Science
AAAS

Supplementary Information for

Quantitative analysis of population-scale family trees with millions of relatives.

Joanna Kaplanis, Assaf Gordon , Mary Wahl, Tal Shor, Omer Weissbord, Dan Geiger, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, Gaurav Bhatia, Daniel G. MacArthur, Alkes L. Price, Yaniv Erlich*

*Correspondence to: erlichya@gmail.com

This PDF file includes:

Supplemental Methods

Fig S1-S21

Table S1-S6

Movie S1

Contents

Supplemental Methods

Data acquisition and cleaning

Initial round of data gathering

We created scripts based on Geni.com `RESTful APIv0` to systematically download public data from the website. This process took several months due to the rate limitation of Geni.com on third party applications. The returned dataset was millions of `JSON` files that represented individual profiles and their immediate families (see an example of the public profile of Sewall Wright). We parsed the familial connections in the returned `JSON` files and represented the data as a graph using the `C++ Boost Graph Library` and custom `Python` scripts.

The topology of family trees can be represented as a bipartite directed graph $G = (P, U, E)$, where the P nodes represent individuals (profiles on Geni.com), U nodes represent union between two individuals (which we assume to reflect marriages), and E represents the set of edges between the nodes. Let $|X|$ denotes the size of set X. G_{raw}, initially the raw graph prior to any processing, had $|P| = 43.72$ million individuals, $|U| = 13.3$ union events, and 4.5 million connected components, each of which denotes a family tree. The largest connected component was a family tree of 15.4 million people.

Removing cycles

Biology dictates that G is a Direct Acyclic Graph (DAG) because an offspring cannot be the ancestor of any of his parents. Consanguinity does not create cycles, rather it induces multiple paths between an ancestor and an offspring. However, G_{raw} had cycles presumably due to errors and the fact that the data was collected over three months while active merging events have been carried by the Geni genealogists. To identify cycles in G_{raw}, we used Tarjan's strongly connected components algorithm (strong connected component is a subgraph where every node is reachable by any other node and contains one or more cycles) (*60*). We found a total of 1018 strong components, the largest component included 3814 vertices (profiles and unions) and the average component had 15 vertices. The most common cycle was an offspring that is the parent of one of his parents. We removed all $15,256$ nodes that were part of the strong components and the edges that connected these nodes to the rest of the graph.

Cycle removal had a minimal effect on the overall topology of the graph. The number of individuals in the largest connected component was reduced by 0.3% to 15.3 million and the number of connected components was increased by the same percentage.

Cleaning up multi-parents

Two parents are necessary for a reproductive event. Thus, the maximal indegree for any node in U cannot be larger than 2 and the maximal indegree for any node in P cannot be bigger than one. Again, we found nodes that violate this assertion and had more than two parents or were assigned to more than one union event. Our initial strategy was to eliminate these events when possible by locally merging profiles that were a mere duplication of each other (fig. S2).

First, we employed a ”merging-up” procedure. We created an algorithm that scanned for: (a) union nodes with indegree of more than 2, or (b) profile nodes with indegree more than 1. When such a node was identified, the algorithm retrieved the genealogical parents and evaluated whether multiple parental nodes represent the same person. For that, the algorithm employed a similarity test for each pair of parents. The test consisted of checking whether the reported sex of the two parental profiles were concordant and whether their first names matched under the phenotypic Soundex system, which can accommodate spelling variants. If the two profile nodes passed the similarity test, the algorithm merged them by transferring the edges of one node to the other node and excluding the former node from the dataset. Next, the algorithm moved one level up and merged the upstream union nodes by transferring the edges from one node to the other. The algorithm kept moving upwards towards the ancestral nodes like a zipper, performing similarity tests to putative profile nodes and merging them and their upstream union nodes. In order to increase the reliability of the merges, the algorithm committed to the merges only if it did not fail in any similarity test along 5 successive generations from the initial merge. Otherwise, the algorithm undid all operations originated in this round, restoring this section of the graph to its original topology.

Next, we employed a ”merging-down” procedure. The algorithm scanned again the nodes that were merged in the previous step but this time in went downwards in order to merge duplicate profiles of descendants. Since the algorithm already decided that there are two duplicated family trees, it just employed the similarity test for each generation until no more profiles were available.

Before the merging procedures, there were $354,502$ nodes with more than two parents. The merging up procedure merged $321,985$ nodes (including ancestral nodes to the one that triggered the procedure) and the merging down procedure merged $132,623$ nodes. In total, the algorithm resolved $180,507$ out of the $354,502$ nodes that had more than two parents.

The last step of the algorithm was to discard nodes with more than two parents that could not be resolved with the local merging procedure. Similar to the cycle clean up, we excluded those nodes from the dataset and pruned the edges connecting them to the rest of the graph.

The clean-up step had an effect on the overall topology of the graph. The number of individuals in the largest connected component was reduced by 15% to 13 million and the number of connected components was increased by 1.17%.

Estimating clean up accuracy

To evaluate the accuracy of the merging procedure, we compared the decisions of the merging algorithm to the decisions of the Geni.com genealogists. We randomly selected 1000 profile pairs that were merged during the merging up step and downloaded them again from the Geni.com website after some time from the initial collection of the data. We found that in 966 cases, the genealogists merged these profiles with each other or simply deleted one of them. In 18 cases, one of the profiles was merged to another profile but not to the one that was predicted by the merging algorithm, and in 16 cases no action was taken. The last two cases represent either potential failures of the merging algorithm or just incomplete work of the genealogists. Therefore, we estimate that at least 96.6% of the up-merges were correct based on agreement with the genealogists decisions.

We also evaluated the accuracy of the merging-down algorithm using a similar procedure of comparing our algorithm to decisions by human genealogists in 1000 cases. Here, 84% of the merges by our algorithm were concordant with the the decisions made by the genealogists. The size of the 'no action' cases was much larger (65 cases). We presume that the lower concordance in the down merges is in part attributed to a lack of a clear trigger for merging for the genealogists. It is easier for the genealogists to see more than two parents in their shared trees, which will initiate a merging up event. However, the genealogists do not always finish the process and eliminate duplicated nodes from the newly merged trees.

To get an additional independent estimator of the merging concordance, we also examined the profile photos that were associated with pairs of merged nodes. The hypothesis of this procedure was that profiles that are true duplicates of each other will have profile photos of the same person. We asked two human raters to independently inspect pairs of photos of profiles that were merged and to report whether the same person appears in both photos. To increase the reliability of the analysis, we included only cases where the reports of the two raters was identical. In 161 out of 171 merge cases with pairs of facial photos, the two raters concluded that the same individual appeared in both photos. With this result, the correct merging rate is estimated to be around 94%.

Second round of data collection

The initial round of data collection was done in 2011. In 2015, we scanned the website again to update our dataset. This round increases the size of our collection to 86 million unique profiles. We merged this new profiles with the old using the profile-id numbers of Geni. When the genealogists merge two profiles together, the Geni data flags one of the profiles a obsolete and indicated the newly consolidated profile (”merge-to”).

We followed all ”merge-to” events and consolidated the information of the profiles in our clean tree. Using this procedure, we took the most up to date data and copied to the clean trees. To estimate the number of genealogists that contributed the data, we used the ”created-by” field. All data was stored in a PostgreSQL database and is available on the FamiLinx website.

Estimating the accuracy of the tree using genetic markers

We sought to evaluate the quality of the pedigree data in Geni.com using genetic markers. To this end, we downloaded publicly available records from websites such as Ysearch.org and Mitosearch.org that we used in our previous studies (*61*). The records in these website include either Y chromosome STR markers or mitochondrial D-loop markers and also a small pedigree that can be matched to our records. We only obtained data that is publicly available to any user with Internet connection and were voluntarily shared by individuals.

If two individuals are truly related, the distribution of the number of mismatches in their genetic markers can be approximated as a Poisson process with a parameter $\lambda = \mu \cdot g \cdot n$, where μ is the mutation rate per locus per meiosis, g is the total number of meiosis events between the records, and n is the number of loci that were considered. Following previous studies (*62*), we set $\mu_{Y-STR} \approx 1/500$ mutations per meiosis per marker and $\mu_{mito} \approx 3 \times 10^{-5}$ mutations per meiosis per nucleotide.

The distribution of the number of mismatches between individuals that are not related was calculated by taking random pairs of records in our dataset and empirically measuring the number of mismatches.

We determined for each pairs of records the number of meiosis events between them and the observed number of mismatches and compared the two hypotheses whether the individuals are likely to be related or not. For that, we approximated the two distributions of the number of mismatches above as Gaussians. Next, we performed a likelihood ratio test with an equal prior to discriminate between the hypotheses that the pair of individuals is related versus unrelated. For the mitochondrial data, we were able to test 209 pairs of allegedly related individuals through their maternal lines, which spanned 1,768 meiosis events. Only five pairs of individuals were unrelated according to the likelihood ratio test. As the prior expectation for non-maternity is low, we assumed a single non-maternity event per erroneous maternal line. With that, the maximum likelihood rate of non-maternity per is 5/1768 = 0.3% per meiosis event. For Y-STR data, we were able to test 28 pairs of allegedly related individuals, which spanned 324 meiosis events. Six pairs of individuals were unrelated according to the same procedure. Assuming a single non-paternity event per erroneous line, the maximum likelihood rate of non-paternity rate was 6/324 = 1.9% per meiosis event.

The rate of non-paternity matches previous estimates of Europeans. A meta-analysis based on 67 studies estimated the non-paternity rate to be 2% in Europeans (*24*). More recently, a surname study in the UK with over 1,500 samples found a quite similar non-paternity rate (*25*).

Validating key demographic parameters

Obtaining age of death

The year of birth and death for each profile was extracted by using the `birth_date` and `death_date` fields. Full date was defined as (i) entries that have non-empty strings in their corresponding month, day, and year fields AND (ii) the ”circa” flag in the Geni JSON is ”0” AND (iii) the month field is between 1 to 12 AND (iv) day field is between 1 to 31.

The age of death of a profile was defined as the difference between the year of death to the year of birth. We filtered a few thousand entries with negative life span or life span above 100. Those entries are usually due to simple typos such as reporting the year of birth with four digits (e.g ”1910”) and the year of death with two digits (e.g. ”64”).

Comparing Geni to traditional demographic data

The expected life span from the Oeppen and Vaupel (O&V) study was taken from the male column in their Supplemental Table 2. We used the male column due to slight excess of males in our data. The life expectancy data from Geni was on average 1.4 year lower than their study across all years. This finding is quite expected as the O&V dataset was ascertained from countries with the highest life expectancy, whereas the Geni life expectancy is from a wide range of countries. One notable difference in the datasets is around 1840-1860, where the O&V curve jump very fast, where the Geni life expectancy was quite flat. This interval in the O&V data was nearly entirely taken from Norway. According to the Economy History Association, Norway experienced an economic boom during this period. Specifically, the repeal of the British Navigation acts in 1849 led to a large increase of the Norwegian exportation market. The Norwegian Per Capita GDP growth of approximately 1.6%, much higher than the European average of 0.95% and led to a significant reduction in infant mortality and a rapid increase of seven years in Norwegian life expectancy. However, this spike in life expectancy seem to be largely confined to Norway. Therefore is not picked in our data that reflects major trends rather than local ones. For example, during the same time of economic boom in Norway, the life expectancy in France and Belgium stayed largely the same, similar to the flat trend in our data. However, the France and the Belgium data showed a reduced life expectancy compared to our data. This is likely the result of underestimation of infant mortality as explained in the paragraphs below.

Data from the Human Mortality Database (HMD) was downloaded from France, Belgium, England and Wales, and Denmark for 1820, 1850, 1880, 1910, and 1940. Those countries were selected because they had data for the 19th century and since most of the population in Geni has Western European heritage. The life span probability density function for each country and time point was calculated using the `"dx"` column of the HMD file, which lists the number of deaths in each age. We then averaged the histograms from all four countries, except for 1820, in which we only had data from France (fig. S5).

Let $p_t(x)$ and $q_t(x)$ be the Geni and the HMD age of death distributions for year t, respectively. Define $\mathrm{BC}_t \triangleq \sum_x \sqrt{p_t(x)q_t(x)}$, where BC_t is the Bhattacharyya coefficient for the two distributions for year t. The worst Bhattacharyya coefficient reported in the main text was set to $\mathrm{BC}_{worst} = \min(\mathrm{BC}_{1820}, \ldots, \mathrm{BC}_{1940})$.

While the Bhattacharyya coefficient was high (> 0.95), we did detect a systematic difference of an approximately 50% reduction in the mortality rate before the first birthday in the Geni data (fig. S5B). This difference presumably reflects the inherent difficulty for genealogists of obtaining documentation of neonatal and infant deaths that occurred several generations ago. As our analysis of the genetic of longevity involves only adults over 30 years old, this difference in infant mortality should not affect our conclusion regarding the genetic architecture of adult longevity. However, it should be taken into consideration for future studies with our dataset that investigate life expectancy, infant mortality rates, or fertility rates, especially in early periods.

Annotating geographic information

In order to analyze the geographical origin of the Geni profiles, we used two sources of information: user approved annotations and automated geo-parsing.

User approved annotations: the Geni website allows users to enter the location of a genealogical event in a foreign language and immediately presents potential annotations of place in English. The user has the opportunity to review the annotation before saving the information. The saved location is then automatically assigned to longitude and latitude coordinates based on a pre-compiled table. This feature supports over 120 languages and dialects including historical languages such as Yiddish and ancient Greek. For example, consider a user that copy-pastes ”東京都” from an historical document and places this string in the place of birth of a profile. The website instantaneously presents the user with the following annotation: “Tokyo, Japan”; after the user’s approval, the website assigns the location to Tokyo’s coordinates (fig. S6A). The user-approved annotations is a relatively new feature for Geni that was added to the website about five years after its launch and not all profiles have this information.

Automated geo-parsing: for the old profiles, we only had free text for the following fields: `birth_location`, `current_residence`, `death_location`, and `burial_location` fields of the `JSON` files. This text displayed substantial heterogeneity. In some cases, the genealogists entered the location as part of a short sentence (”Census reports say he was born in Pennsylvania, Cleveland, Ohio, United States”) or in an inconsistent format (”, , New York, USA”). Also some of the text was in foreign languages, such as Russian, French, and Hebrew. To deal with this heterogeneity, we used the geoparsing capabilities of Yahoo! Placemaker service. This web service accepts unstructured text in 11 languages and returns structured location information that includes: the annotated location in a canonical format, the type of the location (e.g. town or state), longitude & latitude coordinates, and the quality of the annotation as a score between 1 (poor) to 10 (excellent). To obtain the most probable location when more than one place could match, we turned on the `autoDisambiguate` option of Yahoo! Placemaker service. In order to process the data efficiently, we constructed a corpus of locations that was comprised of all unique entries in the location fields of the `JSON` files and submitted the unique entries to Yahoo! Placemaker. To increase the reliability of the annotation, we excluded annotations that were too broad (such as continent or country). Next, we filtered annotations with quality level of 8 or above.

Importantly, we kept the user approved annotation from Geni if this information was available. If not, we annotated the data with the Yahoo geo-parser. Fifty percent of the trio data and 63% of couples were derived from user approved annotations.

Consistency of geoparsing versus manual curation

One potential concern is that the results are biased due to difficulties to geo-parse various languages, morphological differences in the name of places, and instances where the same name refers to different places such as Bethlehem in the West Bank versus Bethlehem in United States.

The user approved annotations have the advantage of having a human with a vested interest (the user) approving each annotation. To evaluate potential biases in the automatic geoparsing pipeline, we repeated the analysis of familial dispersion but restricted the profiles to individuals whose geographic information was provided by the high quality user approved annotations. We did not observe any major difference after restricting the data to these high quality annotations that could affect our conclusion (fig. S6B).

Validating the annotation accuracy using historical events

We placed profiles in the time-space domain by associating either their birth year with their birth location or their death year with their death/burial location, with the former given a priority over the latter. The year of birth or death was extracted only from fields that had full date information. The years of settlement were taken from the Wikipedia page of each city and reflected the earliest documented Western settlement in the area, which in some cases was before the official incorporation of the city.

Validating the key demographic parameters with the Vermont death certificate collection

We further evaluated potential ascertainment biases regarding socioeconomic status and causes of death. To this end, we were assisted by the Vermont death certificate registry. This resource is publicly-available upon request from the Vermont Department of Health. The registry contains a digitized collection of each of the 78,029 death certificates in Vermont between 1985 to 2000. These records include detailed demographic data on each deceased person, including exact birth date, death date, education, state of birth, place of death, and an ICD-9 code describing the cause of death.

We found Geni profiles that deceased in Vermont area during 1985-2000. To this end, we compared the birth date and death date between the two collections and used the first name to validate the match. This process yielded close to 1000 records that overlap between Geni and the Vermont vital record.

To assess potential socioeconomic biases, we compared the distribution of the education level, birth state, and ICD-9 cause of death between: (i) the entire Vermont collection and (ii) Geni profiles who we can find information in the Vermont collection. For education levels, we found 99.3% concordance (Bhattacharyya coefficient) between the subset of the Geni records and the entire Vermont death records (Table S1). For state of birth, we found 98.8% concordance between the Geni records and the Vermont death records (Table S2). For ICD-9 codes, we first grouped each cause of death into their broad categories such as heart conditions, neoplasm, and injuries based on the ICD-9 numbers and then compared the subset of Geni profiles to the entire Vermont collection. This process revealed a 99.4% concordance between the two groups.

The genetics of longevity

Modeling the expected life span

Longevity was defined as the deviation of the age of death from the expected life span based on temporal and environmental factors. As we were interested only in adult death, we restricted the analysis only to profiles whose age of death is above 30. We evaluated a series of nested models with a range of non-genetic factors that have been associated with lifespan, namely sex, year of birth, country of birth, the exact longitude and latitude of birth location, and the average temperature (fig. S8). The following R code describes these models:

```
# only gender
model1<- bam(long ~ gender, samfrac=1, data=training_set)

# birth year
model2<- bam(long ~ birth_year ,samfrac=1, data=training_set)

# gender and birth year
model3<- bam(long~gender+birth_year ,samfrac=1, data=training_set)

# gender, birth year, and birth country
model4<-bam(long~gender+birth_year+birth_country, samfrac=1, data=training_set)
```

```
# gender, birth year, temperature (mean and sd)
model5<- bam(long~gender+birth_year+tmp_mean+tmp_sd , samfrac=1, data=training_set)

#gender, birth year, geo location using spline regression
model6<- bam(long ~ gender + birth_year +s(birth_location_latitude,
    birth_location_longitude, bs="sos",k=splines), samfrac=1, data=training_set)

# gender, birth year, birth country, and temperature (mean and sd):
model7<- bam(long ~ gender+ birth_year+ birth_country+ tmp_mean+ tmp_s
    ,samfrac=1,data=training_set)


# gender, birth year, temperature (mean and sd), and geolocation using spline regression:
model8<- bam(long ~ gender + birth_year+ tmp_mean+ tmp_sd+ s(birth_location_latitude,
    birth_location_longitude, bs="sos",k=splines), samfrac=1, data=training_set)


# gender, birth year, birth country, and geolocation using spline regression:
model9<- bam(long ~ gender +birth_year + birth_country + s(birth_location_latitude,
    birth_location_longitude, bs="sos",k=splines), samfrac=1, data=training_set)

# gender, birth year, birth country, and temperature (mean and sd), and geolocation
    using spline regression:
model10<- bam(long ~ gender+ birth_year+ birth_country+ tmp_mean+ tmp_sd+
    s(birth_location_latitude, birth_location_longitude, bs="sos", k=splines),
    samfrac=1, data=training_set)
```

Temperature information as GIS was obtained from WorldClim in 2.5min resolution. The mean temperature reflects the average annual temperature and the standard deviation reflects the per month spread around the mean. We used the birth location of each profile to query the GIS grid and obtain the average temperature and standard deviation. While individuals can migrate from their birth locations, our data show that most migration events are very small especially before 1850. Therefore, we posit that the temperature at place of birth largely reflects the climate that the person experienced.

The country of birth was defined as a factor variable. To increase the power to associate this covariate, we restricted the data to individuals that were born in the top 25 countries with most profiles, namely: ”United States of America” ,”Netherlands” ,”Sweden” ,”Germany” ,”United Kingdom” ,”Norway” ,”Estonia” ,”Finland” ,”Canada” ,”Denmark” ,”France” ,”Australia” ,”Belgium” ,”Poland” ,”South Africa” ,”Russia” ,”Italy” ,”Czech Republic” ,”Switzerland” ,”Ireland” ,”New Zealand” ,”Austria” ,”Croatia” ,”Hungary” ,”Spain”.

The training set for these models consisted of approximately 3 million individuals from our resource who had exact data regarding date of birth, death, and location, and lived at least to the age of thirty. For validation, the models reported their goodness of fit (R^2), mean squared error, Bayesian Information Criterion (BIC), and the statistical significance of each factor using an independent set of 300,000 individuals.

The best model included sex, birth year, and the exact longitude and latitude; simpler models had higher MSE, lower R^2, and worse BIC levels, while more complex models that included the country of birth or the temperature did not improve the goodness of fit, and the additional factors did not reach statistical significance in most cases. The best model explained about 7% of variance of longevity in the validation set.

To calculate the longevity of individuals, we simply subtracted their age of death from the expected life span using this model.

Filtering individuals before evaluating genetic models

To further reduce the impact of environmental factors, we employed the following filtration steps:

(a) Including only individuals with full year of birth and year of death for which we can calculate their life expectancy using the model above.

(b) Removing individuals that were born after 1910 to avoid ascertainment bias towards early lifespans or before 1600, which can have lower reliability.

(c) Removing individuals who died during the American Civil War, WWI, and WWII, which showed a marked increase in death rates of military age individuals.

(d) Removing individuals without precise geographical assignment.

After finding pairs of relatives (see next section), we filtered additional pairs:

(e) We filtered potential twin pairs from the dataset to avoid erroneous IBD estimation as it is impossible to distinguish between MZ and DZ pairs.

(f) We filtered pairs of individuals with an expected IBD over 60% as these are likely to represent genealogy errors.

(g) We filtered pairs of individuals that died within 10 days from each other. We observed that pairs of individuals that were born in the same town had higher death rates within the same 10 days (fig. S10). This pattern was mostly evident in siblings and first cousins but also even in fourth cousins that were born in the same town (but not fourth cousins that were born far away from each other). This pattern is presumably the result of local environmental hazards such as natural disasters or violent activity. We were concerned that these local hazards can confound the analysis as more related individuals tend to live closer, which can induce spurious association between genetic similarity and longevity. To further validate our hypothesis, we examined the cause of death inserted by the users for siblings that died up to 10 days apart versus siblings that died within the same year (but not 10 days apart). The cause of death field was reported for about 2% of the profiles and was given as a free text based on the genealogists' knowledge. Despite these limitations, we examined the presence of a few potential catastrophes such as plagues and fire in this field between the two conditions. This analysis showed that plagues and fire are reported in significantly higher rates (Fisher exact test, $p < 10^{-15}$, odds-ratio: 11.7×) in the 10 days death pairs versus the rest of the year. This provides an additional evidence that this filtration method is likely to remove environmental catastrophes. Thus, to mitigate the effect of such catastrophes, we removed pairs of individuals that died within 10 days from each other. Indeed, filtering these pairs removed the over-representation of deaths within the same year.

Heritability with nuclear families

We sought to further validate the quality of our data by estimating the narrow-sense heritability of longevity (h^2) according to the mid-parent design. Let i be an individual whose father is f and mother is m. The mid-parent heritability was measured by `lm`, the least-square regression package of `R` to infer α_0 and β of the following model:

$$y_i = \frac{\alpha_0(y_m + y_f)}{2} + \beta$$

h^2 was set to the estimator of α_0.

We analyzed nearly $130,000$ parent-child trios with the highest quality data (exact date of birth, death, and town resolution for place of birth). This process yielded $h^2_{mid-parent} = 12.2\%$ (s.e.=0.4%), which falls within the range of previous heritability estimates, but on the lower end (Table S5).

We also repeated the mid-parent heritability estimates on a narrower range of mid-parent values. We tested the mid-parent heritability when the mid-parent excess longevity deviated between $(-30,30)$ years and $(-20,20)$years. The narrow ranges did not dramatically changed the heritability estimates. With $(-30,30)$, we found that $h^2_{mid-parent} = 12.5\%$ and with $(-20,20)$, we found that $h^2_{mid-parent} = 13.0\%$. As such, these results show that outliers in the mid-parent results cannot explain the lower heritability of our experiments compared to previous studies with nuclear families.

Similar to previous findings (Table S5), the data did not show any significant linear correlation ($p > 0.1$) between heritability and the decade of birth of the parents. For this analysis, the years of birth of the parents were averaged and grouped into bins of 10 years. We employed the mid-parent regression for each bin of data and calculated the maximum likelihood estimate of the heritability and its standard error. Then, we regressed the heritability results on average year of birth of the parents using linear regression with weights that correspond to the inverse of the standard error of each heritability measurement.

We also did not find any significant difference ($p > 0.4$) between the heritability of longevity from mother-offspring pairs ($h^2_{mother} = 12.8\%$, $s.e. = 0.4\%$, $n = 220,000$) versus father-offspring pairs ($h^2_{father} = 13.2\%$, $se = 0.4\%$, $n = 271,000$). For this analysis, The h^2_{mother} and h^2_{father} were calculated similarly to the mid-parent design, by inferring α_0 and β of the following models: $y_i = \alpha_0 y_m + \beta$ and $y_i = \alpha_0 y_f + \beta$, respectively. The reported h^2 was set to $2\alpha_0$ in each model.

We did observe a significant difference ($p < 10^{-11}$) between the heritability of concordant-sex parent-offspring pairs (e.g. mother-daughter) versus discordant-sex pairs (e.g. mother-son), with $h^2_{concordant/parent-child} = 15.0\%$ ($se = 0.4\%$, $n = 254,000$) and $h^2_{discordant/parent-child} = 10.7\%$ ($s.e. = 0.4\%$, $n = 236,000$). Testing whether the difference between $h^2_{concordant}$ and $h^2_{discordant}$ is significant was done by measuring the p-value of the interaction term α_2 in the following model using linear regression:

$$y_i = \alpha_0 y_p + \alpha_1 I(s_i, s_p) + \alpha_2 [y_p \times I(s_i, s_p)] + \beta$$

where individual p is either the mother or father of i, $s_i \in (\text{male}, \text{female})$ and I is an indicator function that returns 1 for $s_i = s_p$ and 0 otherwise.

The significant difference between the concordant and discordant pairs most likely arose due to differences in the distribution of longevity between sexes; women show higher mortality than men around child bearing ages but lower mortality after these ages (fig. S9). Our model only adjusts for the average life expectancy of each sex and we posit that the reduced $h^2_{discordant/parent-child}$ stems from the differences between the distributions.

Adjusting relationships

In order to collect and measure the IBD between relatives, we employed the following steps:

(a) Finding relative pairs: familial ties can be defined by three basic relationships: A is a parent of B, A is a full sibling of B, and A is an offspring of B. For example, C is the uncle of A, if C is the full sibling of B, and B is the parent of A. For each individual that passed the multiple inclusion criteria, we scanned for all the profiles that are related to him by alternating between these three basic steps to identify siblings, parents, grandparents, great-grandparents, uncles, cousins or cousins once removed up to 4th cousins.

(b) Maximizing relationships: in the presence of consanguinity, the genealogical relationships are not unique since a relative can be reached by multiple paths. To address this issue, we assigned the closest possible relationship if more than one path was detected. For example, if A and B could be second cousins or fifth cousins, we assigned them to the second cousin group.

(c) Calculating identity coefficients: another complexity of consanguinity is that the expected IBD between related individuals can be higher than their genealogical relationships (fig. S11).

In addition to assigning pairs of relatives to genealogical classes, we also calculated the expected IBD taking into consideration all potential paths using relatives up to nine generations. For this task, we employed IdCoeff (*63*), which calculates Jacquard's 9 Condensed Coefficients of Identity (fig. S12). These coefficients are an extension of the IBD probabilities and represent all possible configurations of a bi-allelic autosomal site between a pair of individuals, given that the parent of origin does not matter. Since IdCoeff was designed for relatively small pedigrees, we modified the source code to accept topological sorted pedigrees, which removed the $O(n^2)$ pre-processing time of registering the pedigree in the computer memory to $O(n\log(n))$ (code is available from the authors). Running the modified IdCoeff on all possible genealogical pairs of relatives took approximately 25000 hours of CPU time (a month of a server with 15 parallel processes).

(d) Calculating the expected IBD: Let r_{ij} be the IBD probability between individuals i and j in the presence of consanguineous marriages. Then (see *64*):

$$r_{ij} = 2\Delta_1 + \Delta_3 + \Delta_5 + \Delta_7 + \frac{\Delta_8}{2}$$

where Δ_i is the i-th identity coefficient for the i, j pair according to the IdCoeff output.

Measuring dominance variance

Throughout this manuscript, the phenotypic correlation is defined as:

$$\text{Corr}(y_i, y_j) \triangleq \frac{y_i y_j}{\sigma_i * \sigma_j}$$

where y_i, y_j are the longevity of individual i and j, respectively, and σ_x is the standard deviation of all individuals in class x.

The Δ_7 allelic configuration mediates the variance of v_d and mainly appears in full sibs (fig. S13). To estimate this factor, we sought to compare full-sibs to parent-child pairs, in which $\Delta_7 \approx 0$. This comparison is immune to confounders due to additive and epistatic factors as the average IBDs of both relative groups were the same with $\mathbb{E}_{sibs}(r) = 0.5010$ and $\mathbb{E}_{parent-child}(r) = 0.5010$, and therefore, should mediate on average similar additive and epistatic vavriance. To further reduce potential biases due to sex differences, we limited the analysis only to males, resulting in $157,600$ father-son pairs and $159,000$ brother pairs.

However, one potential caveat in this setting is that brothers can be more correlated due to transient environmental effects, such as local catastrophes that are unaccounted in our model. To illustrate this, consider a family of a father and two sons of ages 70, 35, 30 years old, respectively that live in the same town. In the case of an extrinsic catastrophe that kills all of these individuals, the two siblings will be more correlated than the father-son pairs. Importantly, the correlation difference simply reflects the fact that the two sibs were born in similar years compared to their father and has nothing to do with dominance. This excess in correlation is not unique to our data and was documented in previous studies. For example, a classic paper by Rao et al. (*65*) inspected the correlation of height and weight in nuclear families. Despite adjustments of the phenotype to the age of the child, they found a strong decline in the phenotypic correlation as a function of the difference in year of birth between siblings.

To address that, we tested the two models using non negative least square regression based on the `R` package `nnls`:
Model 1: $\text{Corr}(y_i, y_j) = \Delta_7(i,j)v_d + \beta$ (subject to $v_d \geq 0$)
Model 2: $\text{Corr}(y_i, y_j) = \Delta_7(i,j)v_d + \Delta\text{gen}(i,j)\alpha + \beta$ (subject to $v_d \geq 0$ & $\alpha \leq 0$)

where Δgen is the absolute difference in the year of birth of the two profiles, which can reflect changes in the environment over time. The first model estimated $v_d = 24\%$. The second model estimated $v_d = 4\%$ and

$\alpha = -0.0018$.

Several pieces of evidence support the second model: first, the $\text{BIC}(\text{model1}) - \text{BIC}(\text{model2}) = 41$, suggesting that Model 2 is more appropriate. Second, we also regressed:

$$\text{Corr}(y_i, y_j) = \Delta\text{gen}(i,j)\alpha + \beta$$

but this time only with pairs of brothers. Again, we found that $\alpha = -0.0015$ and was statistically different than 0 ($p < 0.0005$). This shows that Δgen explains similar phenotypic correlation to Model 2 while v_d is nearly fixed. Third, Model 1 estimated $v_d = 0.24 \pm 0.03$. This implies that the entire correlation observed in the MZ twins (0.24) is attributed to dominance variance. However, the heritability of sex-concordant parent-offspring pairs was estimated to $h^2_{concordant} = 0.21 \pm 0.01$. This estimator is robust to dominance variance, suggesting that v_d was largely overestimated by Model 1. Therefore, we used the maximum likelihood of Model 2 and estimated the that the dominance variance of longevity is around 4%.

Model fitting for epistatic interactions

We evaluated the following three nested models:

$$\text{Corr}(y_i, y_j) = \beta + v_1 r_{ij} + v_d \Delta_7 \qquad (n = 1)$$
$$\text{Corr}(y_i, y_j) = \beta + v_1 r_{ij} + v_2 r_{ij}^2 + v_d \Delta_7 \qquad (n = 2)$$
$$\text{Corr}(y_i, y_j) = \beta + v_1 r_{ij} + v_2 r_{ij}^2 + v_3 r_{ij}^3 + v_d \Delta_7 \qquad (n = 3)$$

These models are based on Kempthorne analysis of variance components for high order interactions in outbred populations (*66*). Fourth cousins, which display almost zero IBD probabilities, showed longevity correlation of 2.0% that was significantly different than 0 ($p < 10^{-24}$). This correlation might be due to the fact that fourth cousins live much closer to each other (median distance: 75km) than complete random pairs of individuals in our data (median: ˜2000km). β was included in the models to allow positive longevity correlation in these far related individuals.

The models were fitted to the data using non negative least squares with the `R` package `nnls`, forcing $\beta, v_1, \ldots, v_3 \geq 0$. The value for v_d was set to 4.0 according to the results above. For testing whether the model was statistically significant, we used nested ANOVA. Both the ANOVA and the BIC calculation were done using custom functions written in `R` by the authors. For cross-validation, we removed each class of relatives in Table S4 and fitted the three models with the remaining data. Then, we calculated the mean squared error based for the C class by:

$$e_n(C) \triangleq \frac{\sum_{i,j \in C}[Y_{ij} - m_n(r_{ij})]^2}{N}$$

where Y_{ij} is the observed longevity correlation for the i, j pair that is part of the C class. $m_n(r_i)$ denotes the longevity correlation prediction of the n model (e.g. n=1) for the i, j pair based on their IBD readout, which is equal to r_{ij}. N is the number of pairs in the class. The total MSE of the k-th model was set to $\text{MSE}_k = \sum_{c \in C} \frac{e_k(C)}{k}$.

For consistency with the calculation of dominancy, we also repeated the regression of the three configurations with Δgen (difference in age of birth) as an additional covariate. This procedure had no effect on our conclusion. The epistatic terms were not significantly different than zero, did not improve the MSE, and still showed a large difference in their BIC values compared to the additive model.

The Danish twin pairs underwent the similar adjustment and pre-processing steps as the Geni data. Namely, we converted the twin age of death to longevity using the same model that takes into account sex, year of birth, and the geolocation of birth. As we did not know the exact geo-location, we set the birth place to Copenhagen, the most populated metropolitan in Denmark. We note that due to the small size of this country (without Greenland) assignments to other coordinates with Denmark has virtually no effects on the life expectancy estimation of our model. In addition, similar to the Geni relatives, we removed MZ-twins that died before age of 30 or during WWI or WWII. Finally, we adjusted the twin correlation to reflect the measured v_d.

The only difference from the Geni processing steps was that in the absence of access to the exact dates of death, we could not filter individuals that died within 10 days of each other. This caveat could create a slight increase in the longevity correlation of MZ twins compared to other pairs in our. Such bias would be in favor for epistatic models and therefore does not affect our conclusion about the role of additivity.

MZ prediction was measured by evaluating each of the $n = 1,2,3$ models in $r_{ij} = 1$.

To address the correlated data points in our analysis, all standard errors of the model estimators are based on a bootstrapping technique using 100 iterations.

We also repeated the analysis of the three models assuming $v_d = 0$. This process yielded slightly higher additivity of 17.9% but the epistatic terms converged to zero and where disfavored by BIC and MSE analysis. All three models were 3% short of predicting the MZ twin correlation.

Addressing potential environmental confounders and more complex models

1. **Adjusting household effects**

   The correlation of longevity between relatives can be induced to shared household effect such as access to the health care system, socio-economic status, or dietary habits. To mitigate these effects, we adjusted the life expectancy of an individual using data from their spouse similar to (*35*). The idea is that if the spouses are not related to the individual of an interest, than the correlation of longevity in couples includes shared household effects. To best fit the data, we extended our life expectancy model and considered the following configurations in R:

   ```
   # sex, birth year, and the spouse life span:
   model_s1 <-bam(long1~ birth_year + sex1+long2, samfrac=1, data=training_set)

   # sex, birth year, the birth geolocation (with splines), and the spouse excess
   ```

```
    longevity
model_s2 <-bam(long1 ~ birth_year + sex1+ s(lat1,lon1,bs="sos",k=1000) + ex_s
    ,samfrac=1, data=training_set)

# sex, birth year, the birth geolocation (with splines), and the spouse life span
model_s3 <-bam(long1 ~ birth_year+ sex1+ s(lat1,lon1,bs="sos", k=1000)+ long2,
    samfrac=1, data=training_set)

# sex, birth year, birth geolocation (this is the previous model and it was
    evaluated as a control)
model_s4 <-bam(long1 ~ birth_year+ sex1+ s(lat1,lon1,bs="sos",k=1000) , samfrac=1,
    data=training_set)

# the sex, birth year, and spouse life span:
model_s5 <- bam( long1 ~ birth_year+ sex1+ ex_s, samfrac=1, data=training_set)
```

The data consisted only of couples that are not genetically related according to our tree data to avoid confounding the household correlation with genetic correlation. In addition, we only considered couples that both people died after age of 30. Finally, if the individual remarried to another person, we averaged the life expectancy of their spouses before calculating the model. The parameters were inferred using 90% of the data (approximately 2.5 million data points) and the evaluation was done with the other 10% of the data points.

We found that model_s3 gave the best results (fig. S14). It has the lowest BIC value, the lowest mean squared error in the test set, and explained the highest fraction of the longevity variance. In addition, all of the covariates were statistically significant according to ANOVA ($p < 0.01$). In overall, spouse longevity helps to explain an additional 1% (s.e. = 0.2%) of the longevity variance compared to the original model that just considered sex, year of birth, and geolocation. These results are consistent with previous longevity studies in twins (*32*) and nuclear families (*35*) that also estimated a similar contribution of the shared sizable household effect.

The decomposition of genetic variance of the spouse-adjusted longevity (n=450,000) had no effect on our conclusions. Similar to our previous setting with all the data and longevity without spouse-adjustment, additivity explained $h^2_{concordant/relatives} = 16.0\%$ (s.e.= 2.2%) of the variance, dominancy explained 3% (s.e. = 5.9%), and the epistatic values converged to zero (quadratic s.e. = 0.1%, cubic s.e = 7.4%). Again, BIC and RSE estimates disfavored the epistatic terms. We did not compare this condition to the Danish MZ twin dataset since we could not apply the same spousal correction for the Danish MZ twins.

The only difference was that the new model estimated a lower dominance variance $v_d = 2.3\%$ to longevity based on the difference in the longevity correlation in sib pairs versus parent-offspring. The lower observed dominance is not surprising. In the the current analysis, the father's phenotype includes an adjustment based on the longevity of his wife, which shares 50% of her genome with the offspring. Thus, the parent-offspring correlation is expected to be higher. Since dominance variance is calculated as the difference between the correlation of siblings minus the correlation of parent-offspring, the outcome is expected to be lower. In any case, we also evaluated the three nested models (additivity, epistasis of two genes, and epistasis of three genes) after adjusting the the dominance variance to 4% and using the spouse-adjusted phenotypes. The results did not change and the epistatic terms converged to zero.

2. **Only female lines** Our results with the Y chromosome and mitochondria data showed that our pedigree has a low percentage of errors. We wondered whether our results are affected by these errors. To address this issue, we focused on relatives that are purely due to shared maternal lines, because these edges in our graph are much more reliable (0.3% error rate versus 1.9%).

   We scanned all pairs of relatives and find the shortest genealogical path between them. Next, we filtered any pair of relatives that the intermediate people along their genealogical path did not include females. This process resulted with over $300,000$ pairs of relatives.

   Finally, we fitted the three nested Kempthorne models to the data. Again, the results were highly similar to the full model that included pairs of relatives due to male lines. Additivity explained 16.9% (s.e.= 2.2%), the epistatic values converged to zero (quadratic s.e.= 0.5%, cubic s.e=6%), and all model diagnostic strongly disfavored any epistatic interactions.

3. **No inbred lines** Abney et al. (*64*) showed that a population that is the product of consanguineous marriages contains additional components of variance that can contribute to the phenotypic correlation: v_h, $\text{Corr}_h(a,d)$, and μ_h^2.

   v_h denotes the variation component due to sharing two pairs of identical alleles between a pair of relatives. $\text{Corr}_h(a,d)$ denotes the correlation of the additive and dominant effects. μ_h^2 denotes the inbreeding depression in homozygous individuals. The phenotypic correlation of a relative pair due to dominancy in an inbred population is:

$$\text{Corr}(y_i,y_j) = v_1 r + \Delta_7 v_d + \Delta_1 v_h + (4\Delta_1 + \Delta_3 + \Delta_5)\text{Corr}_h(a,d) + (\Delta_1 + \Delta_2 - f_a f_b)\mu_h^2$$

   where f_a and f_b denote the inbreeding coefficient of each individual in the pair.

   While the average values for Δ_1 to Δ_6 are extremely small in our pedigree ($< 0.1\%$), we sought to evaluate epistatic interactions in the absence of such confounders. For that, we retained only pairs of individuals whose $\Delta_1 + \ldots \Delta_6 = 0$. With this process, we still had over 2.6 million pairs of relatives to evaluate our models.

   We did not find any difference in the results between this outbred data points and the model with the 3.2 million relatives that included also inbred relatives. Additivity explained 15.8% (s.e.=4.2%) and the epistatic terms converged to zero (quadratic s.e.= 0.1%, cubic s.e=0.1%).

Another method to measure the contribution of shared household environment in longevity

We also sought to measure the shared household environment from a different angle. To this end, we analyzed the longevity correlation in (a) pairs of same-sex individuals whose respective spouses are siblings and (b) pairs of same-sex individuals whose spouses are first cousins. For simplicity, we will dub the former pairs as Class I and the latter pairs as Class II. For example, consider a scenario in which Alice married Bob, Bob and Casey are brothers, and that Casey and Dora are married. Alice and Dora represent class I pairs (pair of individuals whose spouses are sibling). Similarly if Bob and Casey were first cousins, then Alice and Dora were considered as case II pairs (pair of individuals whose spouses are first cousins) (see S16A under ”Household” for graphical representation).

We filtered the pairs according to the previous steps (no deaths during major wars, only sex-concordant pairs, etc). We also filtered any pairs that are genetically related (up to fourth cousins). For example, if Alice and Dora happen to be third cousins, we removed them from the analysis. We were concerned that Class II could show lower correlation due to larger geographic distances of first cousins. To address this issue, we measured the distance in the place of birth of the kids of each of these pairs. For example, if Alice gave birth to Adam and Dora gave birth to Diane, we measured the distance of the place of birth of Adam and Diane. For that two classes, we only retained pairs whose kids were born at the same place to reduce differences in the phenotypic correlations due to geography. Our hypothesis was that child birth better represents the environment of individual later in life and therefore would help to capture additional environmental correlations on top of our adjustment to the place of birth for life expectancy.

After controlling for these potential environmental correlations, the extra longevity correlation of Class I over Class II can represent:

1. Similar household environments due to the extra correlation in the household of sibs versus first cousins.

2. Correlations due to assortative mating patterns. For example, if Bob is tall then it is likely that his brother Casey is also tall since the heritability of height is 80%. Due to assortative mating, it implies that both Alice and Dora are also likely to be taller than average. However, since the correlation of height in cousins is likely to be smaller, then we expect that correlation due to assortative mating with respect to height should be smaller in cousins. Similarly, if there are certain factors of longevity that play a role in assortative mating, they are likely to be more correlated in Class I pairs than Class II pairs.

3. Correlations due to grieving effects. Previous research have reported an increased likelihood of a person to die after the death of their spouse (*67*) or their sibling (*68*). For example, if Bob and Casey are brothers, the death of Bob can affect Alice and Casey, which could eventually affect Dora. We hypothesize that these effects are less likely between first cousins. Therefore, they should contribute to the excess of correlation in Class I over Class II.

Taken together, the difference between the longevity correlation of Class I pairs and Class II represents the net effect of the three conditions above. Since these conditions are likely to create non-negative correlations, the difference between the class is likely to reflect an upper bound of the shared household environment.

We found that Class I had only 1.5% additional correlation in their longevity than Class II, which is only tenth than the measured additive effects. This likely upper bound argues against strong shared household effects in longevity.

It is also important to note that household effects are more likely to inflate the correlation of close relatives in our model, such as sibs and half-sibs. As such, they can create spurious epistasis signal rather than inflating the additive signal. Therefore, our results regrading absence of epistasis are unlikely to be affected by if unaccounted household effects still exist.

Restricted maximum likelihood estimation of the heritability of longevity

The linear regression method to derive the additive versus epistatic components is basically an extension of the Haseman-Elston (HE) linear regression framework. Previous work has shown the equivalence between HE and Linear Mixed Models (LMM) estimators (*69*,*70*). However, one concern is that HE only uses pairs of individuals to make the estimates and not the entire data structure that has complex inter-dependencies that can improve the prediction and provide better power.

To further validate our results, we also estimated the heritability of longevity via LMM using a restricted maximum likelihood (REML) estimation procedure. REML estimation requires first estimating matrices of kinship coefficients between every pair of individuals, and then estimating the heritability via an LMM which uses these matrices. However, both tasks are complicated due to a matrix inversion step that has an $\mathcal{O}(n^3)$ complexity, which is computationally impractical for our data with $500,000$ individuals.

To overcome the computational complexity of regular LMM, we developed a technique called sparse Cholesky factorization linear mixed model (Sci-LMM). Sci-LMM takes advantage over the high sparsity in the relatedness matrix and uses sparse Cholesky factorization to circumvent the need to invert the matrix (*43*). The algorithm is available on GitHub under GPLv3 (`https://github.com/TalShor/SciLMM`).

In the next paragraphs, we will briefly introduce Sci-LMM and the results using the Geni data for longevity.

The kinship coefficient of two individuals is defined as the probability that two alleles drawn from the two individuals are identical by descent. The kinship coefficient is given by $0.5\sum_P(1+f_P)2^{-n_P}$ (*71*). Here, P iterates over individuals that are most recent common ancestors of the two individuals (meaning that there is at least one path of direct family relationships between the two individuals passing through P which does not pass through any other most recent common ancestor), f_P is the kinship coefficient of the parents of P, and n_P is the length of the path connecting the two individuals.

To compute the matrix of kinship coefficients efficiently, we used the technique of Henderson (*72*, *73*) and one of its extensions (*74*). Briefly, this technique decomposes the kinship coefficient matrix A into

$A = LHL^T$, where L is a lower triangular matrix such that L_{ij} contains the fraction of genome sharing between individuals i and j, H is a diagonal matrix containing the within-family additive genetic variance of individuals, and the matrices are ordered such that ancestors precede their descendants. These matrices can be stored efficiently via sparse matrix routines, and can be computed efficiently via dynamic programming.

Using these techniques, we were able to compute the matrix A in less than 10 hours.

We estimated the heritability of longevity via an LMM, using a restricted maximum likelihood (REML) technique. Briefly, we assume that the vector y of longevity records for n individuals follows a multivariate normal distribution:

$$y \sim \mathcal{N}(C\beta\,; \sigma_A^2 A + \sigma^2{}_{A\times A} A^2 + \sigma_e^2 I). \tag{1}$$

Here, $C \in \mathbb{R}^{n\times c}$ is a matrix of c covariates, $\beta \in \mathbb{R}^c$ is a vector of fixed effects, $A \in \mathbb{R}^{n\times n}$ is a a matrix of kinship coefficients, $I \in \mathbb{R}^{n\times n}$ is the identity matrix, σ_A^2, $\sigma_{A\times A}^2$ denote the additive variances and the squared epistasis variance, respectively, and σ_e^2 is the variance of the non-shared environmental effects of y. We are interested in finding the REML of β and of σ_A^2, $\sigma_{A\times A}^2$, σ_e^2.

Unfortunately, a straightforward (restricted) maximum likelihood estimation in LMMs requires inverting $n\times n$ matrices. This operation scales cubically with n, which renders it computational infeasible for samples with hundreds of thousands of individuals as in the current study. Fortunately, the matrices A and A^2 are extremely sparse, which enables efficient computations. Specifically, denoting $V = \sigma_A^2 A + \sigma_{A\times A}^2 A^2 + \sigma_e^2 I$ as the combined Sci-LMM covariance matrix, the Cholesky factorization of V can be computed efficiently using the CHOLMOD software package (*75*). We additionally employ a sampling-based approximation, as proposed in (*76*, *77*).

For convenience, we first reparameterize the LMM as follows. First, we define $\sigma_g^2 = \sigma_A^2 + \sigma_{A\times A}^2$, $\mu = \frac{\sigma_A^2}{\sigma_g^2}$. Next, we define $B = \mu A + (1-\mu)A^2$. Following (*78*, *79*), we introduce the variances ratio $\delta = \frac{\sigma_e^2}{\sigma_g^2}$. Finally, we define $U = B + \delta I$. The Sci-LMM covariance matrix is therefore given by $V = \sigma_g^2 B + \sigma_e^2 I = \sigma_g^2 U$.

To find the REML efficiently, we perform a grid search over μ, δ. The other model parameters can be computed efficiently given these two parameters, as we now demonstrate.

The non-restricted and restricted LMM likelihood are given by (*79*):

$$\begin{aligned} \ell(\beta,\mu,\sigma_g^2,\delta) &= -\frac{1}{2}(y - C\beta)\left(\sigma_g^2 U\right)^{-1}(y - C\beta) - \frac{1}{2}\log\left|\sigma_g^2 U\right| - \frac{n}{2}\log(2\pi). \\ \ell_R(\beta,\mu,\sigma_g^2,\delta) &= \ell(\beta,\sigma_g^2,\delta) + \frac{c}{2}\log(2\pi) + \frac{1}{2}\log\left|C^T C\right| - \log\left|C^T V^{-1} C\right|. \end{aligned} \tag{2}$$

Following (*78*, *79*), the fixed effects estimator given the other parameters is given by:

$$\hat{\beta} = \left(C^T U^{-1} C\right)^{-1} C^T U^{-1} y. \tag{3}$$

This estimator can be computed efficiently by using the CHOLMOD routines.

The derivative of $\ell_R(\sigma_g^2)$ with respect to σ_g^2 is given by:

$$\begin{aligned}\frac{\partial \ell_R(\sigma_g^2)}{\partial \sigma_g^2} &= \frac{\partial \ell(\sigma_g^2)}{\partial \sigma_g^2} + \frac{1}{2}\text{tr}\left(\left(C^T V^{-1} C\right)^{-1} C^T V^{-1} \frac{\partial V}{\partial \sigma_g^2} V^{-1} C\right) \\ &= \frac{\partial \ell(\sigma_g^2)}{\partial \sigma_g^2} + \frac{1}{2}\sigma_g^{-2}\text{tr}\left(\left(C^T U^{-1} C\right)^{-1} C^T U^{-1} B U^{-1} C\right).,\end{aligned} \tag{4}$$

where $\frac{\partial \ell(\sigma_g^2)}{\partial \sigma_g^2}$ is given by:

$$\begin{aligned}\frac{\partial \ell(\sigma_g^2)}{\partial \sigma_g^2} &= \frac{1}{2}\tilde{y}^T V^{-1} \frac{\partial V}{\partial \sigma_g^2} V^{-1} \tilde{y} - \frac{1}{2}\text{tr}\left(V^{-1}\frac{\partial V}{\partial \sigma_g^2}\right) \\ &= \frac{1}{2}\tilde{y}^T \left(\sigma_g^{-2} U^{-1} B \sigma_g^{-2} U^{-1}\right)\tilde{y} - \frac{1}{2}\sigma_g^{-2}\text{tr}\left(U^{-1}B\right),\end{aligned} \tag{5}$$

and where $\tilde{y} = y - C\hat{\beta}$.

By setting $\frac{\partial \ell_R(\sigma_g^2)}{\partial \sigma_g^2}$ to 0 and doing some algebra, the REML of σ_g^2 is given by:

$$\hat{\sigma}_g^2 = \frac{\left(U^{-1}\tilde{y}\right)^T B \left(U^{-1}\tilde{y}\right)}{\text{tr}\left(U^{-1}B\right) + \text{tr}\left(\left(C^T U^{-1} C\right)^{-1} C^T U^{-1} B U^{-1} C\right)}. \tag{6}$$

The term $\text{tr}\left(U^{-1}B\right)$ can be computed via the Monte Carlo routine described in (*76*). All the terms in the second term of the denominator are easy to compute, assuming that the number of covariates c is very small compared to n. Finally, σ_e^2 is trivially given by $\sigma_e^2 = \delta\sigma_g^2$. Hence, we estimate the REML by performing a two dimensional grid search over μ in the range [0,1] and over $\exp(\delta)$ in the range [-5, 5].

We estimate the estimator variance via the diagonal of the inverse Hessian of the restricted log likelihood, which can be approximated efficiently via the average information technique (*76*). Specifically, we use the following approximation for every pair of parameters θ_i, θ_j:

$$\frac{\partial \ell}{\partial \theta_i \theta_j} \approx -\frac{1}{2} y^T P^{-1} \frac{\partial V}{\partial \theta_i} P^{-1} \frac{\partial V}{\partial \theta_j} P^{-1} y \tag{7}$$

where $P = V^{-1} - V^{-1}C\left(C^T V^{-1} C\right)^{-1} C^T V^{-1}$ is the Sci-LMM covariance matrix after regressing out the covariates.

To summarize, our analysis included two variance components corresponding to kinship and squared kinship matrices and covariates corresponding to sex, year of birth, and the top 20 principal components of the kinship coefficients matrix. The average non negative entry in the kinship matrix was 9.47×10^{-3}. We used pairs of individuals up to 40 meioses away from each other.

The resulting additive heritabiltiy was 17.8% (s.e. 0.84%) and the epistatic component estimated to be 1.6% (s.e. 1.7%), which is not significantly different than zero ($p = 0.17$). We could not measure the dominance component using this approach due to strong collinearity to the identity matrix. Therefore, we speculate that the slight increase of the maximum likelihood of the epistatic values can also be explained by lack of correction of dominance variance.

Analyzing familial dispersion

For all migration events, we only used data from individuals that we had an exact date of birth AND high quality birth location as defined above. All of the results below are available as a single R script from the authors.

Measuring migration distance

The migration distance corresponds to the great circle distance between the birth location of each pair of profiles. We transformed the distance to log scale according to $\log_{10}(1+x)$, where x is the distance in km. The year of birth was averaged between the pair of individuals. The migration distance of males was defined as the distance between the birth locations of father-offspring pairs. The migration distance of females was defined as the distance between the birth locations of mother-offspring pairs. The marital radius was defined as the distance between the birth locations of spouses. The plots were smoothed using a rolling average with a window of ten years. The raw data before smoothing can be seen in fig. S18 - S19.

Profiles from Europe were defined as profiles that were born in the following country codes (according to current political borders): ’AL’, ’AT’, ’BA’, ’BE’, ’BG’, ’CH’, ’CY’, ’CZ’, ’DE’, ’DK’, ’EE’, ’ES’, ’FI’, ’FO’, ’FR’, ’GB’, ’GI’, ’GR’, ’HR’, ’HU’, ’IE’, ’IS’, ’IT’ ,’LT’, ’LU’, ’LV’, ’MC’, ’MK’, ’NL’, ’NO’, ’PO’, ’PT’, ’RO’, ’SE’, ’SI’, ’SK’, ’SM’, ’VA’.

Profiles from North America were defined as profiles that were born in the following country codes (according to current political borders): ’US’, ’CA’.

Measuring identity by descent between couples

To measure identify by descent, we employed the procedure in section ”Adjusting relationships” to individuals that were married to each other. We only included individuals that are non-founders. The plot was smoothed using a rolling average of two years. To test isolation by distance, we regressed the average identity by descent per year on the martial

Fusing datasets with FamiLinx

The public data in FamiLinx includes the profile-id of each person in our database (without the names). Our data use agreement allows not-for-profit researchers to consent participants in order to obtain their Geni profile-id to identify them in our the FamiLinx data. Such usage is conceptually similar to the collaborative nature of Geni.com. However, it is important to note that we strictly prohibit any re-identification without the consent of the participant.

To simplify the collection of the Geni profile-id, we constructed a simple web button that researchers can integrate into their website. The button is based on client side Java script and uses the Geni SDK. If the user consent to contributing their Geni profile, they can click on the button. This sends a signal to the Geni.com server and creates a pop-up for the user that asks for their Geni username and password over secure HTPPS communication. Users without a Geni profile can create a new profile as part of this process. Importantly, this pop-up window is served directly from Geni and the researcher cannot see this transaction. Next, the user's client receives a Json message from Geni that includes the profile-id and some basic account information. Our client side script transmits this message to the server (the researcher), which can now register this information along with other information form the user (e.g. genome, phenotypes, etc...). Then, the researcher can search the FamiLinx tree for the user profile-id and overly the phenotypic or genomic information with the tree structure.

Fig. S21 shows the sequence of events and the experience of the user. We tested and validated this button in DNA.Land and were able to collect profile-ids of over one thousand individuals using this process. Researchers that are interested to test this method can visit:

https://teamerlich.org/geni-integration-example/.

The code is available on: https://github.com/TeamErlich/geni-integration-example.

Supplementary Figures

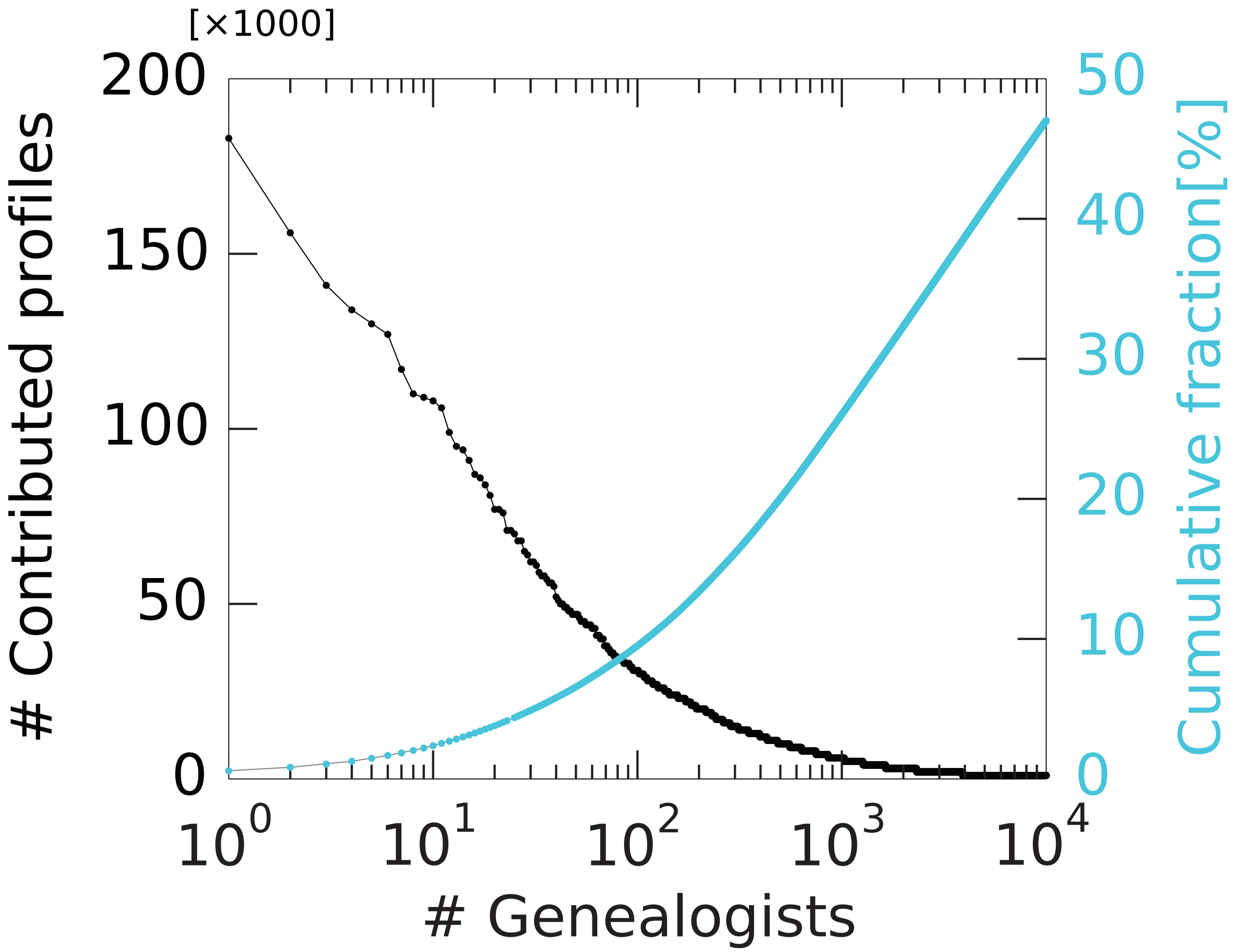


Figure S1: The inferred contribution of profiles by the top ten thousand genealogists (black: number of profiles contributed by each genealogist sorted based on their contribution; light blue: the cumulative distribution of profiles with a known genealogist).

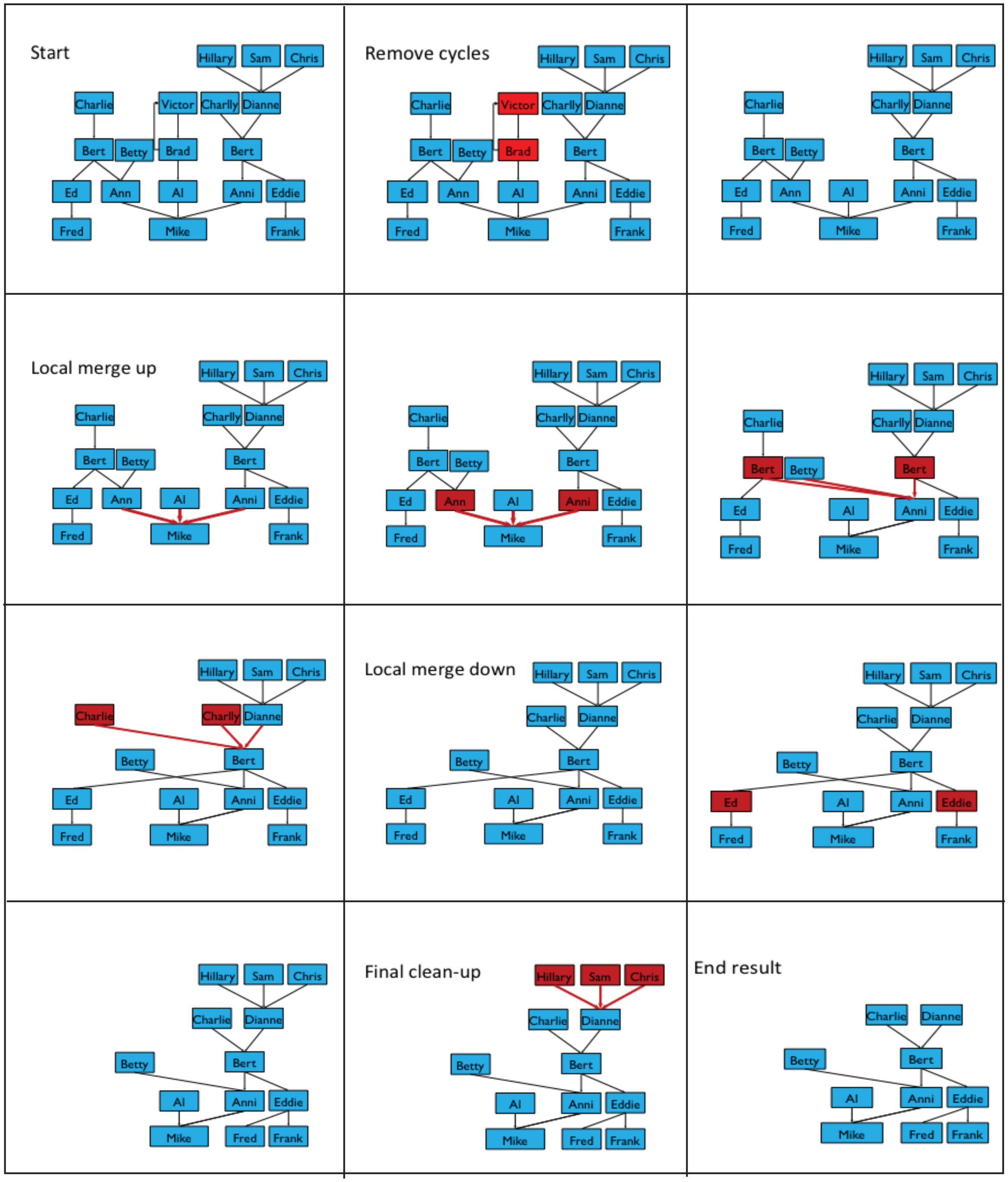


Figure S2: An illustration of the steps taken to clean the Geni graph. Union nodes are not shown for clarity.

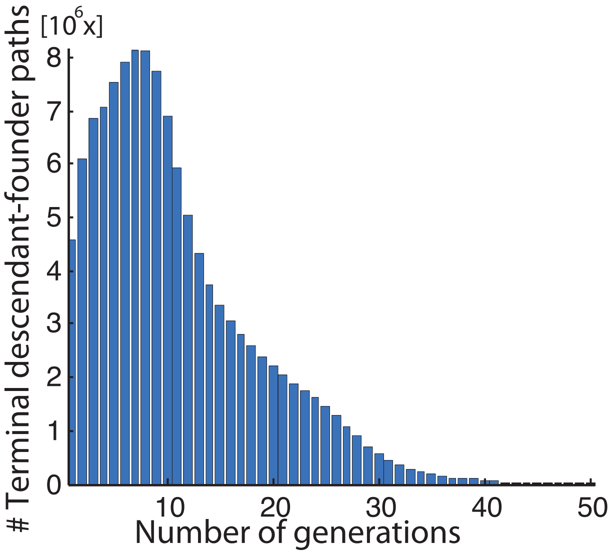


Figure S3: The distribution of number of generations between terminal descendants and founders.

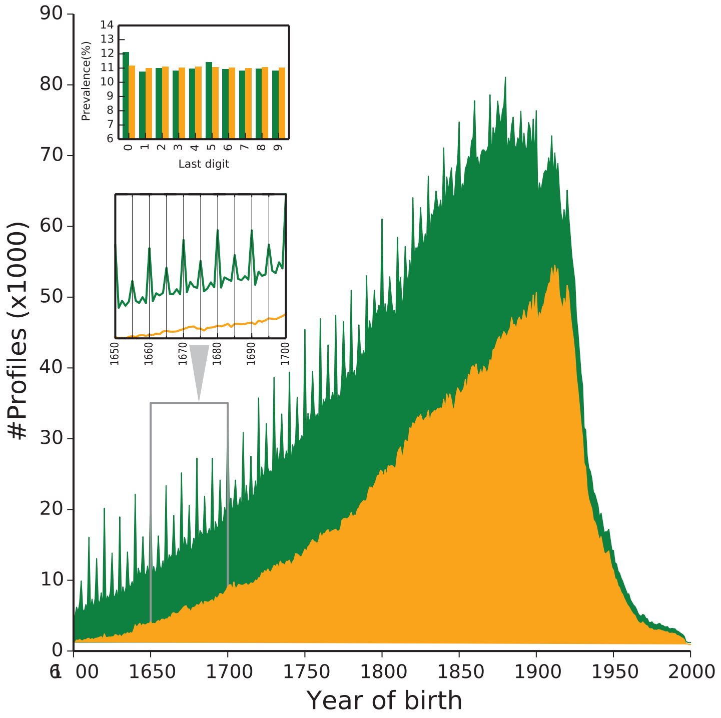


Figure S4: Distribution of the profiles' year of birth since the 16th century. Without filtering (green), round decades are overrepresented, suggesting imprecise data. Retaining profiles with exact dates (yellow) removes this pattern.

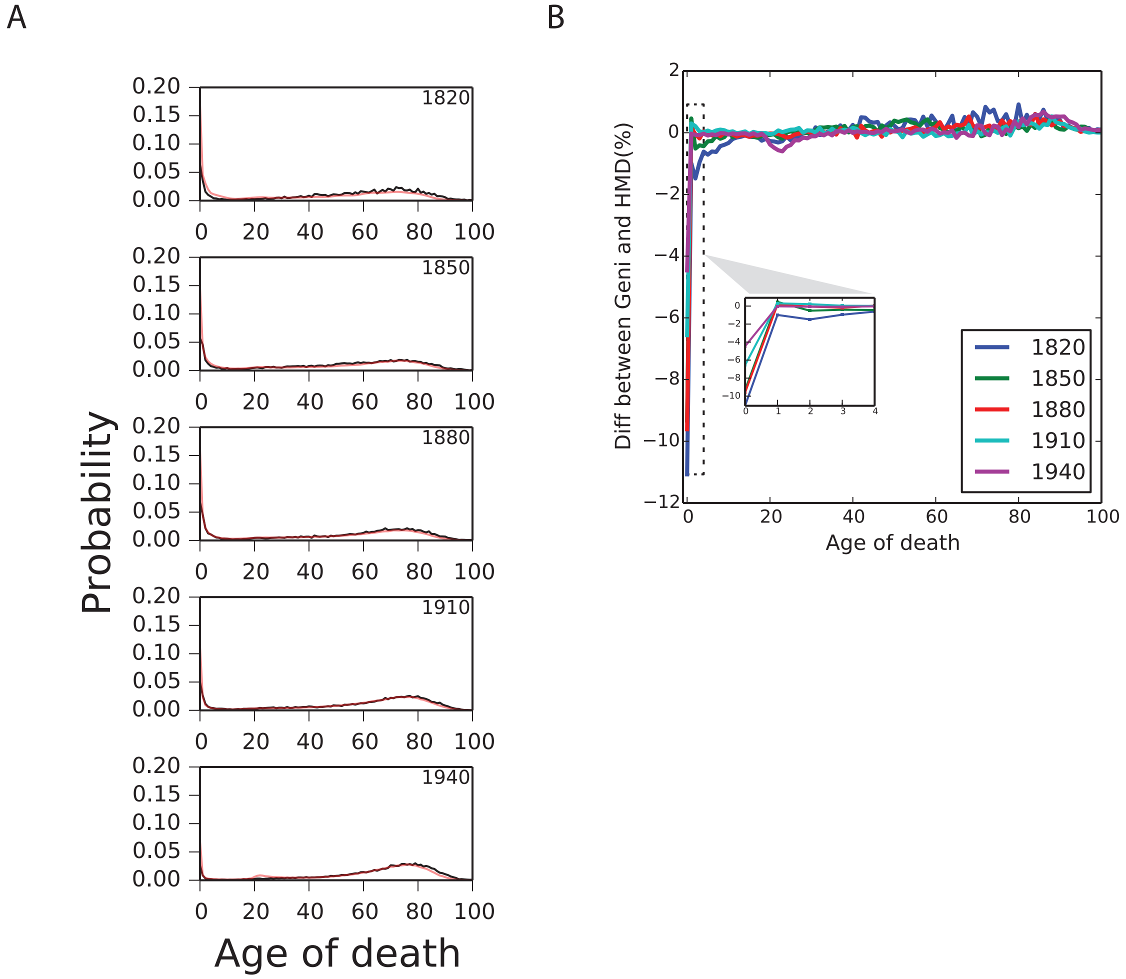
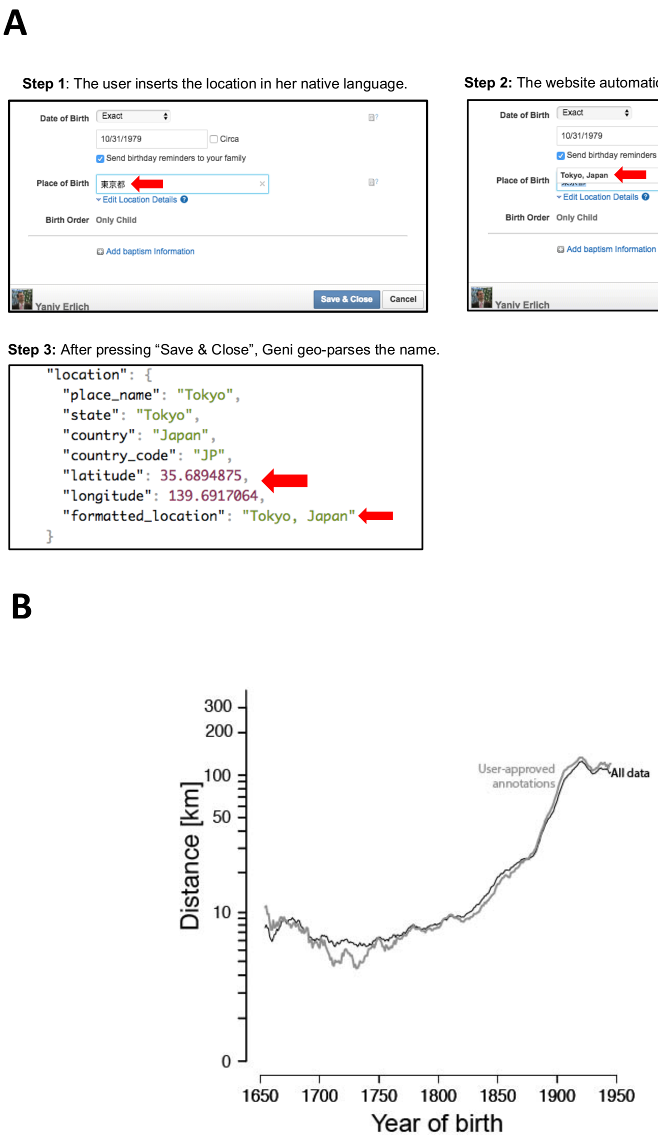


Figure S5: Comparing Geni to HMD (A) The age of death probability density function (PDF) in Geni (black) versus HMD (red) for 1820-1940 (B) The difference between the HMD and Geni data. Each curve represents the subtraction of the HMD's age of death PDF from the Geni one. The only systematic difference is underestimation of infant deaths in Geni.



Step 1: The user inserts the location in her native language.

Date of Birth: Exact
10/31/1979 □ Circa
☑ Send birthday reminders to your family
Place of Birth: 東京都
Edit Location Details
Birth Order: Only Child
Add baptism Information

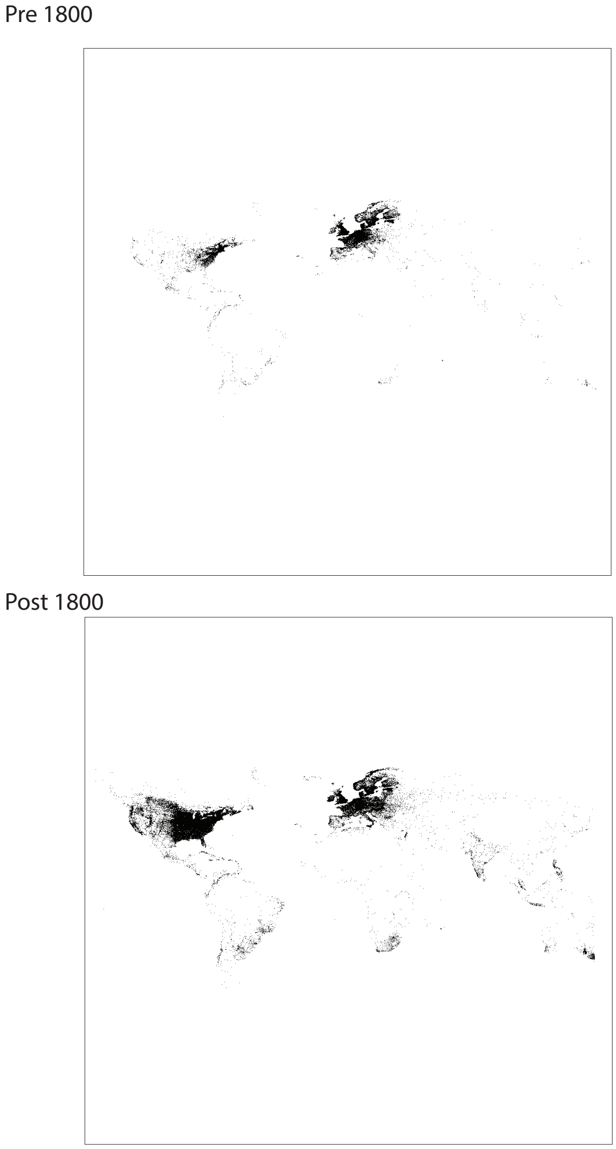
Step 2: The website automatically converts the name to its English form.

Date of Birth: Exact
10/31/1979 □ Circa
☑ Send birthday reminders to your family
Place of Birth: Tokyo, Japan
Edit Location Details
Birth Order: Only Child
Add baptism Information

Step 3: After pressing “Save & Close”, Geni geo-parses the name.

```
"location": {
  "place_name": "Tokyo",
  "state": "Tokyo",
  "country": "Japan",
  "country_code": "JP",
  "latitude": 35.6894875,
  "longitude": 139.6917064,
  "formatted_location": "Tokyo, Japan"
}
```

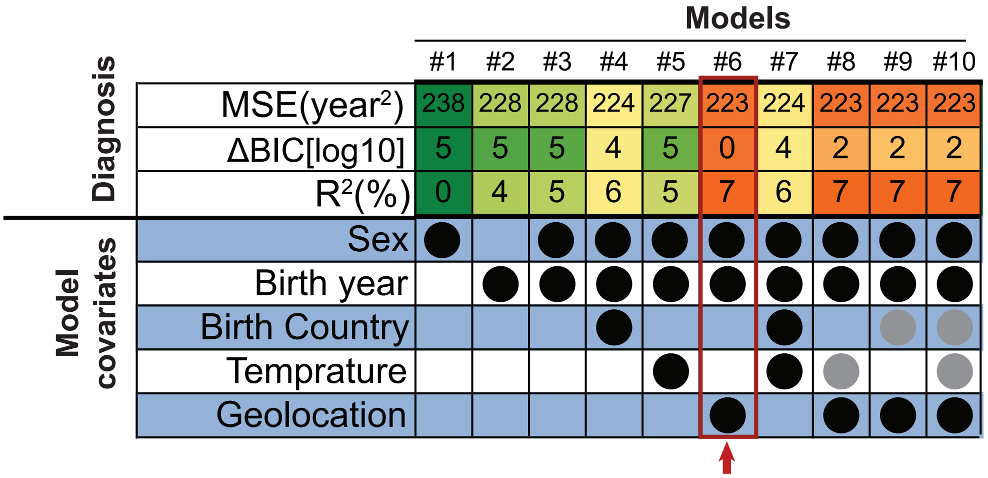
Figure S6: Assessing biases in migration using user approved annotations (A) User-approved annotations start with the user reporting an event in her own native language. Immediately after inserting the text, the website offers a standard text in a canonical format in English. Upon approval, the standard text and longitude and latitude are saved in a JSON format and accessible via the Geni API (B) Median martial distance using all data (automatic geoparsing and user-approved annotations; black) as presented in the main text versus the same analysis with user-approved annotations (gray) that are expected to have higher quality.



Pre 1800

Post 1800

Figure S7: Geographic distribution of Geni profiles pre and post 1800.



		Models									
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Diagnosis	MSE(year^2)	238	228	228	224	227	223	224	223	223	223
	ΔBIC[log10]	5	5	5	4	5	0	4	2	2	2
	R^2(%)	0	4	5	6	5	7	6	7	7	7
Model covariates	Sex	●		●	●	●	●	●	●	●	●
	Birth year		●	●	●	●	●	●	●	●	●
	Birth Country				●			●		◍	◍
	Temprature					●		●	◍		◍
	Geolocation						●		●	●	●

Figure S8: Adjustment of longevity using various environmental models. MSE is the mean squared error per individual; ΔBIC is the difference of the Bayesian Information Criterion from the best model after $\log_{10}(x+1)$ transformation. Sex is a two level factor (male/female). Temperature is average and standard deviation of the temperature at place of birth, as obtained from WorldClim in 2.5min resolution. Geolocation is the longitude/latitude location of birth and was modeled using splines on spheres. Closed circles represent covariates in each model. Gray/Black: statistically in/significant ($p < 0.01$) covariate. Orange to green: most desired to least desired diagnosis outcome. The best model is #6, which adjusts longevity based on sex, year of birth, and geolocation. This model had the lowest MSE, smallest BIC, and best R^2.

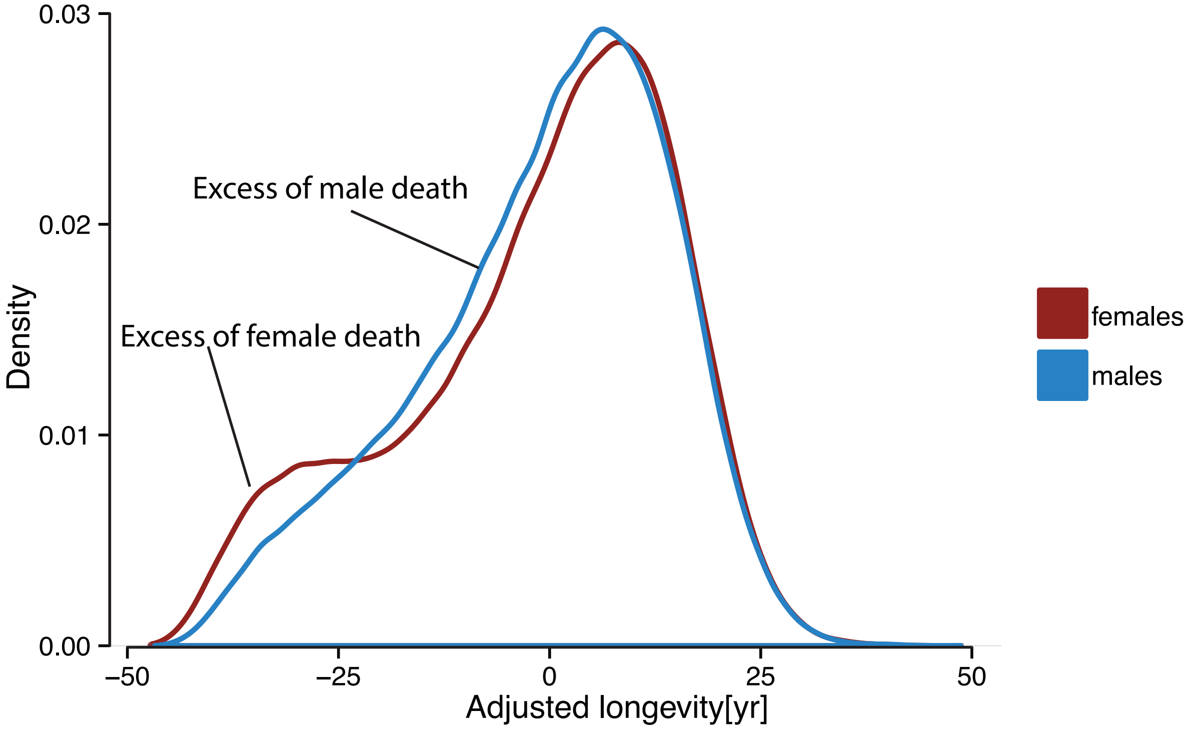


Figure S9: The distribution of the adjusted longevity in all children included in the mid-parent design. The distribution shows excess of very early female death around child bearing ages whereas males show higher rates of death after those years.

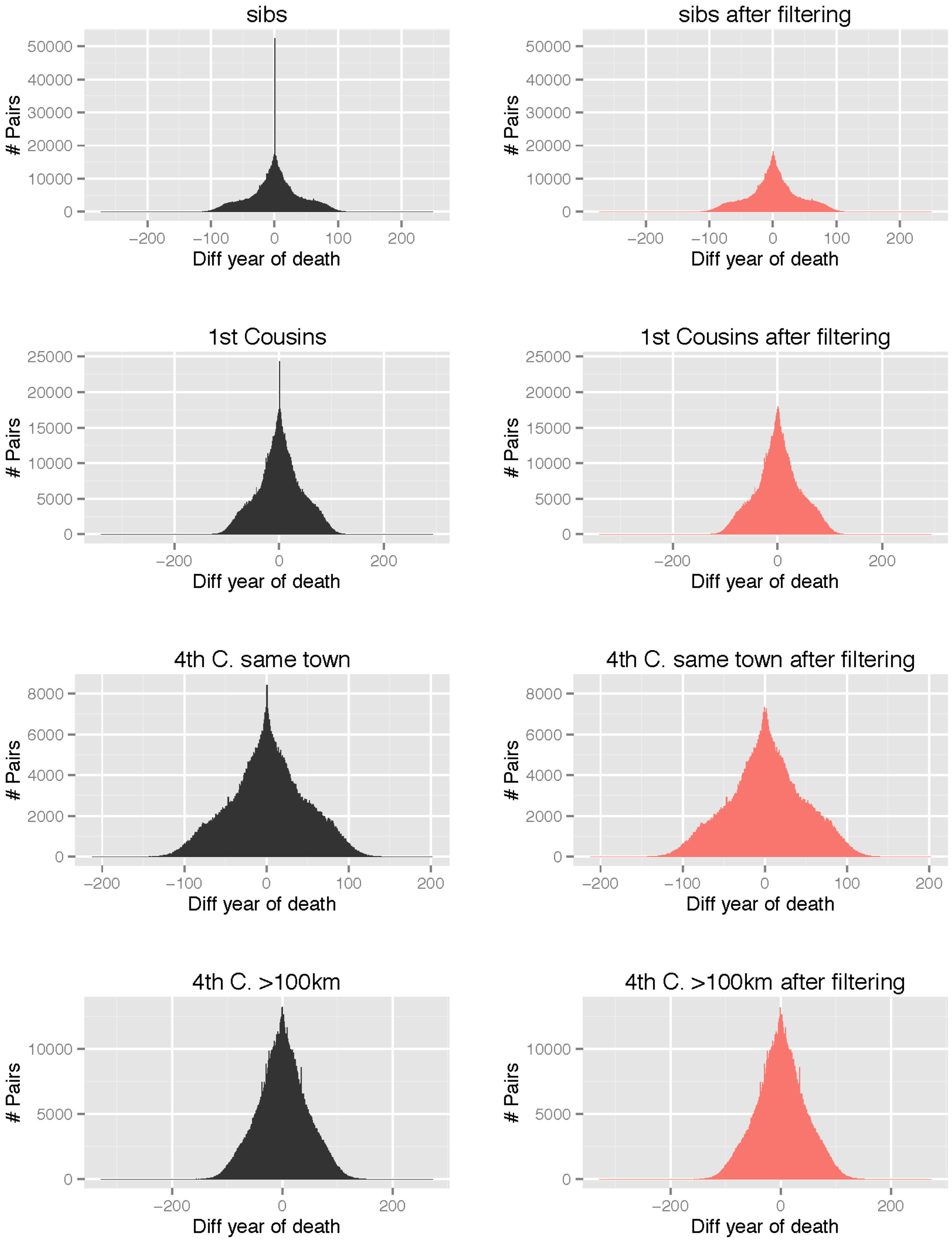


Figure S10: The effect of potential environmental hazards. The histograms present the difference in year of death between various types of pairs of relatives (from top to bottom: siblings, first cousins, 4th cousins that were born in the same town, 4th cousins that were born more than 100km from each other). Left (black): histograms before filtering. Higher death rates within the same year (arrow) are evident in all cases of relatives that were born in the same town. Right (light red): same data after filtering pairs of relatives that died within 10 days. The over-representation is removed, suggesting that this effect is attributed to abrupt environmental hazards such as natural disasters.

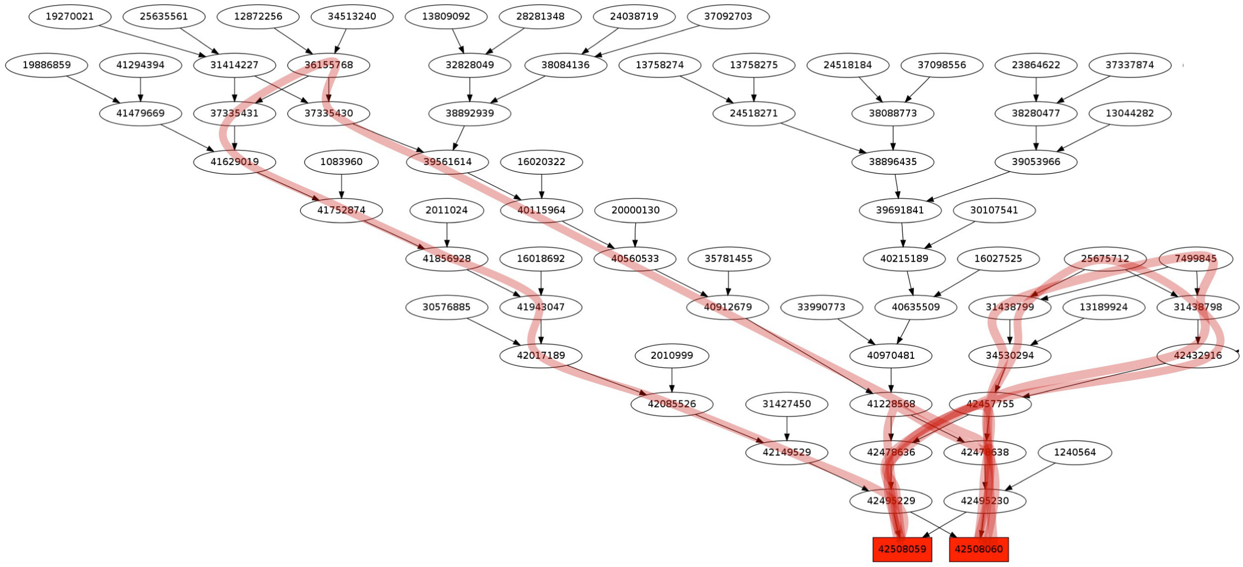


Figure S11: An example of two siblings that are the product of multiple relative marriages. Red: additional paths due to consanguinity. The expected IBD of this pair is 0.565 instead of 0.5. For de-identification, the Node IDs represent the internal representation in our analysis script and not the actual profile ID in Geni.

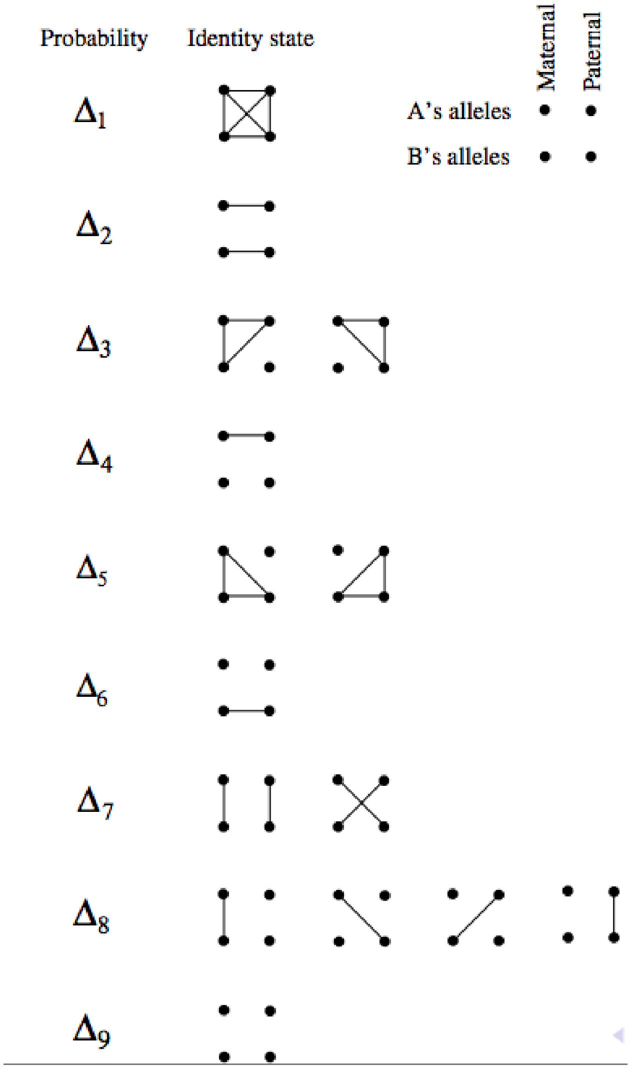


Figure S12: Jacquard's nine condensed coefficients of identity.

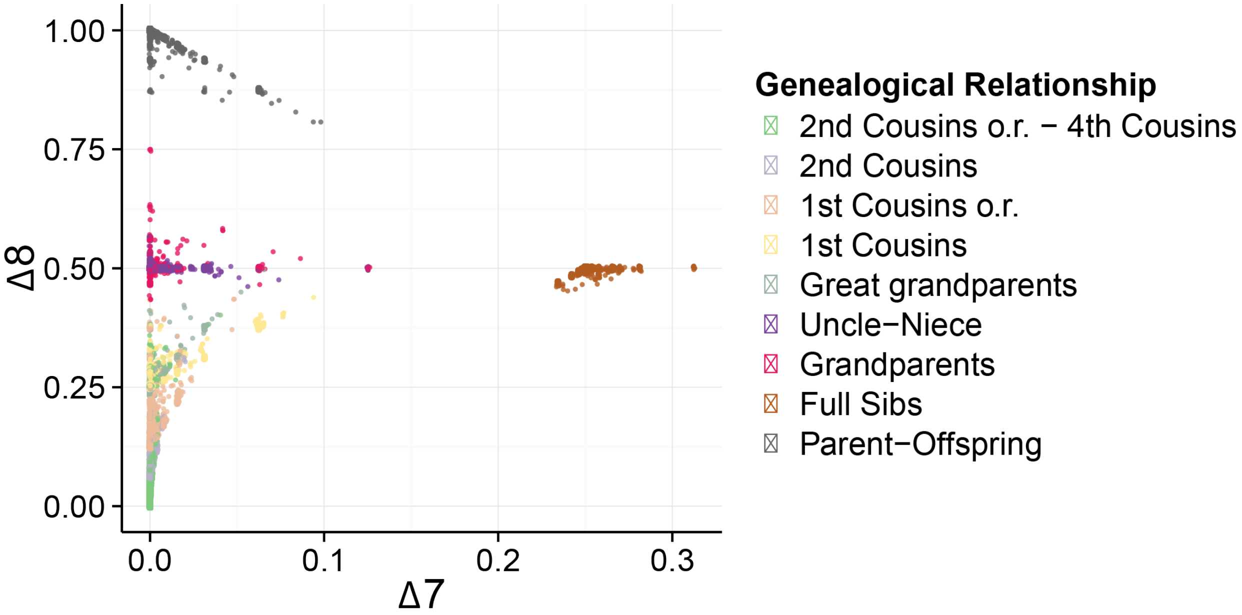
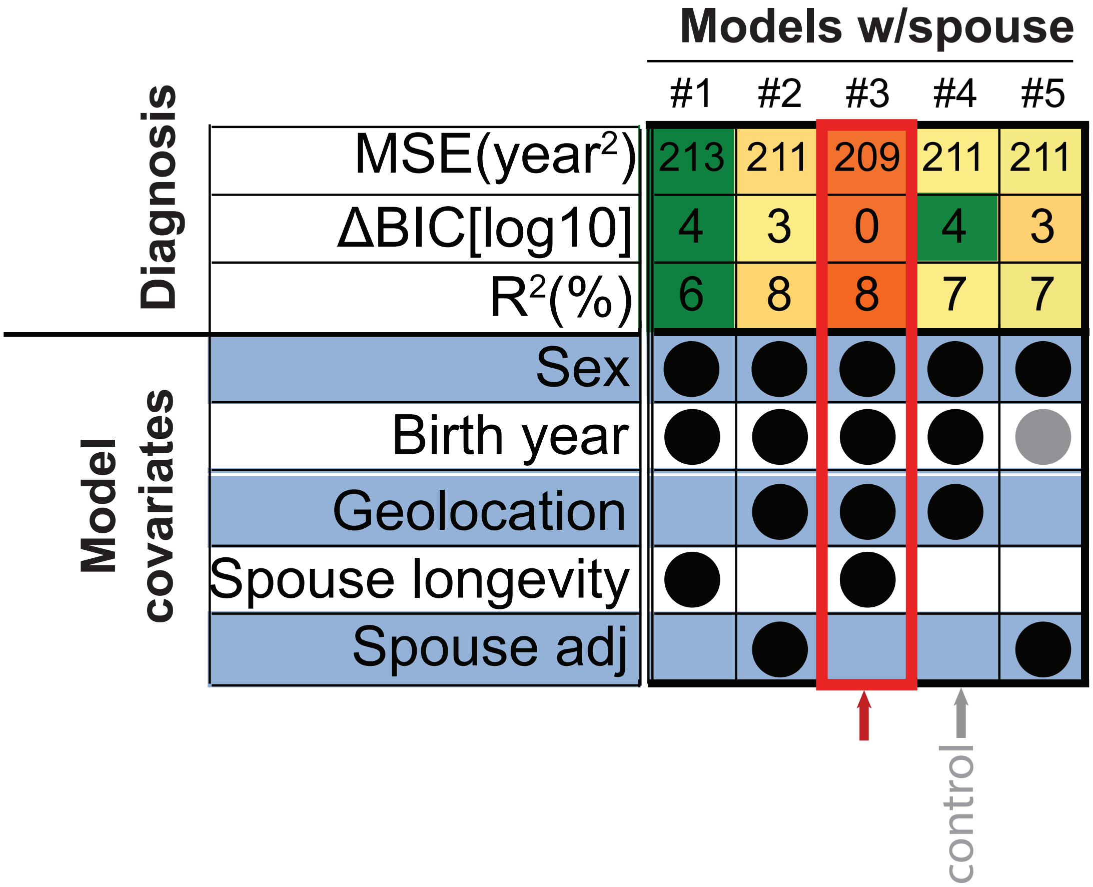
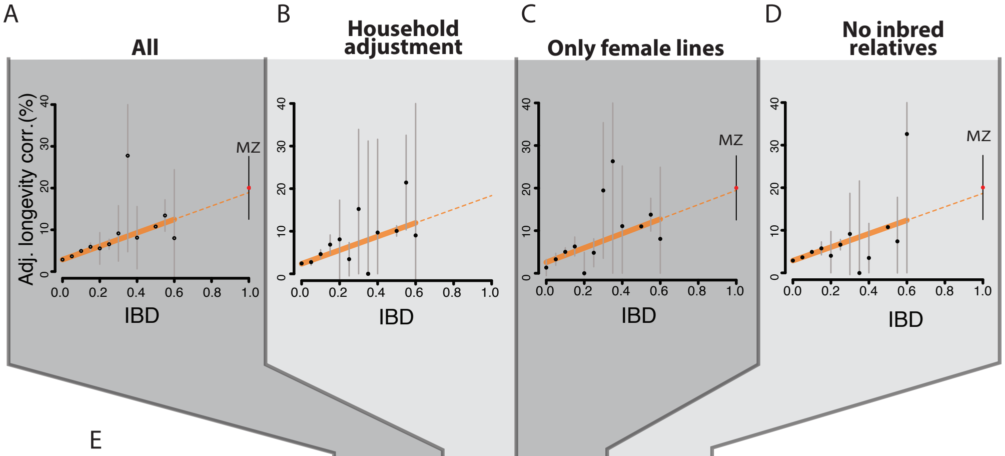


Figure S13: The distribution of Δ_7 vs. Δ_8 in our data. Notice that only full sibs exhibit high frequency of the Δ_7 configuration.



		Models w/spouse				
		#1	#2	#3	#4	#5
Diagnosis	MSE(year^2)	213	211	209	211	211
	ΔBIC[log10]	4	3	0	4	3
	R^2(%)	6	8	8	7	7
Model covariates	Sex	●	●	●	●	●
	Birth year	●	●	●	●	●
	Geolocation		●	●	●	
	Spouse longevity	●		●		
	Spouse adj		●			●

Figure S14: Adjustment of longevity using information from spouses of individuals. MSE is the mean squared error per individual; ΔBIC is the difference of the Bayesian Information Criterion from the best model after $\log_{10}(x+1)$ transformation. Sex is a two level factor (male/female). Geolocation is the longitude/latitude location of birth and was modeled using splines on spheres. Spouse longevity is the age of death of the spouse. Spouse adj. is the longevity of the spouse based after adjustment using model #4 of the fig. S8. Closed circles represent covariates in each model. Gray/Black: statistically in/significant ($p < 0.01$) covariate. Orange to green: most desired to least desired diagnosis outcome. The control model (#4, no spouse information) as a reference point. The best model is #3, which adjusts longevity based on sex, year of birth, geolocation, and spouse longevity. This model had the lowest MSE, smallest BIC, and best R^2.



E

Data	#Relatives	3,279,463	450,533	336,500	2,642,025
Var. decomposition	h^2(%)	16.1	16.0	16.9	15.8
	V_{AA} (%)	0	0	0	0
	V_{AAA}(%)	0	0	0	0
	Dominancy(%)	4.0	2.3	4.0	4.0
Model diagnostic	p(h^2)	10^{-318}	10^{-35}	10^{-66}	10^{-287}
	p(V_{AA})	1	1	1	1
	p(V_{AAA})	1	1	1	1
	ΔMSE $V_{A\times A}$	0	0.32	0.05	0.01
	ΔMSE $V_{A\times A\times A}$	0	0.39	0.06	0.01
	ΔBIC $V_{A\times A}$	15.0	13.0	12.7	14.8
	ΔBIC $V_{A\times A\times A}$	30.0	26.0	25.4	29.57
	ΔMZ (%)	1.1	-	0.5	1.3

Figure S15: The results of variance partitioning of longevity with various adjustment of potential confounders. The figures show the longevity correlation (after dominancy adjustment) as a function of IBD. In all cases, the epistatic terms converged to zero. The table displays the number of pairs of relatives in each condition and the maximum likelihood estimators for each component of variance. The dominancy component was evaluated only for "All" and "Household" conditions. We used the same value measured in "All" for "Female Lines" and "No Inbred Relatives". p-values denote the results of a nested ANOVA for the additive component, squared-, and cubic- epistasis. ΔMSE denotes the average residual error per sample between each epistatic model and the additive model. ΔBIC denotes BIC difference between each epistatic model and the additive model. ΔMZ shows difference between the observed correlation of longevity in the Danish MZ twins and the extrapolation of the additive model.

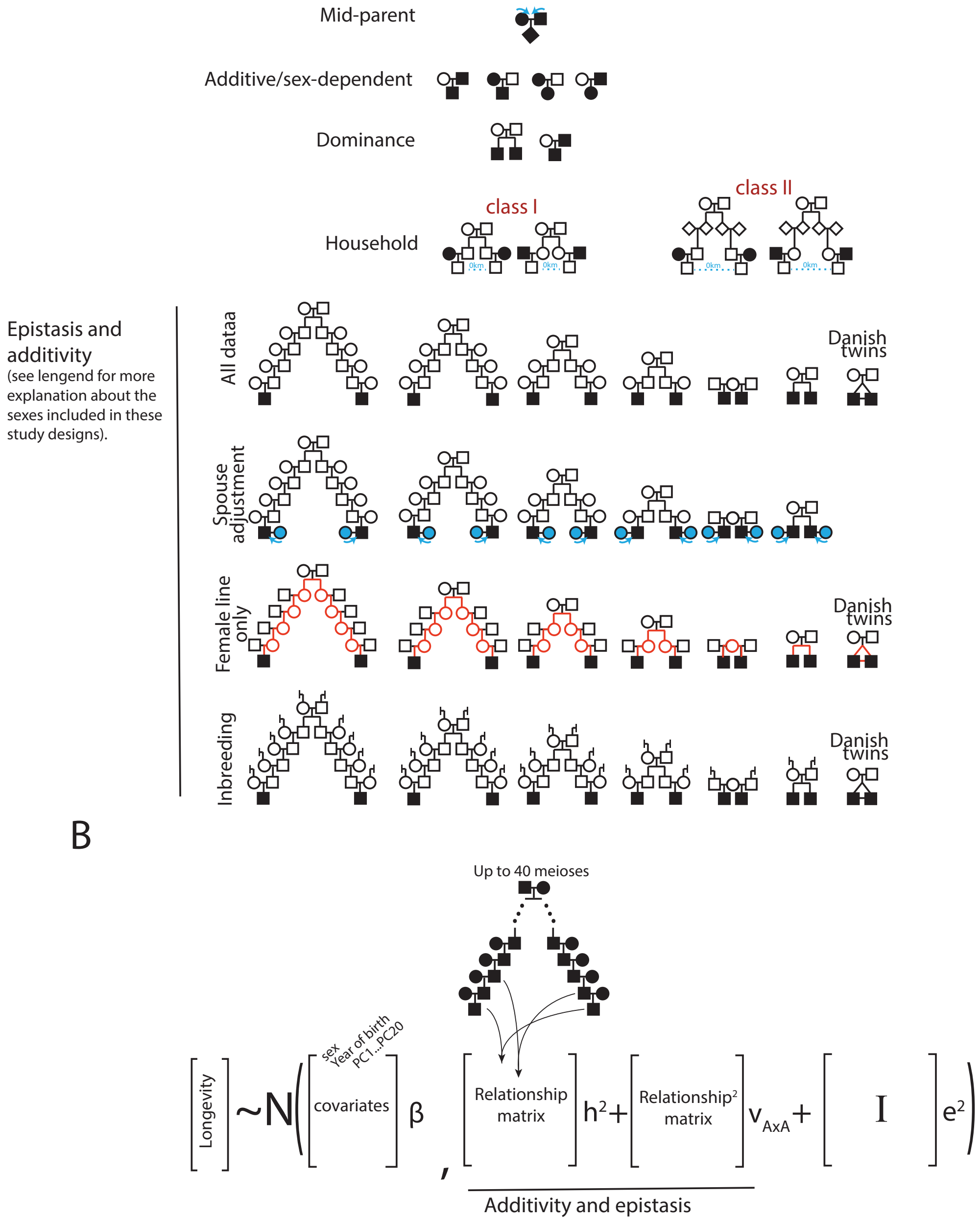


Figure S16: A graphical representation of all study designs used to infer the variance components of longevity. Pairs included in the analysis are filled in black (A) Study designs using least square regression and correlation differences. For clarity, we simplified the pedigrees in the epistasis and addivity section. In reality, this section involved both pairs of sex-concordant relatives, females and males. Also, for ”All data”, ”Spouse adjustments”, and ”Inbreeding” conditions, both maternal, paternal, and mixed lines of familial relationships were considered. (B) The Sci-LMM study design. All possible pairs of individuals were included in the analysis without any restriction on familial lines. A dominance relationship matrix was not included because it was almost indistinguishable from the identity matrix, owing to the relatively small fraction of parent-child pairs in the matrix.

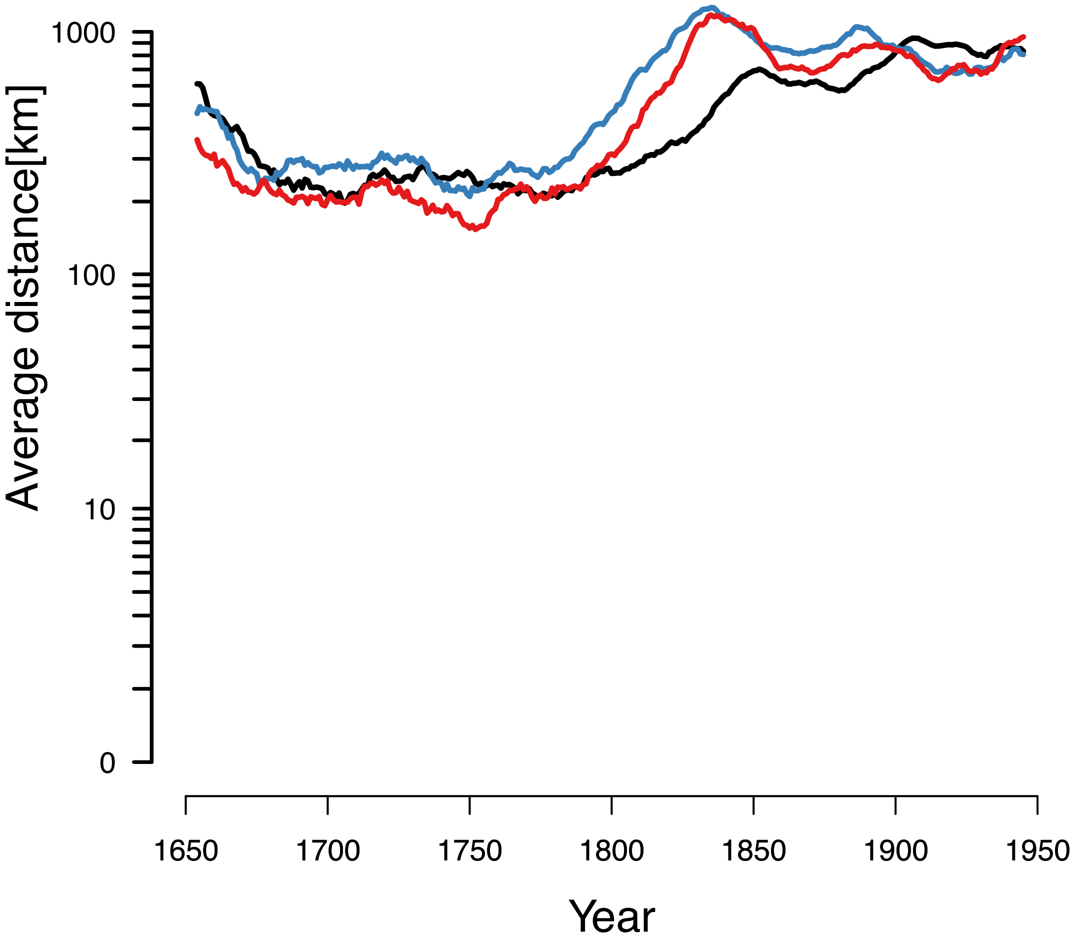


Figure S17: Average migration distance of individuals as a function of time. Red: mother-offspring, blue: father-offspring, black: marital radius. Dots represent the data before smoothing.

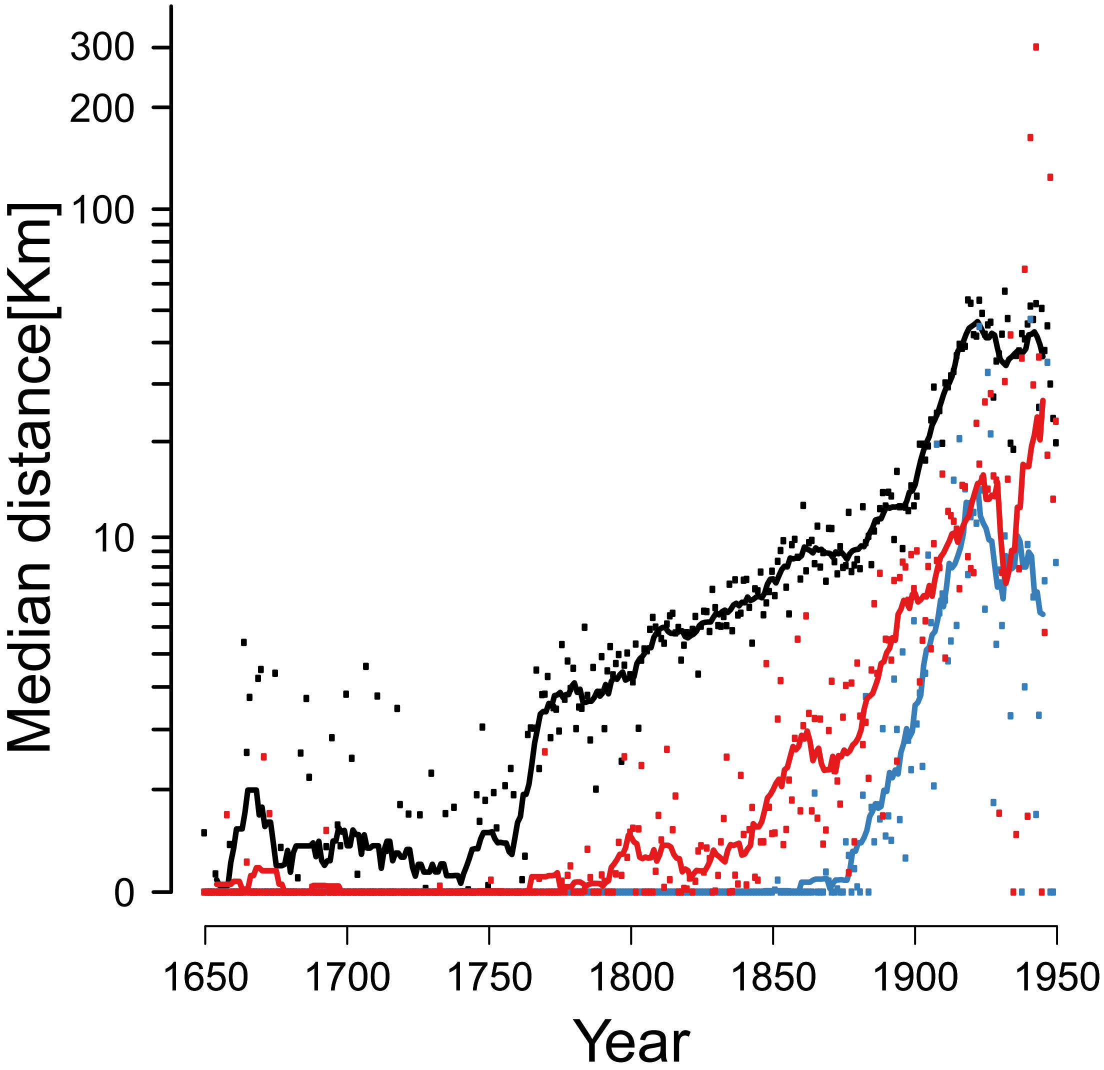


Figure S18: Median migration distance in European-only born individuals as a function of time. Red: mother-offspring, blue: father-offspring, black: marital radius. Dots represent the data before smoothing.

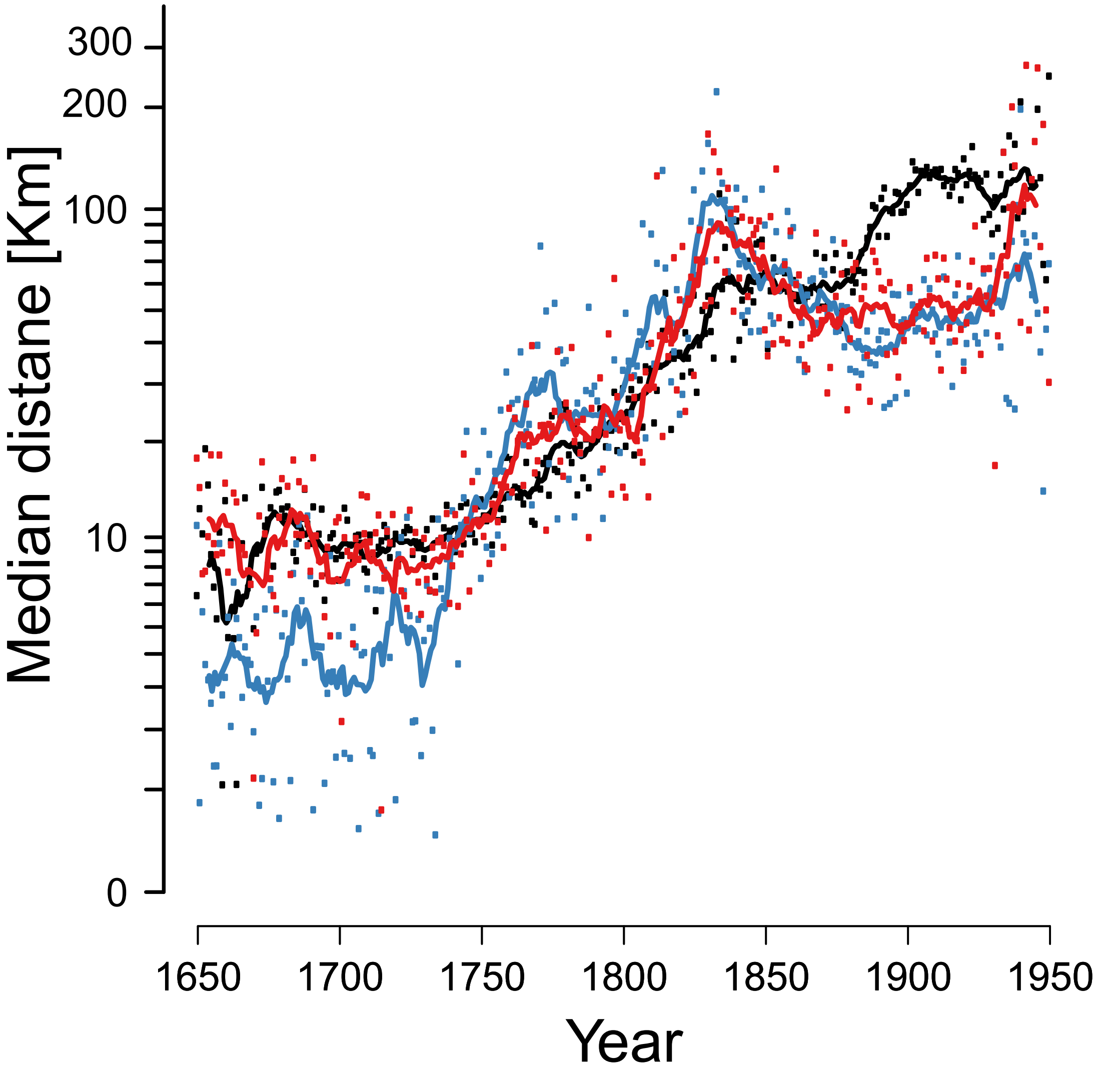


Figure S19: Median migration distance in North American-only born individuals as a function of time. Red: mother-offspring, blue: father-offspring, black: marital radius. Dots represent the data before smoothing.

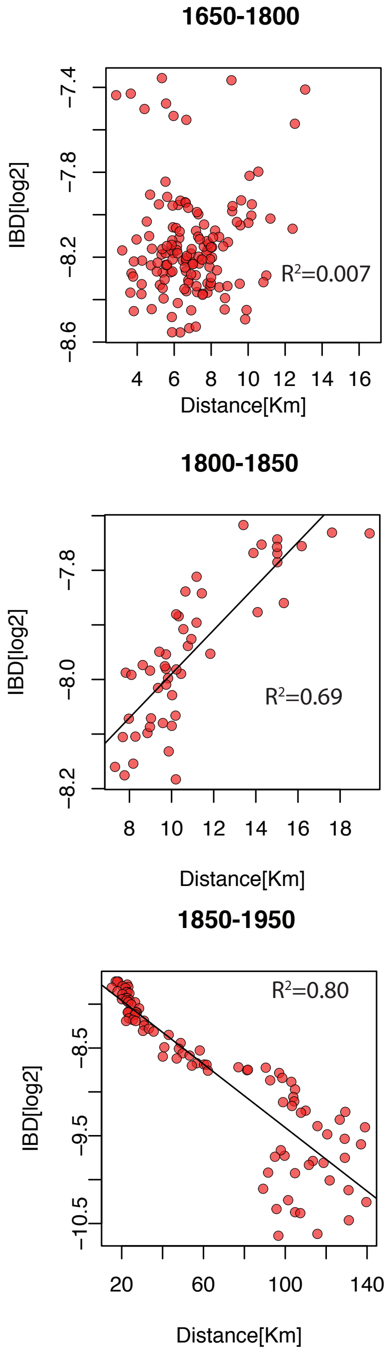
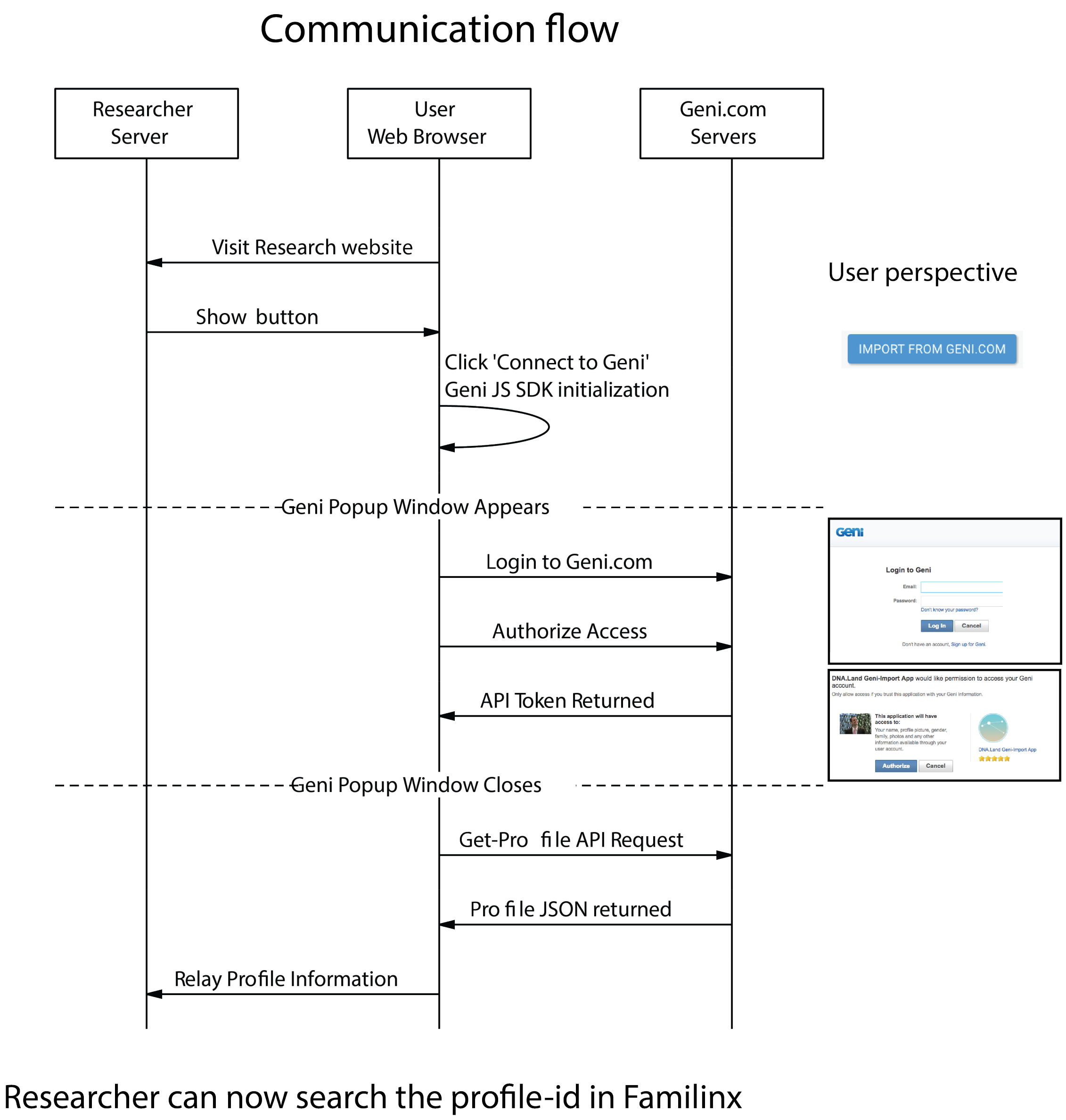


Figure S20: The expected kinship between couples as a function of the median martial distance stratified by time (A) Couples that were born between 1650-1800 (B) Couples that were born between 1800-1850 (C) Couples that were born between 1850-1950.



Login to Geni

Email:

Password:

Don't know your password?

Don't have an account, Sign up for Geni.

DNA.Land Geni-Import App would like permission to access your Geni account.

Only allow access if you trust this application with your Geni information.

This application will have access to:

Your name, profile picture, gender, family, photos and any other information available through your user account.

Figure S21: Communication flow to obtain the profile-id from users. Please see a live demo on `https://teamerlich.org/geni-integration-example/`https://teamerlich.org/geni-integration-example/

Supplementary Tables

Category	All Vermont	Geni in Vermont
Elementary / secondary	32.0-32.7	29.4-36.3
High School	37.0-37.8	30.8-37.8
1yr College	2.9-3.2	2.9-6.0
2yr College	7.4-7.8	6.3-10.4
3yr College	1.7-1.9	0.2-1.7
4yr College	8.2-8.6	6.2-10.2
5+yr College	4.7-5.1	3.4-6.6
No education	1.4-1.6	0.0-0.6

Table S1: 95% **C.I. for education levels in entire Vermont death collection vs. Geni profiles who deceased in Vermont**

Category	All Vermont	Geni in Vermont
VT	53.1-59.7	53.3-54.0
NY	8.1-12.1	12.4-12.9
YY†	5.3-8.6	8.6-9.0
MA	4.8-8.0	6.8-7.2
NH	5.1-8.4	4.3-4.6
CT	1.7-3.9	2.6-2.9
NJ	0.6-2.1	2.0-2.2
PA	0.7-2.3	1.6-1.8
ME	0.6-2.1	1.3-1.5
RI	0.1-1.2	0.6-0.7
IL	0.5-1.9	0.5-0.6
OH	0.2-1.3	0.5-0.6
MI	0.0-0.7	0.4-0.5
CA	0.0-0.3	0.2-0.3

Table S2: 95% **C.I. for state of birth rates between the entire Vermont death collection and Geni users who deceased in Vermont. Only the states with more than 0.2% profiles in the entire Vermont death collection are presented**

† birth outside of the US.

Category	All Vermont	Geni in Vermont
Diseases of the circulatory system	37.3-44.3	41.8-42.5
Neoplasm	23.9-30.3	23.5-24.1
Diseaes of the respiratory system	7.6-11.8	9.9-10.4
Injury and poisoning	2.1-4.8	6.5-6.9
Diseases of the digestive system	2.3-4.9	3.5-3.8
Endocrine, metabolicand immunity disorders	3.5-6.7	3.3-3.6
Diseases of the nervous system	1.9-4.5	2.5-2.8
Mental disorders	1.0-3.1	2.0-2.2
Diseases of the genitourinary system	1.0-3.1	1.6-1.8
Infectious and parasitic diseases	0.4-1.9	1.3-1.5
Certain conditions originating in the perinatal period	0.0-0.8	0.9-1.0
Symptoms, signs, and ill-defined conditions	0.0-0.4	0.4-0.6
Diseases of the blood and blood-forming organs	0.1-1.2	0.4-0.5
Congenital anomalies	0.1-1.2	0.3-0.4

Table S3: 95% C.I. for ICD-9 cause of death categories between the entire Vermont death collection and Geni users who deceased in Vermont. Categories with less than 0.1% are not presented.

Degree of relationship	Relationship	Expected IBD	Adjusted IBD	#Pairs	Distance	ΔGen
1	P	0.5	0.50112	637,653	9	33
1	S	0.5	0.50114	879,694	0	7
2	Av	0.25	0.23849	1,500,143	9	31
2	G	0.25	0.25216	566,747	35	65
2	HS	0.25	0.25438	94,062	0	14
3	1C	0.125	0.12281	1,269,077	9	11
3	2G	0.125	0.12798	479,309	74	96
4	1RC	0.0625	0.063376	2,573,284	22	31
4	3G	0.0625	0.065567	400,327	123	128
5	2C	0.03125	0.032895	2,079,209	30	14
5	4G	0.03125	0.033906	322,311	225	158
6	2RC	0.015625	0.017622	4,424,847	46	31
6	5G	0.015625	0.017573	235,245	435	188
7	3C	0.0078125	0.0091457	3,182,460	55	16
8	3RC	0.0039062	0.005657	6,759,744	72	32
9	4C	0.0019531	0.0028056	4,168,540	75	19

Table S4: The number and basic properties of the pairs of relatives in FamiLinx.

These are the pairs with exact date of birth, exact date of death, and exact birth location. More pairs are available but with incomplete data. Relationship: C. denotes cousinship and ”R” denotes once removed. Av, G, HS, P, and S denotes avuncular, grandparent-grandchild, half-sibs, parent-offspring, and full sib relationships. 2G and 3G denote great and great-great granparent relationships. Adjusted IBD: the averaged IBD after taking into account consanguineous relationships. Distance: the median birth distance in Km. ΔGen: the average difference between the birth years of the relatives.

Paper†	**Population**	**Relative Types**	**Sample Size**	**Years of Birth**	**h^2 estimate**
Herskind et al. *(33)* 1996	Danish	Same sex twin pairs	5744	1870-1900	Males: 26% Females: 23%
Kerber et al. *(35)* 2001	Mormon	Various types up to 1st cousins	78994	1870-1907 lived to at least 65	15% (12-18%)
Ljungquist et al. *(32)* 1998	Swedish	Same sex twin pairs	1200	1886-1925	All twin pairs: 35% Male Twin pairs reared apart: 0% Female Twin pairs reared apart: 15%
Mayer 1991 *(31)*	New England	6 large families	13656	1650-1874	Parent-offspring: 10-33% Sibship-Parent: 16-22% Siblings: 33-41%
Mitchell et al. *(34)* 2001	Amish	Parent-offspring, siblings	1655	Before 1890	25% (20-30%)
Phillipe 1978 *(30)*	French Canadian	Parent-offspring	265	Parents married 1820-1899	Parent-offspring: 10.1% Like-sex sibling: 10.1% Unlike-sex siblings: 13.9% (Spouses: 12.1%)

Table S5: Heritability estimates of longevity from previous studies

† For full citation, see main text.

Relationship	Pairs of analysis	Adj. longevity correlation
S	253076	0.09231052
HS	24243	0.04675896
1C	357139	0.02361952
2C	584856	0.01122288
3C	891462	0.00684701
4C	1168687	0

Table S6: The number of pairs and correlation of longevity for each class of relatives. The correlation was adjusted to remove dominant effects and used four cousins as the baseline.

For the movie, see the Supplemental Information on the *Science* magazine website.

Movie S1: **A time lapse of the birth places of the Geni data from 1400 to 1900 in jumps of five years**. Each colored pixel corresponds to a genealogical profile and the intensity indicates the number of profiles. Prominent colonization events are noted.