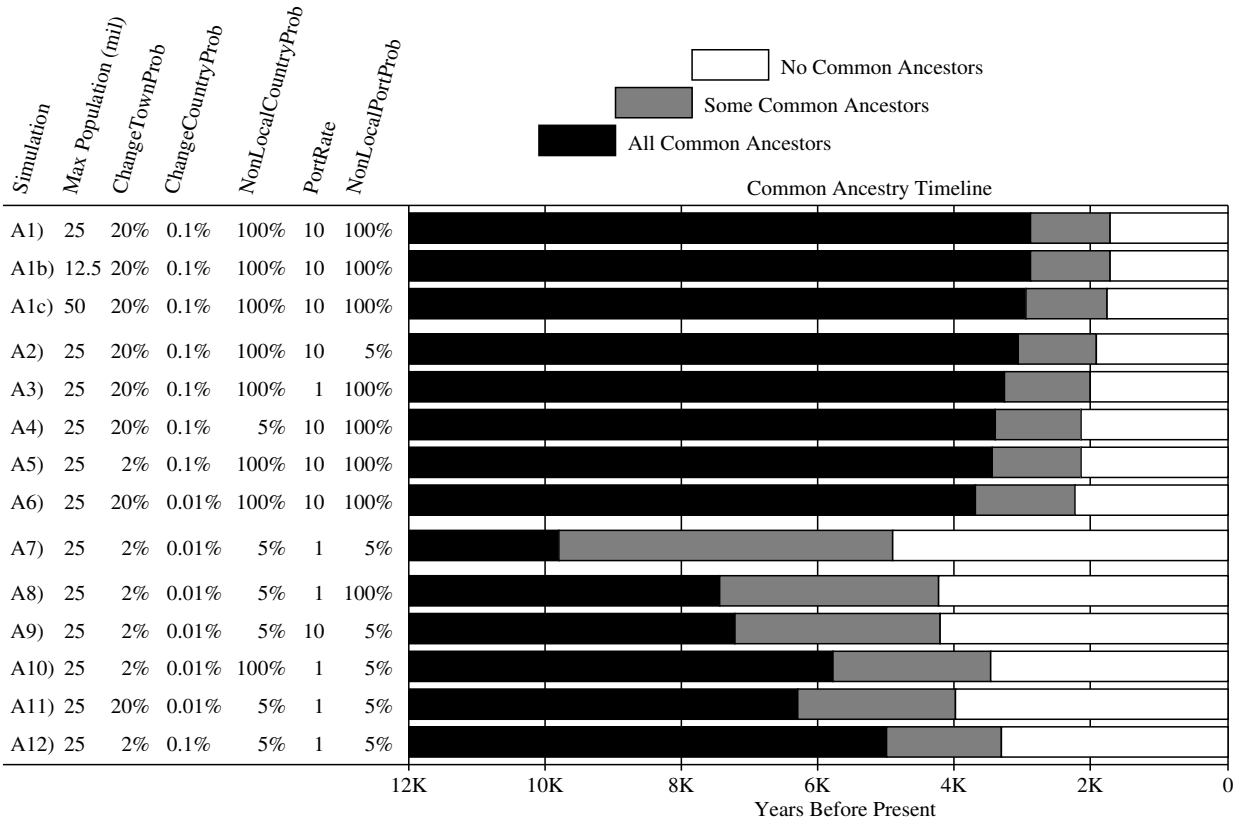

Figure 3: Results of the Model A simulations with various parameter settings. The timelines are in years before present, with the present located on the right. In the white region there are not yet any common ancestors. The border between the white and gray regions is the MRCA point, when the MRCA died. The border between the gray and black regions is the ACA point.

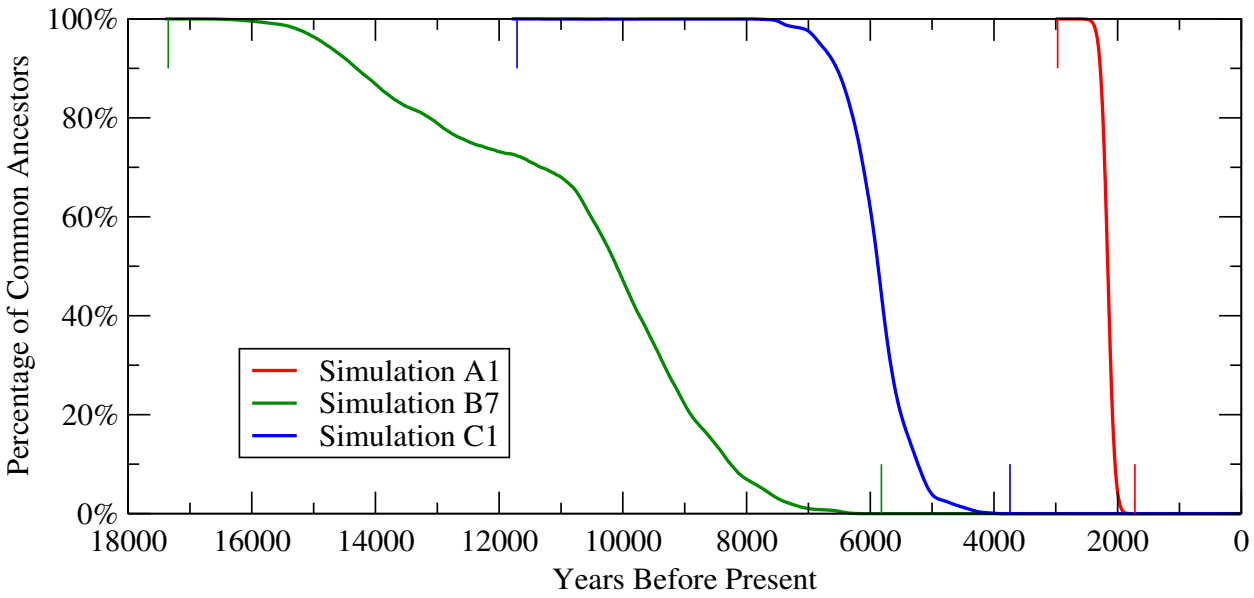

Figure 4: Percentage of non-extinct sims who are common ancestors of everyone living in the year 2000 for three representative trials. The vertical markers show the dates at which the curves reach 0% and 100%.

both of these statements are true. Although the Japanese and Norwegian have the same set of ancient ancestors, they did not receive an equal hereditary contribution from each of those ancestors. The Japanese owes a small proportion of his genetic makeup to people living in northern Europe several thousand years ago, and a large proportion to people living in and around Japan, while the opposite is true of the Norwegian. Thus, their ancestry does differ considerably, but only in distribution. This point will be examined further in Section 5.3.

Simulations A1b and A1c manipulate the maximum population size, keeping the final port rate the same in terms of migrants per generation. Halving the population (A1b) causes almost no change, while doubling it (A1c) causes a very slight, possibly non-significant, increase in MRCA and ACA time. As we'll see in simulation B7c, a larger population can actually lead to more recent dates under different conditions. The larger population has little direct effect on the ancestry coalescence time because it occurs after the MRCA lived. The most important determiner of the rate of spread of a lineage is not the absolute number of people with the lineage but the *percentage* of people. As the population uniformly grows or shrinks over time, this percentage is not affected. The only point at which the total population plays a role is at the time of the original ancestor. At that point, the percentage of the population represented by that ancestor is indeed a function of the population size. Thus, a larger population at the time that an ancestor lived would result in a longer delay for that person to become a CA. But a population that grows uniformly once the ancestor has died does not result in a similar delay. Larger populations do, however, tend to have more migrants across the most difficult barriers, resulting in a faster spread of lineage. In the case of simulations A1–A1c, these effects are either minimal or counteracting.

Simulations A2–A6 were similar to A1, but each manipulated a single parameter, applying a more conservative value to test the sensitivity of the model to that parameter. Simulation A2 lowered the *NonLocalPortProb* from 100% to 5%, so most users of a port must be born in its source country. The effect of this change is quite small. Relative to simulation A1, there was a 11.4% increase in the MRCA time and a 6.2% increase in the ACA time. These percent changes in MRCA date and ACA date are shown in the left-most bars in Figures 5 and 6, respectively.

Simulation A3 lowered the *PortRate* from 10 sims per generation to just one per generation. This resulted in a 17.2% increase in the MRCA time and a 12.9% increase in the ACA time. Thus, the model is not tremendously sensitive to migration rate by itself. Once a lineage has spread throughout most or all of a continent, it only takes a single non-extinct migrant to spread that lineage to an-

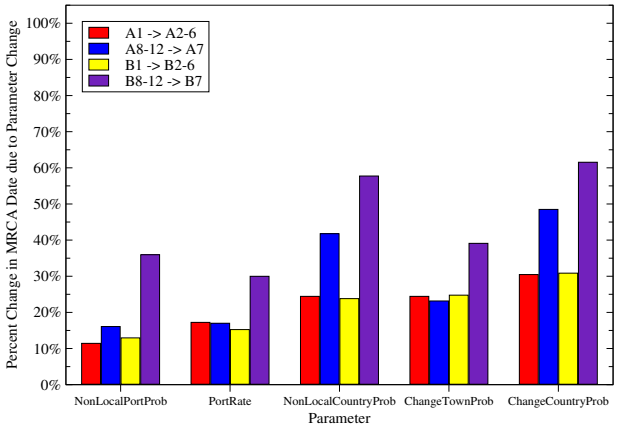

Figure 5: The percent change in the MRCA date resulting from various parameter changes for Models A and B. The red bars indicate the percent change from Simulation A1 to either simulation A2, A3, A4, A5, or A6, depending on the variable in question. The blue bars indicate the percent change from the less conservative simulations A8–A12 to the more conservative A7.

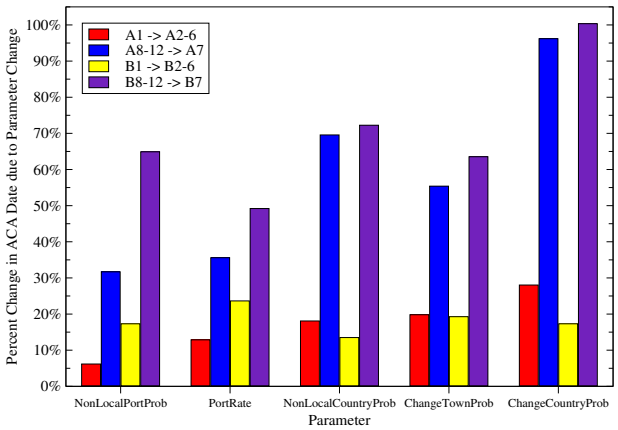

Figure 6: The percent change in the ACA date resulting from various parameter changes for Models A and B.

other continent and even very low migration rates may result in only short-term delays.

Simulation A4 lowered the *NonLocalCountryProb* from 100% to 5% causing most migration between countries to be local. This reduces the rate of admixture within continents, resulting in a 24.4% increase in MRCA time and an 18.1% increase in ACA time. Simulation A5 reduced the admixture rate within countries by lowering the *ChangeTownProb* parameter from 20% to 2%. This has a similar effect on the overall dates, also increasing the MRCA time by 24.4% and increasing the ACA time by 19.8%. Finally, simulation A6 reduced the *ChangeCountryProb* from 0.1% (1 in 1,000) to 0.01% (1 in 10,000). This has the greatest effect of the single-parameter manipulations, increasing the MRCA and ACA times by 30.5% and 28.0%, respectively.

If these five parameter changes have independent effects, we might expect the net effect of combining all of them to be either the sum of their independent additive effects or the product of their multiplicative effects. If the effects were additive, it would result in predicted MRCA and ACA dates of 3570 BP and 5355 BP, respectively. If the effects were multiplicative, the predicted dates would be 4533 BP and 6241 BP, respectively. The actual dates of the combined parameter changes from Simulation A7 are 4910 BP and 9790 BP. Thus, the effects of the parameters appear to be greater than their independent additive or multiplicative combination and we might conclude that there is interaction between them. This is particularly true for the ACA date, which experiences a greater change (240% relative to simulation A1) than does the MRCA date (186%). However, we do not yet know the nature of this interaction.

Simulations A8–A12 start with the same parameter values as A7, but change each of the variables back to its less-conservative setting. The point is to test the sensitivity of the model to each parameter in this new part of the space. The sensitivity is measured as the percent change in MRCA or ACA from the less conservative simulation, A8 for example, to the more conservative A7. If a parameter is acting independently and its effects are multiplicative, we should expect to see the same percentage change in the blue bars in Figures 5 and 6 as we saw in the red bars. If the blue bars are higher, it indicates that the model is more sensitive to the parameter when the other parameters are more conservative, suggesting that the parameters are interacting.

In terms of MRCA, the *PortRate*, and the *ChangeTownProb* appear to be acting independently of the other parameters. However, the *NonLocalCountryProb*, the *ChangeCountryProb*, and to some extent the *NonLocalPortProb* have greater effects in simulations A7–A12. This indicates that these parameters are interacting, probably with one another. These parameters all affect the rate at which lineage can spread long distances across continents. Lineage can spread fairly rapidly if either the *ChangeCountryProb* is high, meaning there are more inter-country migrants or the *NonLocalCountryProb* is high, meaning that there may be only a few migrants but they are more likely to travel long distances. When there are both few migrants and they tend to move short distances, there is a much greater resulting effect on the MRCA date.

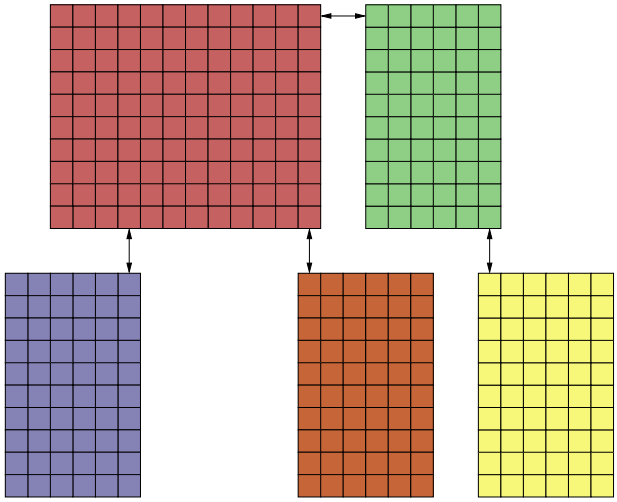

Figure 7: Model B. A highly simplified world map.

Interestingly, the same does not hold true for the ACA date, shown in Figure 6. In fact, all of the parameters seem to interact in determining it. As a result, the ACA date becomes increasingly sensitive and the ratio between it and the MRCA date increases when all of the parameters are assigned more conservative values.

3 Model B: Coarse real-world topology

The fully-connected world of Model A was an interesting forum to experiment with the parameters of the model because of its resemblance to more traditional structured coalescence models. However, it clearly bears little resemblance to the real world. Model B, therefore, takes a small step towards a more realistic model of the world, using the map shown in Figure 7. The continents are intended to resemble, clockwise from the lower left, Africa, Eurasia, North America, South America, and Australia/Oceania. The continents are internally the same as in Model A, except that Eurasia is twice as wide as the others. There are only four bidirectional ports in this model, with South America connected to North America and the other continents connected to Eurasia. We are interested primarily in the effect this structure will have on the spread of lineages

throughout the world.

As in the first model, we assume that the distribution of sims is initially uniform throughout the world and that the port rate is also uniform. However, because there are so few ports and they are intended to resemble fairly major intercontinental passages, the ports were given migration rates 10 times higher than those in the first model. Otherwise, a parallel series of simulations was conducted. The summaries and results of these are shown in Figure 8 and the associated percent changes in MRCA and ACA dates are shown along with those for Model A in Figures 5 and 6. Note that the scale for the timelines in Figure 8 is different than that for Figure 3.

Following the increase in migration rate, simulations B1–6 result in very similar MRCA dates as the corresponding A1–6 simulations, averaging less than a 2% increase in MRCA time. Likewise, the individual parameter changes have similar effects on the MRCA time in simulations A2-6, as shown in the comparison between the red and yellow bars in Figure 5. However, the change in architecture has a greater effect on the ACA time, which averages 25.8% longer for simulations B1–6 than for A1–6.

In order for an MRCA to appear, there must be a single person whose lineage spreads throughout the world. Because it can take quite a while for a lineage to cross a continent or to travel from one continent to another, someone's lineage will be most likely to fill the entire world rapidly if that person lives near the center of the world, in a graph-theoretic sense. In the case of Model B, that center is in northeastern Eurasia, where the tip of South America is two ports and the height of two continents away and the tips of Africa and Oceania are one port and two continents away. This situation does not differ too much from Model A, in which all people were no more than one port and two continents apart. As Joseph Chang has noted in some recent work, the MRCA date in a graphical model is essentially proportional to the radius of the graph (Rohde, Olson, & Chang, in press).

On the other hand, in order for an ACA point to appear, the lineage of everyone alive at that time must have either died out or spread throughout the world. The time for this to occur is limited by the time required for a lineage to travel between the two most distant parts of the world, which is governed by the diameter of the graph. For the fully-connected Model A, the diameter is just a bit larger than the radius, but for Model B the diameter from the tip of Africa to the tip of South America is three ports and four continents, or nearly twice the radius. As a result, Model B has a longer ACA date and is more sensitive to parameter changes, particularly changes in the *PortRate* and *NonLocalPortProb*. The MRCA date for the most conservative simulation, B7, was 18.3% earlier than that for A7, but the ACA date was 74.9% earlier.

Model B also experiences greater interactions between the parameters. In this case, all of the parameters appear to interact, in their effect on both the MRCA and the ACA times. This is true even of the *NonLocalPortProb*, *PortRate*, and *ChangeTownProb* that did not interact strongly in their effect on the MRCA date in Model A.

Figure 4 shows the percentage of non-extinct sims who are common ancestors of the living sims as a function of time. The left-most, green, curve is for one of the B7 simulations. Note that, unlike simulation A1, the period between the MRCA and the ACA point in this case is quite long and the transition is not smooth. This is due both to the much lower migration rates in this model and to the sparsely-connected architecture. Moving backwards in time, the percentage of CAs increases smoothly for a time and then begins to level off just under 70%. Presumably, this results from the delay in the common-ancestry relationship reaching North America from Eurasia. There is another slight leveling at around 80%, which may result from common ancestors starting to appear in South America. Note, also, that there is about a 1700-year difference between the time at which 99% of the people are CAs and the true ACA point.

Simulations B7b and B7c are similar to B7 but vary the maximum population size. Recall that varying the population limit had little effect on the more liberal simulation A1. In this case, however, reducing the population to 12.5 million increases the MRCA and ACA dates, while doubling the population reduces the MRCA by 18.7% and the ACA by 3.4%, because larger populations have more intra-continental migrants. Therefore, although we were not able to simulate a full-size world population, it seems that the use of a reduced population has made these models more, rather than less, conservative, resulting in MRCA and ACA dates that are somewhat older than they should be.

4 Model C: Detailed geography and migration

The first two models enabled us to investigate some properties of common ancestry under relatively simple conditions. We found that, even with different architectures and quite widely varying parameters, the date of the MRCA is relatively stable, roughly falling between 2000 and 6000 BP, while the date of the ACA point is more variable, possibly extending as far back as 18000 BP. However, those models were intentionally abstract and bear little resemblance to the real world. Model A was excessively liberal in that the continents were fully interconnected, which was not the case, in terms of migration patterns, until quite recently. On the other hand, Model B was possibly overly conservative in that it allowed only four intercontinental

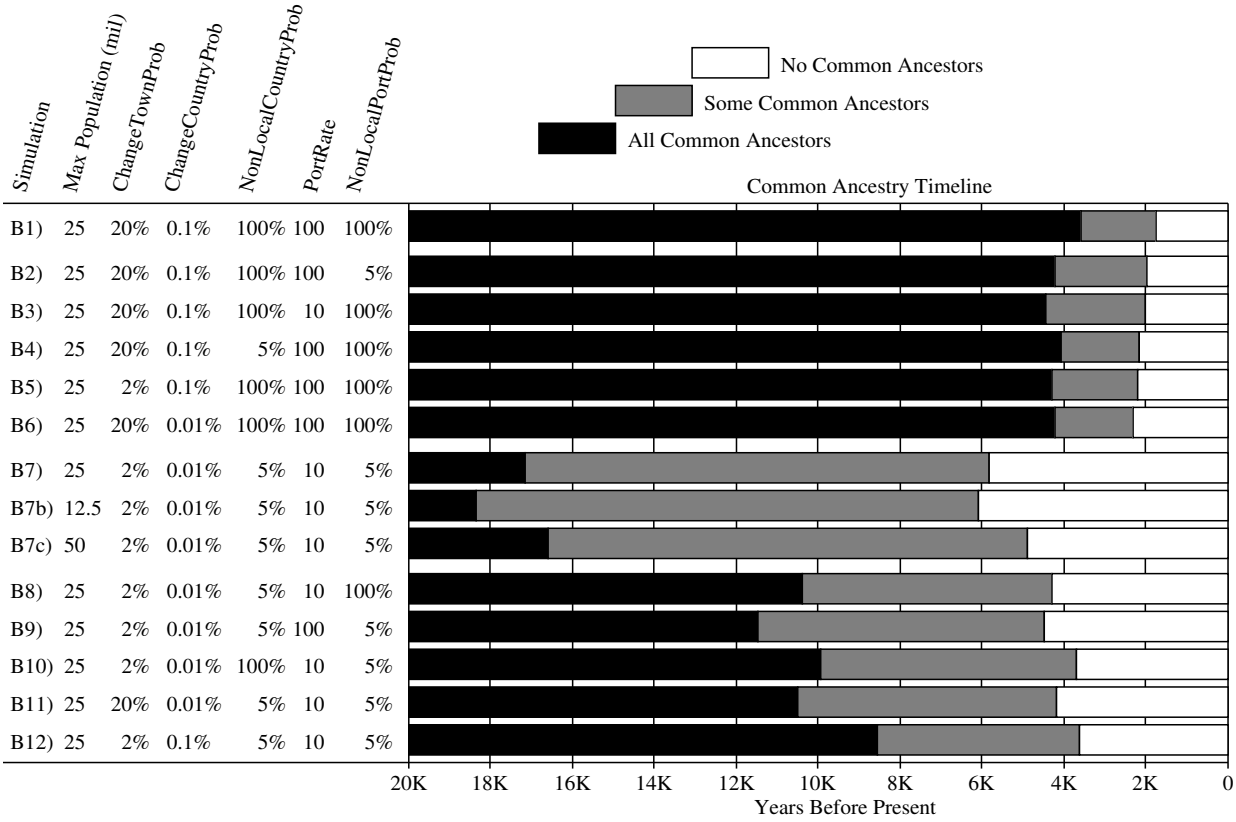

Figure 8: Results of the Model B simulations.

ports with between 10 and 100 migrants per generation across them. Certainly the interchange between Africa and Eurasia has been higher than that, and there are potentially other routes of passage between the continents, such as migration from Borneo to Madagascar and potential contact between Greenland Inuit and Vikings and between South America and Polynesia.

Another limitation of the first two models is that, aside from population growth, they are static. The initial population, even at 20000 BP, was assumed to be spread evenly throughout the world, which certainly was not the case. In reality, population expansion into the Americas and the Pacific probably occurred later and in successive waves, while long distance migration rates have increased with time, particularly with the advent of widespread oceanic navigation and, more recently, flight. Model B did not provide a reasonable depiction of Oceania, treating it as a single, contiguous continent, presumably in conjunction with Indonesia, Australia, and New Zealand. This leaves open the possibility that the relatively low migration rates throughout the Pacific have had a profound effect on the world's common ancestry. Because of the observed interactions between the parameters of the first two models, it remains unclear how the results would translate to a more realistic model with heterogeneous geography, population density, and migration routes.

Therefore, Model C was created in an attempt to provide a very detailed and flexible representation of the world. This will enable us to produce more accurate estimates of where and when our MRCA lived and also to test specific scenarios, such as the potential effect of hypothesized contacts between South America and Oceania or between Vikings and indigenous people of Newfoundland.

4.1 Details of the model

The world of Model C is depicted in Figure 9, with each continent, or independent island, rendered in a different color. The continents are no longer rectangles but are based on projections of real world geography, with each country representing approximately 119,000 square miles.[1] An exception to this is Oceania, where the countries are intended to resemble the major island groups and are typically much smaller in terms of both area and population. Clearly, not all of the "continents" in this model are actual continents, but the name is retained for continuity. The distances shown between continents in Figure 9 are arbitrary, the only important factor being the number and migration rate of the ports connecting them.

[1] The area of sparsely-inhabited northern Siberia has been reduced in the model.

4.1.1 Migration

The Model C simulations begin in the year 20000 BC, but the populated areas at that time 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. The ports are shown as arrows in Figure 9, labeled with their migration rates, in sims per generation. Between Africa and Eurasia, there are ports between 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 a rate of just 5 s/g.

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 migration 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.

In the previous models, immigrants using a port could settle in any random town in the destination country. As a result, immigrants were immediately assimilated into the host community. It is more often the case in modern times, and presumably throughout history, that immigrants 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 will tend to cluster, though they will not be entirely segregated.

Aside from those already mentioned, the remainder of the ports in the model only open at particular points in time, indicated in Figure 9 by the dates in parentheses. Unlike the previous models, in which the entire world was inhabited from the start, we must now deal with the problem of the initial colonization of new territory. The main

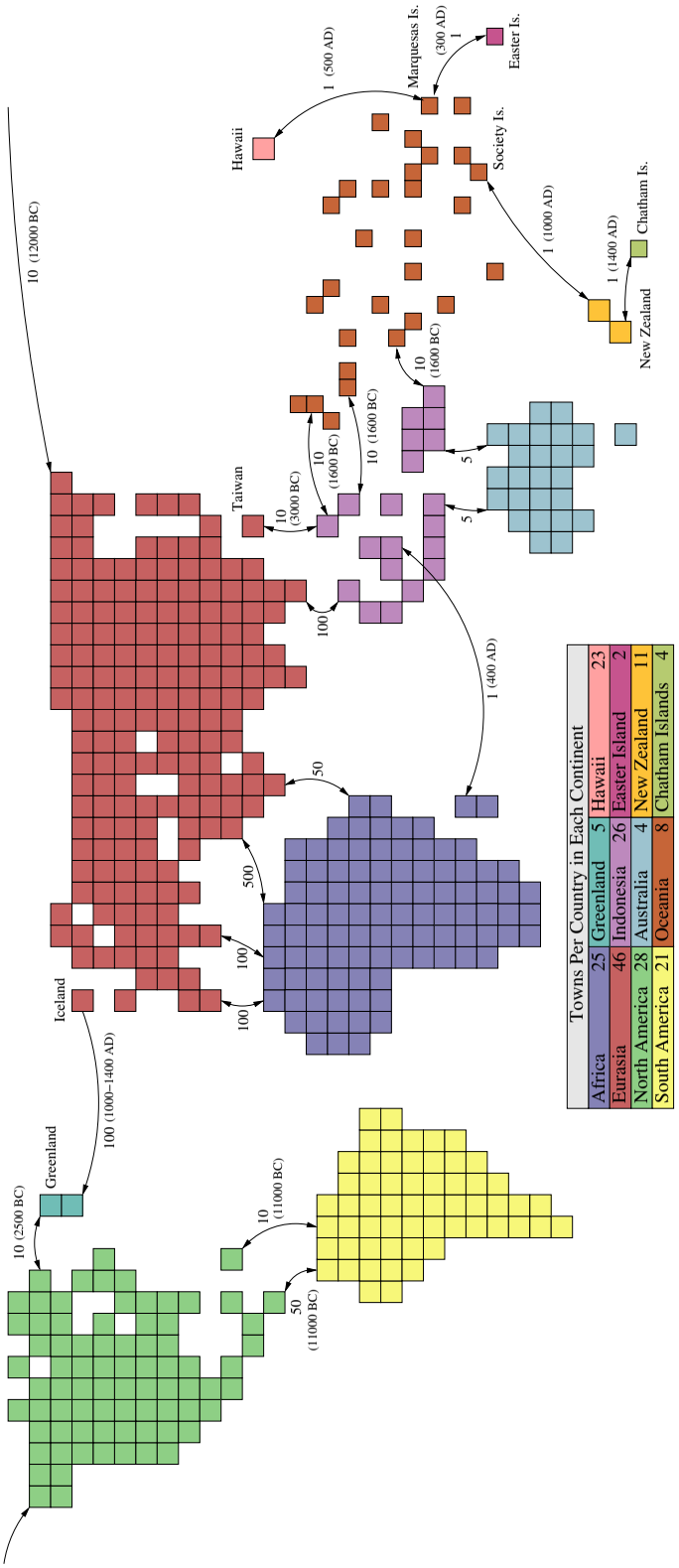

Figure 9: Model C. Arrows denote ports and the adjacent numbers are their steady migration rates, in sims per generation. If given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.

issue this raises is how pioneers are to gain a foothold. A basic assumption of the model is that sims act independently in their migration decisions. They cannot organize a sustainable colony in advance, and, because the rate of migration to new countries is typically very low, individual migrants will often 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 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.

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 colonies 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 4.1.2

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, an 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. 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.[2]

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 just 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 et al., 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 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

[2]The model is somewhat inaccurate in that the smaller islands of "Near Oceania", west of and including the Solomons, are not colonized until 1600 B.C., although they are believed to have been inhabited by Pleistocene-era people for several thousand years prior to that (Terrell, Hunt, & Gosden, 1997). It is unlikely that this has an effect on the results because there is believed to have been significant contact between the Polynesians and these earlier inhabitants.

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 Figure 9, 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 4.1.2, 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. The initial value of this parameter ranges from 2% to 20% in the models tested. In most of the simulations, it 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* also increases to simulate greater mobility, doubling in the years 1500, 1750, and 1900.

In addition to the more complex geography and intercontinental ports, a few other details were changed in producing Model C. The method of selecting fathers in the first two models may not have sufficiently taken into account the preference of women to marry single men. As a result, the process was overly unfair, resulting in too many men with no children or many children and not enough with a few children. This, and some computational considerations, led to a new method of choosing fathers that results in a slightly more fair distribution of children per man, as shown in the purple bars in Figure 2. As a result, the percentage of women who will become extinct decreases to 32.45% and the percentage of extinct men decreases to 42.92%.

The process of inter-country migration in the first two models was more seriously flawed. They used a two-tiered system, with one rate of migration to neighboring countries and a second rate to all other countries, with the difference governed by the *NonLocalCountryProb* parameter. As a result, it was just as likely for someone to migrate two countries away as it was for them to migrate clear across the continent. Instead, the new model uses a distance-based approach. The overall probability that a sim will leave his or her home country is still determined by the *ChangeCountryProb*. However, the probability of reaching any other country in the continent is now 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 still extremely 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%, which is in the middle of the range to be tested, 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 most liberal model to be tested, which has five times this rate of inter-country migration, is still quite conservative in this respect.

4.1.2 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.[3] It is important, therefore,

[3] Higher density, more advanced, societies may also have a larger *proportion* of their citizens migrants, although that is not assumed in the

that the model take into account differing population density throughout the world.

The roman numbers in Table 2 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.[4]

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 B.C, and reduced populations were used thereafter to achieve a maximum world population of 55 million. As we saw in the first two models, if the population is reduced after the death of the MRCA, it should have little effect on the results. 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 1000 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 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 Table 2 give the scaled populations that were actually used in Model C. 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 smaller continents are scaled less: 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 (Diamond, 1997), 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 Table 2 were actually the targets used for 1400. At that point, the birth rate was reduced,

current model.

[4] The final numbers in Table 2 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.

Table 2: Continental populations, in thousands, at various points in time. Roman numbers are estimates of the true populations. The italic numbers below them are the rescaled values used in the Model C 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*

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 Figure 9. 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, Model C contains 497 countries and 15,059 towns.

4.2 Results

Again, it should be stressed that we do not have sufficient data to fix the parameters of the model with much certainty. Therefore, a series of simulations was conducted to determine a range of possible outcomes. The results are shown in Figure 10. For each set of parameters, at least six trials were performed and their resulting dates averaged. The results tend to be somewhat less consistent across trials in this model than in the previous ones, with MRCA and ACA dates having a coefficient of variation of between 6% and 10%.

The first three simulations were intended to test liberal, moderate, and conservative parameter sets. In the "liberal" simulation, C1, the *ChangeTownProb* was 20%, the *ChangeCountryProb* was 0.25%, the *NonLocalPortProb* was 20%, and the port rates were actually 10 times the values shown in Figure 9. The resulting MRCA date was 1945 BP and the ACA date was 4158 BP, which do indeed seem remarkably recent. The case could be made that simulation C1 is not excessively liberal. The *ChangeTownProb* of 20% is in line with the finding of Adams and Kasakoff (1976) that a natural threshold exists across societies for groups with an 80% endogamy rate. The *ChangeCountryProb* of 0.25% means that the average country throughout the final 2500 years of the simulation has just 5.3 emigrants per year, with many having much fewer. Prior to 500 B.C., emigration rates were even lower because of the smaller populations. Furthermore, even with port rates that are ten times those shown in Figure 9, most

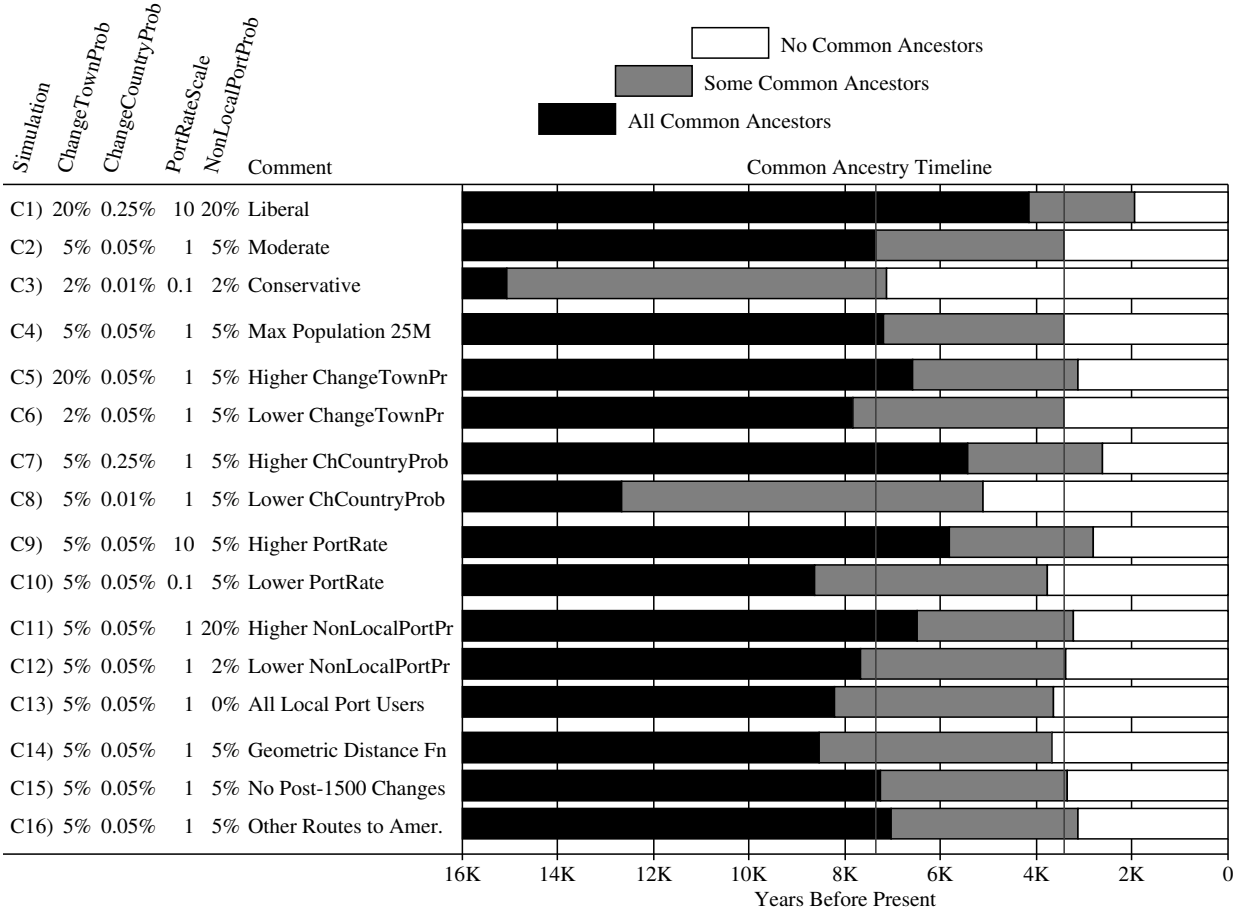

Figure 10: Results of the Model C simulations.

of the land bridges, with the exception of Sinai, still have only about 33 migrants per year, while the long-term nautical bridges average 3.3 or fewer migrants per year.

Nevertheless, it is possible that some of the parameters used in simulation C1 permit too much mobility. Therefore, C2 was conducted with more conservative parameters. The *ChangeTownProb* and *NonLocalPortProb* were both reduced to 5%, the *ChangeCountryProb* was reduced by a factor of 5 and the port rates by a factor of ten. These changes increase both the MRCA and ACA dates by about 75%, resulting in an MRCA date of 3415 BP and an ACA date of 7353. But the primary finding remains that our most recent common ancestor most likely lived within the past few thousand years, much more recently than Mitochondrial Eve or Y-Chromosome Adam.

The blue curve in Figure 4 shows the percentage of common ancestors as a function of time for one of the C2 trials. After an initially slow climb, the percentage of CAs increases quite rapidly and smoothly until it begins to slow at around 75%. This is the point at which most of the countries in Eurasia, Indonesia, and Africa contain CAs and the growth in the percentage of CAs slows as these continents saturate. This occurs at around the 90% mark, at which point the curve noticeably levels. The remainder of the growth involves the slower spread of common ancestry throughout North and South America and Australia. On average, the C2 sims reach 95% CAs in the year 5920 BP, 98% in 6399 BP, and 99% in 6584 BP, but do not reach 100% CAs until 7353 BP. Thus the point at which nearly everyone is a CA and the point at which truly everyone is a CA are separated by perhaps 1000 years.

In the interest of more firmly bounding the range of possible CA dates, a third simulation was conducted that is even more conservative than C2. C3 uses a *ChangeTownProb* of 2%, a *ChangeCountryProb* of 0.01%, a *NonLocalPortProb* of 2% and port rates that are 1/10 those shown in Figure 9. The resulting MRCA appears in 7110 BP and the ACA point moves back to 15042 BP. In this case there is a proportionately longer tail on the transition from the MRCA to the ACA point, with 99% CAs by 13087 BP. The role of the individual parameters in producing these less recent dates, and the reasonableness of those settings, are discussed in the next section.

4.2.1 Parameter effects

Simulations C4–C16 are variations on C2, exploring the effects of individual parameter changes to better understand the model. To aid the comparison, vertical lines in Figure 10 mark the approximate C2 MRCA and ACA points. The first question investigated is the possible effect of the population cap of 55 million. Although running simulations with larger populations was not possible on the available computers, simulation C4 was conducted with a population cap of 25 million, reached in 2000 BC. The result is absolutely no change in the MRCA date, while the ACA date was 2.4% more recent, a difference that does not approach significance. Therefore, it is likely that further growth beyond 55 million would not substantially change the model's predictions. As we saw in Models A and B, larger populations can actually result in more recent dates with conservative migration, but have little effect with more liberal migration.

Simulations C5 and C6 vary the *ChangeTownProb*, or the percent of sims who leave their home town. Increasing this value from 5% to 20% causes a 8.3% reduction in the MRCA time and a 10.4% reduction in the ACA time. Decreasing it to 2%, which is perhaps the minimal reasonable level given our definition of a town, causes no change in MRCA, although the ACA time increases by 7.9%. Because there is reasonably frequent migration within a country even at the lowest reasonable *ChangeTownProb* setting, this parameter has relatively little effect on the model.

In contrast, the model is more sensitive to variations in the *ChangeCountryProb* within the rather wide range of plausible values. Increasing this parameter by a factor of 5 to 0.25% in simulation C7 causes a 23.7% reduction in MRCA and a 26.2% reduction in ACA. Conversely, decreasing this parameter by a factor of 5 pushes the MRCA to 5103 BP in simulation C8, with an ACA point of 12648 BP. With a *ChangeCountryProb* of 0.01%, the average country in the year 1500 has only one emigrant every 4.7 years, most of which will head to neighboring countries and some of which may not even produce children. Less populated countries will have even fewer emigrants and there will also be many fewer early in the simulation because of the lower population densities. In the year 5,000 BC, the average inhabited country will have just one emigrant every 48 years. A *ChangeCountryProb* any lower than this may be unreasonably conservative.

Another important factor is the migration rate across inter-continental ports. The default rates shown in Figure 9, which were used in Simulation C2, were intended to be quite low. Scaling these rates up by a factor of 10, in C9, results in an 18.2% decrease in the MRCA time and a 21.1% decrease in the ACA time. Scaling down by a factor of 10, C10, produces a 10.1% increase in MRCA time and a 17.5% increase in ACA time. Therefore, the model is fairly sensitive to the port rates, but not nearly so much as it is to the *ChangeCountryProb*, which governs the majority of the migration throughout the world.

The percentage of port users who do not originate in the port's source country, the *NonLocalPortProb*, was manipulated in simulations C11–C13. Increasing this value to 20% or reducing it to 0% has very little effect on the MRCA time and a somewhat larger effect on the ACA time. Thus, the existence of migrants who travel long dis-

tances to use a port is not integral to the outcome of the model.

One reasonable concern, however, is that the model permits too much long distance migration within continents. In most of the Model C simulations, the probability of emigration to another country is proportional to the inverse square of the distance between the countries. As a result, really long distance journeys across continents are rare, but not exceedingly so. A stricter assumption might be that migration falls off geometrically with distance. In simulation C14, the probability of migration between two countries separated by distance d is proportional to 2^{-d}. As a result, very long distance migration is much less common. This increases the MRCA time by 7.2% and the ACA time by 16.0%.

4.2.2 Other experiments

Much has changed in the world in the last 500 years. Of particular interest to us is the continuing increase in geographic and social mobility in general and the large movements of people to particular areas including the European colonization of the Americas, Australia, New Zealand, and parts of Africa. Ultimately, these changes will result in a much more recent common ancestor, eventually approaching the predictions of Chang's panmictic model. One question, however, is what effects these relatively recent changes may have had so far on the common ancestors of people living today.

As described in Section 4.1.1, the model C simulations reported thus far involved numerous post-1500 changes, such as the opening of 16 large ports to the Americas and other areas and overall increases in the migration rates. But these additions were far from comprehensive. So we might ask whether the current additions have had any substantial effect on the results and, therefore, whether more additions could be expected to further alter the outcome.

Simulation C15 was designed to test the effect of the current post-Columbian changes by eliminating them entirely. This simulation was identical to C2 until 1400, after which the population sizes remained stable and no additional ports were created or parameters were altered. This had virtually no effect, actually decreasing the MRCA time by 2.3% and the ACA time by 1.3%, both highly non-significant. This suggests that the changes in human migration patterns over the past 500 years may still be too recent to have had much of an effect on our common ancestry, though they will certainly do so in the centuries to come.

There has been considerable debate over the likelihood that the Americas were reached, prior to Columbus, via routes other than the Bering Strait. These include possible contact between Vikings and native Americans in Newfoundland, in addition to those that may have occurred in Greenland, and, more significantly, contact between South America and the Pacific Islands. The primary evidence for the latter appears to be the spread of the sweet potato, believed to be of South American origin, throughout much of Oceania around 1500 years ago (Bellwood, 1979). If such alternate routes to the New World existed, how might they have affected our common ancestry?

This question was addressed with a variant of simulation C2. In C16, two ports were created in 400 AD, between Peru and the Marquesas Islands and between Chile and Easter Island, with migration rates of 10 s/g in each direction. In addition, a one-directional port was created from southern Greenland to Newfoundland lasting from 1000 to 1350 AD, also with a rate of 10 s/g. As shown in Table 10, this resulted in a statistically significant 8.8% decrease in MRCA time and a 4.6% decrease in ACA time (n.s.). Therefore, the possible existence of continued contacts with the Americas other than through Beringia may have had some effect on the dates of our common ancestors, but such contacts are not critical to the existence of a recent common ancestor.

5 Further analyses

To this point, most of the discussion in this paper has focused on the dating of our MRCA and the point at which all ancestors were shared in common. But, assuming the model is sufficiently accurate in its design and details, many other interesting predictions about our common ancestors and ourselves can be drawn from it.

5.1 Who was our most recent common ancestor?

So who was our most recent common ancestor? Was it a man or a woman? Where did he or she live? One might predict that the MRCA is likely to be someone, perhaps an ancient king, with many children who quickly migrated far from home. It turns out that the reality may be far less glamorous.

Studying the MRCAs of several Model C simulations, reveals that the most recent CAs lived either in southeast or northeast Asia, nearly always in a port country. The most common sites are Taiwan and Malaysia in the southeast and in the Chukotka and Kamchatka regions of northeast Russia, close to the Bering Strait. Living near a port is an obvious advantage because it allows one's descendants to rapidly reach a second continent. The reason that the MRCAs arise in either southeast or northeast Asia and not, for example, the Middle East, is their proximity to Oceania and North America, respectively. Because Oceania was settled quite late and relatively slowly, and because Australia has infrequent contact with the rest of the

world, it is advantageous, from the perspective of trying to become a CA, to live in this gateway to the Pacific. Likewise, it is also an advantage to live near the gateway to the Americas because the tip of South America is also difficult to reach. As a result, quite recent CAs emerge throughout far eastern Asia, including Japan and coastal China.

There has been considerable debate in recent years over the site of modern human origins. The multitude of hominid fossils found in east Asia has suggested to some researchers that modern humans may have evolved there, either solely or in parallel with their evolution in other parts of the world (Thorne & Wolpoff, 1992; Wu, Poirier, Wu, & Wu, 1995; Etler, 1996). However, the prevailing view, based primarily on the analysis of genetic markers, now seems to be that humans first arose in Africa before spreading to east Asia and throughout the world (Wilson & Cann, 1992; Jin & Su, 2000; Ke et al., 2001). Even if the origin of modern humans is not to be found in east Asia, perhaps it will be some small consolation to the multi-regionalist to know that the much more recent common ancestors of all humans very likely did live there.

Figure 11 shows the year in which the first CA arises in each country in one trial of simulation C2. In this trial, the MRCA was a man living in Taiwan, born in 3536 BP and who died in 3459 BP. Other CAs arose in Kamchatka and southern China within a few decades, working backwards in time, and then at various other locations in eastern Asia, both north and south. Within 600 years of the MRCA, CAs can be found throughout most of Eurasia, much of Indonesia, and some of north Africa. It takes another 2500 years for the CAs to appear in the more remote parts of North and South America. Note that no CAs lived in Greenland or Oceania because those areas were not yet inhabited.

By studying the immediate descendants of the MRCAs, we can get a better sense of who they may have been and how they earned that title. An average sim living at the time of the MRCAs has about 2.09 adult children. In contrast, a sampling of 130 first-generation CAs (FGCAs), none of whose descendants were also CAs, from simulation C2 had an average of 5.0 children. This is more than twice the overall average, but it is not an exceptional number. Only 5% of these CAs had more than 10 children, while 20% had just three children and 12% had only 2 children. The advantage for the FGCAs stems from the fact that their families happen to have maintained higher than average reproductive levels over the course of a few generations. The average number of children per descendant in the second generation is 2.63, followed by 2.28 in the third generation. Subsequently, birth rates are at average levels. As a result, although an average person at that time has 40 great great great grandchildren, the average FGCA has 130. Because the FGCAs do not tend to be the very rare sort of person with many children, there is no special advantage for men in this regard and just about half of the FGCAs were women. This is partially driven by the fact that husbands and wives tend to become MRCAs together.

One might expect, however, that an FGCAs descendants are likely to have disbursed very rapidly, but that is not always the case. 73% of the FGCAs had no descendants move out of the country in the first four generations, and 51% had no descendants migrate in the first five generations. Of the 22,710 descendants of the 130 FGCAs in the first five generations, only 34 moved away from their home country, a rate that is about three times the *ChangeCountryProb*. But of these, about a third reached a different continent. Therefore, the widespread disbursement of the lineages of MRCAs does not usually begin immediately. Their families grow gradually, at a somewhat higher than average rate, and then disburse gradually, also at a somewhat higher than average rate. The main advantage comes not in being extraordinarily prolific, but in living near the geographic center of the world and near a continental boundary so one's descendants reach another continent within a relatively short time.

It is interesting to track the descendants of a single one of these MRCAs throughout the course of a simulation to discover when they first arrived in each part of the world. The results, mapped in Figure 12, are in appearance very similar to those for the first occurrence of an MRCA in each country, but in this case we are working in the opposite direction in time. This particular MRCA was born in Taiwan in 1536 BC. She had a remarkable advantage in that one of her great grandchildren migrated up the coast to Chukotka. Other early descendants migrated throughout southeast Asia, with some heading to central Russia. Her lineage first reached Indonesia in 1206 BC, North America in 1091 BC, Africa in 838 BC, Australia in 652 BC, South America in 95 BC, and Greenland in 381 AD. Some of the last places reached were southern Argentina in 855 AD, and New Zealand in 1116 AD and the Chathams in 1419 AD, in the first wave of their colonizations.

It appears that our MRCAs were not necessarily remarkable people. They could have been men or women of any social standing, who, most of the time, just happen to have been living at a crossroads. However, this finding may simply be a reflection of the model's assumptions. The model does not allow for the fact that an ancient ruler may have achieved the status of MRCA by fathering many sons who, through continued wealth and power, themselves produced many children (Zerjal et al., 2003). Such a dynasty, if it existed in the right time and place, could have reduced the MRCA time. It is also important to keep in mind that the MRCA of all living humans is not a fixed person and will change over time. With the ever

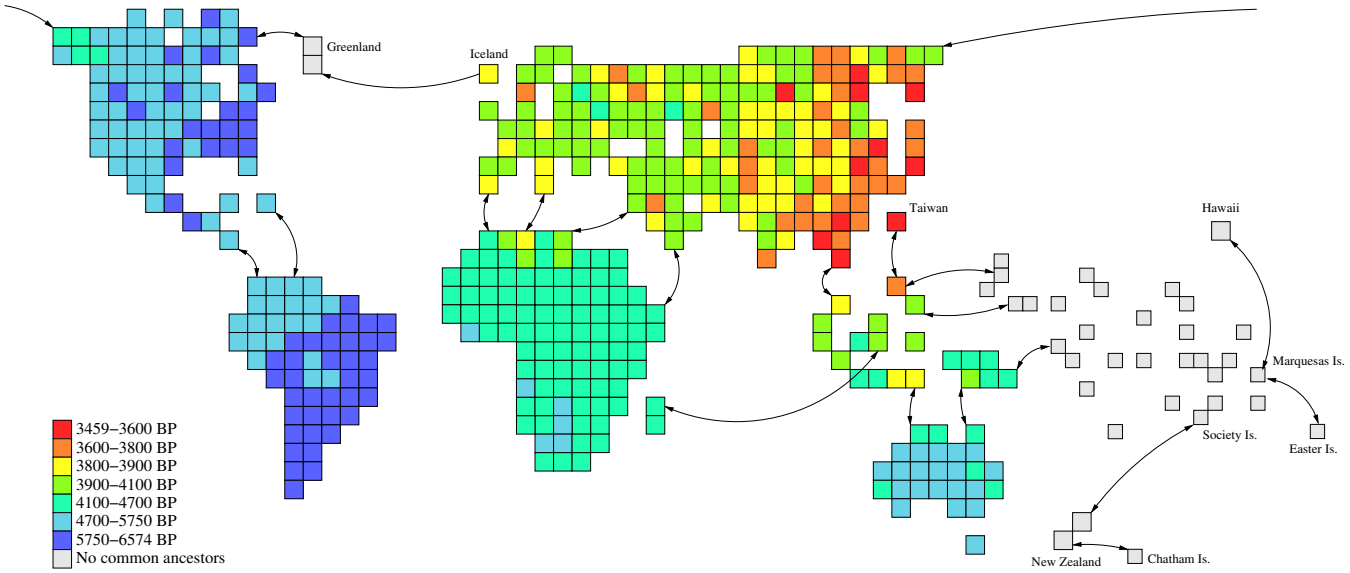

Figure 11: The date at which the most recent common ancestor of everyone alive today appears in each country in a selected C2 trial.

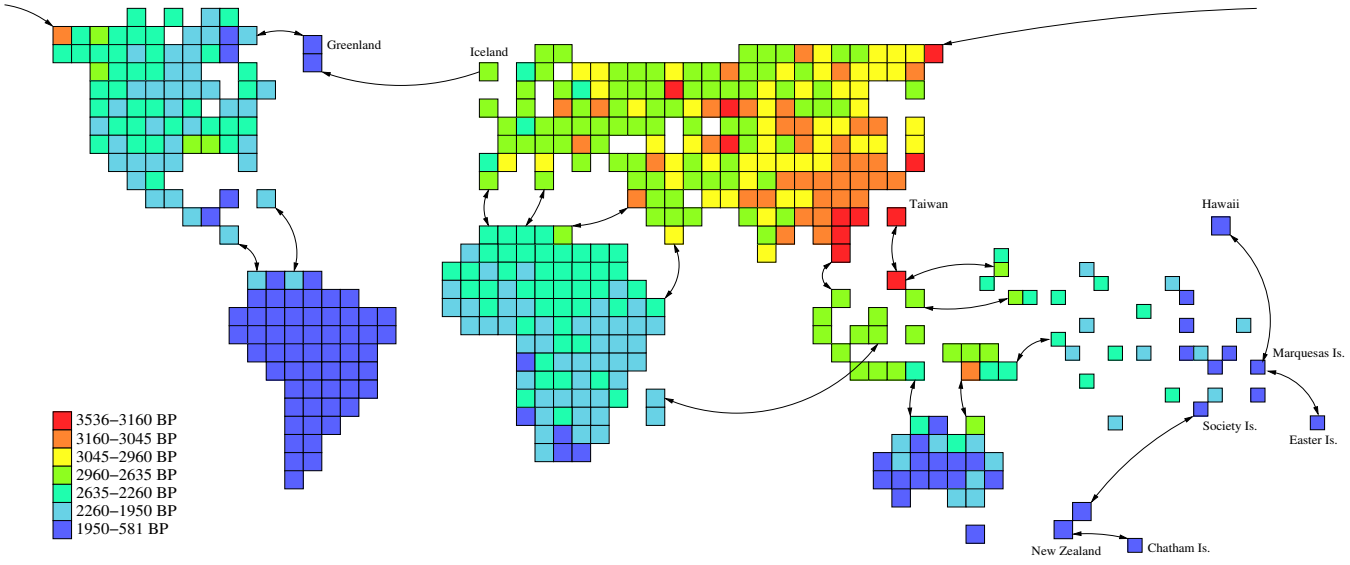

Figure 12: The date at which the first descendant of the MRCA appears in each country in a selected C2 trial.

growing mobility of the world's people, both geographically and socially, the MRCA will, relatively speaking, become increasingly recent.

5.2 Common ancestors of subpopulations

Although we have been looking at the common ancestors of everyone alive in the world today, one might wonder whether certain sub-populations, such as Europeans, share much more recent common ancestors. Starting with a trial from an earlier version of model C, which had a world-wide MRCA date of 3520 BP and an ACA date of 11380 BP, the ancestry of just the present-day sims living in Europe, east of the Black Sea, was tracked. The resulting MRCA appeared in 2120 BP, about 40% more recent than the worldwide MRCA. Although the worldwide MRCAs in the model nearly always live in east Asia, the first few European MRCAs were spread around Eurasia, with the first one appearing close to Israel or Jordan, and with others in Japan in 142 BC, in Siberia in 169 BC, in Burma in 172 BC, and in Iraq in 185 BC. Interestingly, the ACA point is exactly the same for Europe as it is for the world as a whole.

Given that the MRCAs for Europe are scattered throughout Eurasia, it should not be too surprising that the MRCA for all of Eurasia was nearly as recent, dying just 40 years earlier, in 2160 BP. The MRCAs for Eurasia arise first around Thailand and Burma, southern India, and southern China. The MRCA for Eurasia together with Africa in this simulation lived about 300 years earlier, in 2450 BP. They first arise scattered throughout Eurasia, first in Korea, then Uzbekistan, Spain, and Siberia. But these are results of a single trial and there is probably nothing significant about these particular locations. The most recent of the CAs who lived in Africa appear in Egypt, as one might expect, in 2530 BP.

5.3 Genetic inheritance

The point was made earlier that the existence of an ACA date, or a time at which everyone alive today shares the same set of ancestors, does not necessarily imply that we owe the same degree of ancestry to each of those people. Otherwise, it is unlikely that even the most superficial physical differences could have arisen since then. But the question is, to what extent does the ancestral inheritance of various peoples in the world today differ? Are the differences subtle or dramatic?

We can begin to answer these questions by tracing the ancestry of individual present-day sims. But in this case, we are not just interested in the identity of the ancestors, but in the percentage of the sim's genes attributable to each ancestor. We will assume that a sim owes exactly 1/2 of his genes to each of his parents, and thus 1/4 to each grandparent, and so on. Of course, if an ancestor appears more than once on the family tree, she will contribute the sum of the individual proportions. After many generations, the proportion of genes contributed by each ancestor becomes vanishingly small, until some ancestors may contribute no actual genes. But we can sum the proportions over each continent or country to get a picture of the percentage of genes the modern sim owes to ancestors living in various parts of the world at a given time.

We will first trace the ancestry of a randomly selected Japanese sim born in the year 2000 in one of the C2 trials. By 1500 AD, the sim owes 98.8% of his ancestry to his home country, the middle of the three Japanese territories, and much of the rest to the other two countries that form Japan. The remaining 0.4% is traceable to neighboring areas of China and Korea. By 500 AD, 98.9% of the sim's ancestry is still attributable to Japan as a whole. This declines to 97.5% by 2000 BC, 95.7% by 5000 BC, and 88.4% by 20000 BC. The proportion of the sim's ancestry attributable to each country in the world in 5000 BC is shown in Figure 13. The red and orange regions together account for 97.35% of the ancestry, with 2.62% from the rest of Eurasia, 0.014% from Africa, 0.00090% from Indonesia and Australia, and 0.00086% from the Americas.

Figure 14 shows the corresponding ancestry for a randomly selected Norwegian. In this case, 92.3% of the ancestry in the year 5000 BC is attributable to the country in which the sim lives, in central Norway, and 96% to Scandinavia as a whole. The Norwegian has about three times as much African ancestry as the Japanese sim, but much less American, Indonesian, and Australian. The Norwegian owes 0.00044% of his ancestry to 5000 BC Japan, while the Japanese owes 0.00049%, or about 1 part in 200,000, to ancient Norway. That would suggest that, at this rate of mixing, a typical Norwegian might be expected to have inherited about one haplotype block from 5000 BC Japan (Gabriel et al., 2002).

Therefore, because of the fairly low migration rates of simulation C2, the genetic inheritance of this randomly selected sim derives almost entirely from nearby areas, even looking 7000 years into the past. Given the quite recent MRCA date of 1460 BC in this trial, it may be surprising that the sim's ancestry remains so highly concentrated after such a long period. This serves to highlight both how conservative the parameters of the simulations actually are and the remarkable fact that a fairly recent MRCA emerges despite these conservative parameters.

The ancestry of a central African is shown in Figure 15. The pattern of tightly clustered ancestry is similar to that of the previous two examples, but in this case the African had two major ancestral branches, one in his home country and another stemming from a great-grandparent from the Ghana area. The African sim has almost entirely African ancestry, with 0.00092% Eurasian ancestry, pri-

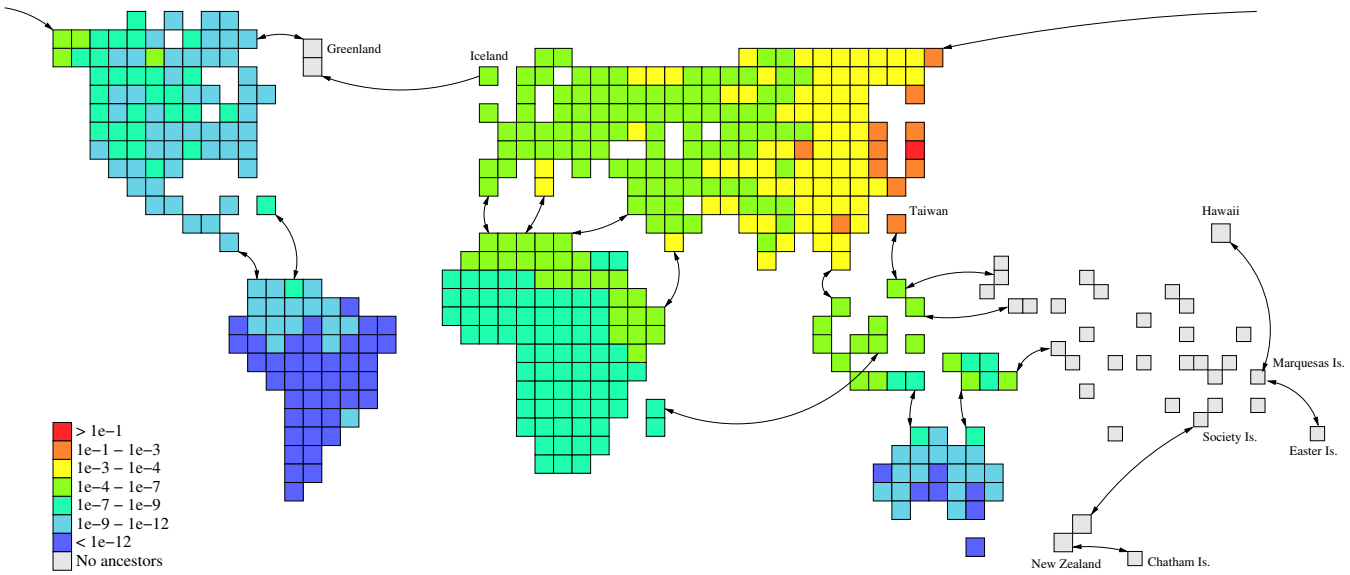

Figure 13: The fraction of ancestry of a randomly selected Japanese sim, who was born in 2000, that is traceable to each country in 5000 BC.

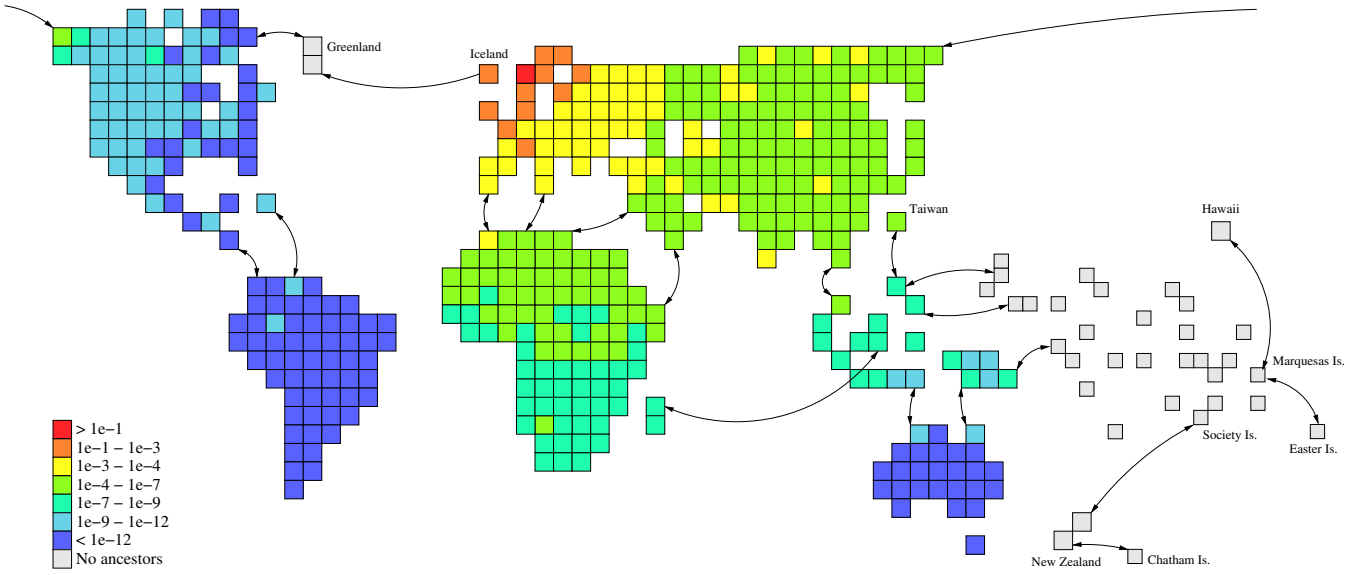

Figure 14: The fraction of ancestry of a randomly selected Norwegian sim, who was born in 2000, that is traceable to each country in 5000 BC.

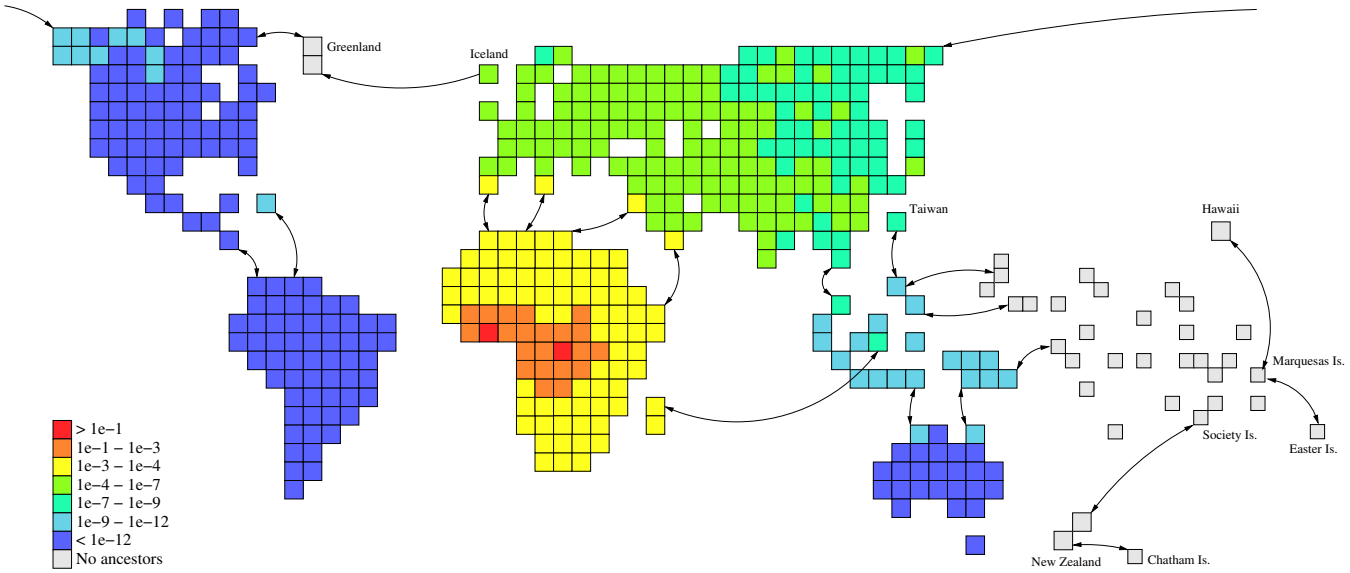

Figure 15: The fraction of ancestry of a randomly selected central African sim, who was born in 2000, that is traceable to each country in 5000 BC.

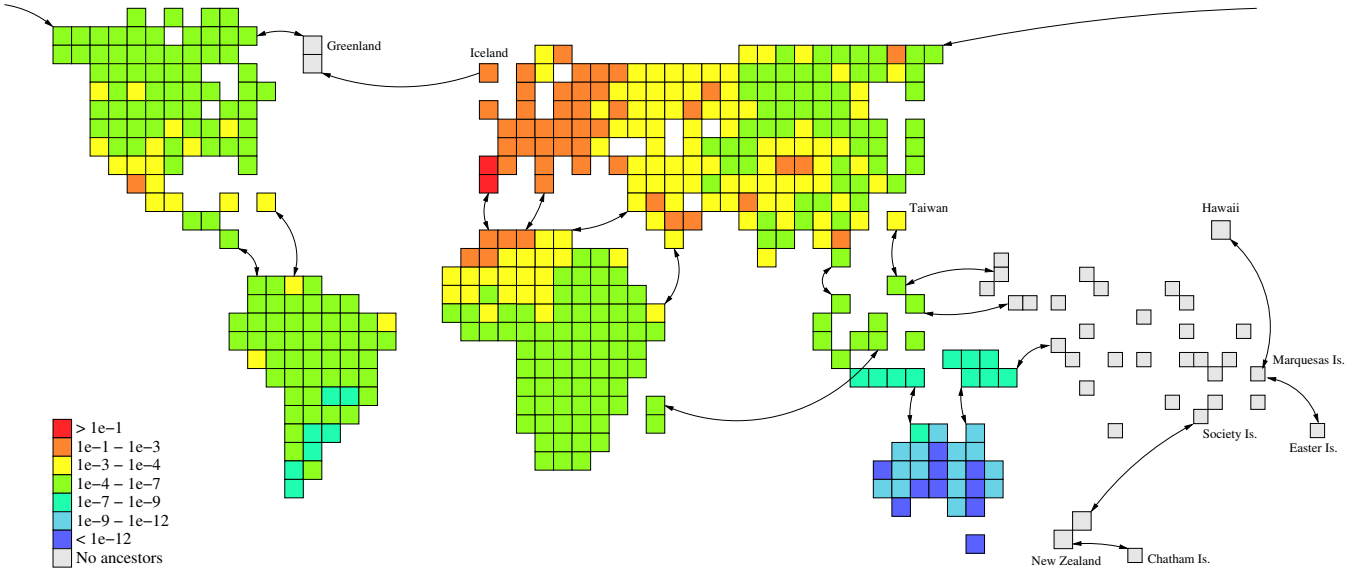

Figure 16: The fraction of ancestry of a randomly selected Mexican sim, who was born in 2000, that is traceable to each country in 5000 BC.

marily from the border countries. Due to the greater isolation of Africa, this sim has much less Indonesian, Australian, and American ancestry than did the Eurasians. The total South American ancestry, for example, amounts to just 1 part in 1.4 trillion.

Finally, Figure 16 shows the ancestry of a sim born in west-central Mexico. Unlike the other three areas, which are well isolated, Mexico is home to a fairly large influx from Spain. By 1800, 9.5% of the sim's ancestry is traceable to Europe. This percentage grows to 41.1% by 1700, 63.4% by 1600, and 85.8% by 1500. In the year 5000 BC, 3.45% of the ancestry is still traceable to natives of the same part of Mexico, with 4.1% from all of North America and 0.044% from South America. 85.3% of the ancestry is Eurasian, 39.9% from southern Spain, and 10.6% is African, primarily from northwest Africa. Thus, because both Mexico and Spain are recipients of a substantial number of migrants, the diversity of the Mexican sim is much greater than that of the Japanese, Norwegian, and African sims. But because the pre-Columbian population of Mexico and the post-Columbian migration rates to Mexico were not carefully controlled, in addition to the fact that we are only looking at a single sim, these results should merely be taken as illustrative. Although the four sims we have examined share nearly the same set of ancestors living in 5000 BC, their proportional inheritance from those ancestors is markedly different.

6 Discussion

This paper has sought to investigate an intriguing aspect of the human species—our common ancestry. Based on the results of a series of computer models, it seems likely that our most recent common ancestor may have lived between 2,000 and 5,000 years ago. This is, perhaps, one tenth to one one-hundredth the length of time to our most recent common ancestors along solely male or solely female lines, which have been the target of considerable recent interest. The point beyond which everyone alive today shares the same set of ancestors is somewhat harder to predict, but it most likely falls between 5,000 and 15,000 years ago, with a significantly more recent date for the point at which we share nearly the same set.

The present study focuses on ancestral, rather than genetic, inheritance, but variants of the model could be employed in studying population genetics by introducing and tracking the rate of spread of genotypes under more complex conditions than are allowed for in current analytical models. Computer simulations could also be valuable in testing the models and techniques currently used in population genetics, such as gene-tree-based methods, coalescence models, or nested-clade analysis (Hey & Machado, 2003). Operating over a longer time period and with the addition of genetic inheritance and mutation, such a simulation could generate hypothetical data samples as well as provide the true history underlying those data. If various analysis methods are then applied to the data, the potential biases of those techniques could be revealed, contributing to their improvement or a better understanding of their appropriate use.

It should go without saying that the models used in this work are based on a number of methodological and parametric assumptions, and the reliability of the predictions are dependent upon the appropriateness of those choices. So let us briefly reconsider some of these assumptions and the possible implications of altering them.

The length of the sims' lives is governed by a formula based on data from the United States between 1900 and 1930, skewed to produce average life spans of 51.8 years for those reaching adulthood. It is not clear if this is a reasonable assumption and may be an overestimate of average lifespans prior to modern medicine and widespread agriculture. The fact that a generation in the model is taken to be 30 years may seem rather long, but it results from the mean age of a newborn's parents, with mothers averaging a bit under 28 years old and fathers just over 33. Reducing the life expectancy would result in younger fathers and thus shorter generations and an even more recent MRCA.

The assumption is also made that women produce children only between the ages of 16 and 40, inclusive. Give or take a few years, this is probably a reasonable range. What may not be reasonable is the fact that the likelihood of a child is constant across these years. Certainly, in modern America, women are most likely to produce children in their mid-20s to mid-30s. However, in other parts of the world and historically, with less availability of birth control and stronger pressures to have large families, women may produce children at a broader distribution of ages or a distribution more skewed towards younger mothers. Further data are necessary to determine if the model's assumptions in this regard are reasonable. As with fathers, younger mothers should result in more recent common ancestors.

Although about a third of men produce no adult offspring, the choice of fathers in the model is reasonably egalitarian. If it were the case that societies have tended to be more elitist, with reproductive rights dominated by a few powerful males, we should expect higher extinction rates overall. This will result in a more rapid spreading of the fewer lineages that remain and therefore will also produce a more recent MRCA.

Population size is one area in which the model is clearly not accurate. Because of computational constraints, limited-size populations were used after 1000 BC. A larger population at the time an ancestor is alive will tend to result in a longer delay before that person be-

comes a common ancestor. However, population growth after the ancestor has died will have little effect because the 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. Because most of the MRCAs in these simulations were born prior to 1000 BC, the limited population should not have resulted in overly recent MRCAs. In fact, because a smaller population has fewer intra-continental migrants, according to the rules of the models, the reduced population may have resulted in an overly ancient MRCA.

A more critical set of parameters are those that directly govern migration. As migration rates decrease, the predictions generated by the model become increasingly sensitive. However, the model only becomes truly sensitive when the number of emigrants from each country falls below about one per generation, which may be an excessively conservative level. With the *ChangeCountryProb* of 0.05% used in simulation C2, the number of emigrants from a country per generation in the year 1500 ranged from 55.3 in Eurasia to just under 1 in Australia. Such rates are probably severe underestimates. Increasing this by a factor of 5 would reduce the MRCA time to about 2600 BP.

The other set of critical factors are the inter-continental migration routes and their usage rates. Model C uses a restricted set of about 48 paired ports throughout the world. If people actually followed a more diverse set of routes, which is likely, it too would lead to a more recent MRCA. The migration rates of the ports could certainly be questioned and it is not clear if they are under- or overestimates. But our simulations suggest that scaling these values up or down by a factor of 10 would probably only change the MRCA date by about 500 years in either direction.

One over-simplification of the model is the assumption of uniform migration patterns within and between continents, with the exception of initial colonization waves. In reality, history has seen the occasional, if not frequent, organized migration of peoples. These include, for example, the invasion of Europe and China by the Huns and Mongols, the Arab occupation of Spain, and the southward spread of Bantu speakers in Africa (Cavalli-Sforza et al., 1994). Such migrations may have resulted in the movement of large numbers of people, but because those people were presumably all related, the resulting effect on the spread of lineages is less significant. Therefore, ignoring such migrations may not have seriously affected the results, although they would have had a major impact on individual ancestry distributions.

Finally, some readers may be concerned with the possibility that the existence of small, extremely isolated island communities, either true island dwellers or cultural isolates, could mean that the MRCA is far less recent than proposed here. Easter Island, Hawaii, and the Chatham Islands were included in Model C to investigate just this possibility, but many other islands were left out. However, in order for a remote island to affect the MRCA date, it must have been colonized prior to the lineage of the mainland MRCA reaching the source population and it must have remained genetically isolated until very recently in order for there to be surviving people with no modern mainland ancestry. This is an unlikely combination because any island that could have been discovered using the seafaring technology available several thousand years ago and large enough to be capable of sustaining a human population would probably have been rediscovered many times since then, if not participating in continuous trade and exchange of people.

Many of the smaller islands, if they had been effectively isolated, were probably colonized by Europeans early enough, and their native populations at the time of discovery so reduced by disease or enslavement, that no purely native descendants remain. Although the Andaman islands are renown for their isolated tribes, the Sentinelese in particular, who may have no recent European ancestry, these islands probably do not present a problem for the model's predictions. They are close enough to Burma, Thailand, Indonesia, and India to have maintained fairly steady contact over the past two thousand years. Furthermore, the Andamans happen to be located very close to the probable home of the MRCA, and therefore may have actually been one of the first places reached by his or her lineage. Finally, even religious isolates such as the Samaritans are known to have accepted occasional outsiders (Bonné-Tamir et al., 2003) and were, in any case, probably already descendants of the MRCA at the time of their formation.

Although even the best of the current models has numerous limitations and underconstrained parameters, efforts were made to bias its design towards overly conservative assumptions, skewing the results towards slower coalescence of ancestral lineages. Nevertheless, the simulations predict that we all share a common ancestor who lived just 70–170 generations ago. A single ancestor living thousands of years ago may have, individually, contributed nothing at all to a modern person's genotype, and the notion of a common ancestor is of little practical importance as far as a geneticist might be concerned. Nevertheless, ancestry is a concept that long predates genetics, and the finding that everyone on earth today shares such recent common ancestors may be, for many, a remarkable and inspiring one. Indeed, we are all related.

Acknowledgments

This research was supported by NIH NRSA 1F32MH65105-01. Correspondence regarding this article may be sent to Dou-

glas Rohde (dr@tedlab.mit.edu), Massachusetts Institute of Technology, NE20-437E, 3 Cambridge Center, Cambridge, MA 02139. Special thanks to Steve Olson and Joseph Chang for their extensive contributions to this work. Thanks also to the Center for the Neural Basis of Cognition at Carnegie Mellon and the University of Pittsburgh and to Dr. Edward Gibson at MIT for the use of their computers.

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.

Bellwood, P. S. (1979). Man's conquest of the Pacific: The prehistory of Southeast Asia and Oceania. In A. V. S. Hill & S. W. Serjeantson (Eds.), *The colonization of the Pacific: A genetic trail* (pp. 1–59). Oxford: Clarendon Press.

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

Bonné-Tamir, B., Korostishevsky, M., Redd, A. J., Pel-Or, Y., Kaplan, M. E., & Hammer, M. F. (2003). Maternal and paternal lineages of the samaritan isolate: Mutation rates and time to most recent common male ancestor. *Annals of Human Genetics*, *67*, 153–164.

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*, 31–36.

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

Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Advances in Applied Probability*, *31*, 1002–1026.

Davis, D. L., Gottlieb, M. B., & Stampnitzky, J. R. (1998). Reduced ratio of male to female births in several industrial countries: A sentinel health indicator? *Journal of the American Medical Association*, *279*, 1018–1023.

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

Dorit, R. L., Akashi, H., & Gilbert, W. (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science*, *268*, 1183–1185.

Etler, D. A. (1996). The fossil evidence for human evolution in Asia. *Annual Review of Anthropology*, *25*, 275–301.

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

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., , & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, *296*, 2225–2229.

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.

Hey, J., & Machado, C. A. (2003). The study of structured populations: New hope for a difficult and divided science. *Nature Reviews: Genetics*, *4*, 535–543.

Jin, L., & Su, B. (2000). Native or immigrants: Modern human origin in East Asia. *Nature Reviews: Genetics*, *1*, 126–133.

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.

Kämmerle, K. (1991). The extinction probability of descendants in bisexual models of fixed population size. *Journal of Applied Probability*, *28*, 489–502.

Ke, Y., Su, B., Song, X., Lu, D., Chen, L., Li, H., Qi, C., Marzuki, S., Deka, R., Underhill, P., Xiao, C., Shriver, M., Lell, J., Wallace, D., Wells, R. S., Seielstad, M., Oefner, P., Zhu, D., Jin, J., Huang, W., Chakraborty, R., Chen, Z., & Jin, L. (2001). African origin of modern humans in East Asia: A tale of 12,000 Y chromosomes. *Science*, *292*, 1151–1153.

McEvedy, C., & Jones, R. (1978). *Atlas of world population history.* New York, pp. 342–351: Facts on File.

Möhle, M. (1994). Forward and backward processes in bisexual models with fixed population sizes. *Journal of Applied Probability*, *31*, 309–332.

Nordborg, M. (2001). Coalescent theory. In D. Balding, M. Bishop, & C. Cannings (Eds.), *Handbook of statistical genetics.* Chichester, UK: Wiley.

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

Rohde, D., Olson, S., & Chang, J. (in press). *Recent common ancestors in structured populations.* Manuscript submitted for publication.

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.

Thorne, A. G., & Wolpoff, M. H. (1992). The multiregional evolution of humans. *Scientific American*, *266*, 76–83.

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.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science*, *253*, 1503–1507.

Wilson, A. C., & Cann, R. L. (1992). The recent african genesis of humans. *Scientific American*, *266*, 68–73.

Wu, H.-C., Poirier, F. E., Wu, X., & Wu, P. (1995). *Human evolution in China: A metric description of the fossils and a review of the sites.* Oxford: Oxford University Press.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M. E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S. Q., & Tyler-Smith, C. (2003). The genetic legacy of the mongols. *The American Journal of Human Genetics*, *72*, 717–721.