File S1

Extended analytical results

S1 The variance of the total sharing

S1.1 The expected number of recent mutations on a shared segment

Consider a segment shared IBD between two individuals. Regardless of the segment length, the two individuals are expected to differ in ≈ 1 site along the segment. This is because for a pair of individuals with MRCA g generations ago, the shared segment is of typical length $100/(2g)$ cM (see, e.g., *Mean total sharing* section in the main text). The number of recent mutations per cM is $2g\mu$, where μ is the mutation rate per generation per cM. The total number of differences is therefore approximately

$$\#\text{ differences} \approx \frac{100}{2g} 2g\mu = 100\mu. \tag{1}$$

For the human genome, $\mu \approx 10^{-8}$ per generation per bp [1], or $\approx 0.8 \cdot 10^{-2}$ per generation per cM (1MB corresponds roughly to 1.25cM). The number of difference is therefore around 1.

S1.2 The assumptions underlying derivation of the variance of the total sharing

We summarize below the assumptions made when calculating the mean and the variance of the total sharing (main text *Mean total sharing* and *The variance of the total sharing* sections).

1. The population is Wright-Fisher with constant (effective) size N. We do not distinguish between male and female history, and all present-day individuals are represented as random pairs of haploids from the current generation.
2. The ancestral process is described by Kingsman's coalescent [2]; specifically, time is assumed to be continuous, and the distribution of coalescence times is exponential with rate 1.
3. Recombination is a Poisson process with rate 0.01 per cM.
4. The recombination rate between markers is proportional to the genetic distance between the them.
5. The markers are equally spaced, in genetic distance, along each chromosome and are dense enough, that when calculating the probability that a segment has length $\geq m$, we can ignore the discreteness of the markers.
6. If two sites are on different chromosomes, they are shared or not independently of each other.
7. Boundary effects at the ends of the chromosomes are ignored.
8. We assume that the events that two sites are in shared segments are independent once we specify the time to the MRCA at each site.

Assumptions 1, 2, 3, and 4 are standard when studying finite, isolated populations [2]. Assumption 5 should present no problem in practice, with SNP arrays covering over a million sites or with whole-genome sequences. For assumption 6, we can, approximately, expect segments on different chromosomes to be shared independently of each other if the individuals are sufficiently unrelated that the average number of segments shared genome-wide is less than one, which is true for 4th (half-) cousins or less related individuals [3]. Assumption 7 is reasonable when $L \gg m$ (L is the length of the chromosome, m is the minimal segment length).

For the last assumption (8), one may suggest that if there was no recombination event in the history of two sites, then they are not independent. The reason why our approximation works is that when the two sites have the same coalescence time, it is usually very short (otherwise there would have been a recombination event and the coalescence times would not be the same in the two sites), increasing the probability that they lie on shared segments. If the sites have different coalescence times, the times would tend to be longer, reducing the probability that the sites are on shared segments, in accordance with the fact that they were separated by a recombination event.

One importance of the derivation presented in the main text is that it sets the framework for a more detailed calculation that eliminates the last assumption. It does so by conditioning the probability $\pi_2(s_1, s_2)$ on whether or not there was a recombination event. For each case, it then proceeds using the Markov chain representation of coalescent with recombination. This is explained in the next subsection.

S1.3 An alternative calculation of the variance of the total sharing

In this subsection, we recalculate the probability $\pi_2(s_1, s_2) = \pi_2(k)$ of two sites separated by k markers to be both on shared segments of length $\geq m$. We use the Markov chain illustrated in Figure 1 of the main text as well as other notation as used in the main text. As mentioned above, we calculate π_2 by conditioning on whether or not the two sites have been separated by a recombination event,

$$\pi_2 = p_{\rm nr}\pi_{\rm nr} + (1 - p_{\rm nr})\pi_{\rm r}, \tag{2}$$

where $p_{\rm nr}$ is the probability of no recombination, $\pi_{\rm nr}$ is the probability of both sites to be in shared segments when there was no recombination, and $\pi_{\rm r}$ is the probability of both sites to be in shared segments when there was recombination.

To calculate the probability of no recombination, we consider the discrete time Wright-Fisher model (as we found that it matches better the discrete-time simulations). In discrete time, the PDF of g, the number of generations to the (single-site) MRCA, is geometric, $P(g) = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{g-1}$. Given an MRCA at generation g at one site, we require that there was no recombination between that site and the other site, in both chromosomes, and in all g generations. Because recombination is a Poisson process and the distance between the sites is $d = k\frac{L}{M}$, there will be no recombination with probability

$$p_{\rm nr} = \sum_{g=1}^{\infty} \frac{1}{N}\left(1 - \frac{1}{N}\right)^{g-1} e^{-dg/50} = \frac{1}{1 + N\left(e^{d/50} - 1\right)}. \tag{3}$$

The scaled recombination rate ρ was defined as in the main text as $\rho = 2Nd/100$ [4].

Consider now the no-recombination probability, $\pi_{\rm nr}$. As long as $d \geq m$, $\pi_{\rm nr}$ is trivially 1. If $d < m$, the segment spanning the two sites is of length $d + \ell_{1L} + \ell_{2R}$, where ℓ_{1L} is the distance to the next recombination event to the *left of the left marker*, and similarly for ℓ_{2R} (see Figure 1 for illustration). Given that the coalescence time (at both sites) was t, both ℓ_{1L} and ℓ_{2R} are exponentially distributed with rate $2Nt/100$. The PDF of the coalescence time is $\Phi(t) = (1 + \rho)e^{-(1+\rho)t}$, since this is the PDF of the time to exit state 1, and we are given that there was no recombination before coalescence. Therefore,

$$\pi_{{\rm nr};d<m} = \int_0^{\infty} (1 + \rho)e^{-(1+\rho)t}dt \int_{m-d}^{\infty} \left(\frac{Nt}{50}\right)^2 \ell e^{-Nt\ell/50} d\ell. \tag{4}$$

These integrals are easily solvable, giving

$$\pi_{\rm nr} = \begin{cases} 1 - \left[\frac{N(m-d)}{N(m-d)+50(1+\rho)}\right]^2 & d < m, \\ 1 & d \geq m. \end{cases} \tag{5}$$

It is easy to see that $\lim_{d\to m^+} \pi_{\rm nr} = \lim_{d\to m^-} \pi_{\rm nr}$, as expected.

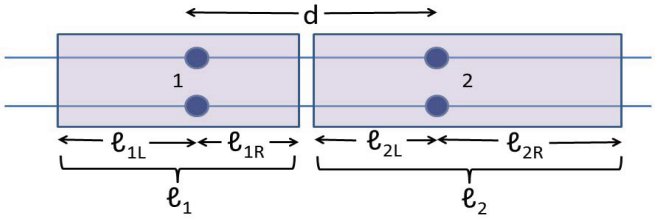

Figure 1: An illustration of the shared segments spanning two sites (numbered 1 and 2). Lines correspond to chromosomes and circles to sites, which are distance d apart. The shaded boxes correspond to hypothetical shared segments. The left segment extends to distance ℓ_{1R} to the right of the site and ℓ_{1L} to the left of it, and similarly for the right segment.

The case of recombination is more complicated. One might think that if there was a recombination event in the history of the two sites, then the two sites will be shared or not independently. However, the presence of a recombination event implies that the sum of ℓ_{1R} and ℓ_{2L} [(the segment length to the right of the left marker) and (the segment length to the left of the right marker)] cannot exceed d (see Figure 1 for illustration). We simplify the analysis by assuming instead that each of those two segments cannot exceed length d, but that their lengths are otherwise independent, resulting in a slight overestimation of π_r. Thus, for a given time to MRCA, t_1, the segment length spanning the left site can be written as $\ell_1 = \ell_{1L} + \ell_{1R}$ (see Figure 1), where ℓ_{1L} is distributed exponentially with rate $Nt_1/50$,

$$P(\ell_{1L}) = \frac{Nt_1}{50} e^{-\frac{Nt_1 \ell_{1L}}{50}} \quad ; \quad \ell_{1L} > 0, \tag{6}$$

and ℓ_{1R} is similarly distributed, except for an upper cutoff at $\ell_{1R} = d$,

$$P(\ell_{1R}) = \frac{\frac{Nt_1}{50} e^{-\frac{Nt_1 \ell_{1R}}{50}}}{1 - e^{-\frac{Nt_1 d}{50}}} \quad ; \quad 0 < \ell_{1R} < d. \tag{7}$$

Using convolution, the probability density function of $\ell_1 = \ell_{1L} + \ell_{1R}$ is

$$P(\ell_1) = \frac{\left(\frac{Nt_1}{50}\right)^2 e^{-\frac{Nt_1 \ell_1}{50}}}{1 - e^{-\frac{Nt_1 d}{50}}} \cdot \begin{cases} d & \ell_1 < d, \\ \ell_1 & \ell_1 \geq d. \end{cases} \tag{8}$$

The probability that $\ell_1 \geq m$ and thus the site is on a shared segment is

$$P(\ell_1 > m) = \frac{1}{1 - e^{-\frac{Nt_1 d}{50}}} \cdot \begin{cases} d\frac{Nt_1}{50} e^{-\frac{Nt_1 m}{50}} & d < m, \\ \left(1 + m\frac{Nt_1}{50}\right) e^{-\frac{Nt_1 m}{50}} - e^{-\frac{Nt_1 d}{50}} & d \geq m. \end{cases} \tag{9}$$

For large d, $P(\ell_1 > m) \to \left(1 + m\frac{Nt_1}{50}\right) e^{-\frac{Nt_1 m}{50}}$, which is exactly the single-site expression (Eq. (1) in the main text), as expected. We then simplify again by approximating the denominator of $P(\ell_1 > m)$ with 1,

$$P(\ell_1 > m) \approx \begin{cases} d\frac{Nt_1}{50} e^{-\frac{Nt_1 m}{50}} & d < m, \\ \left(1 + m\frac{Nt_1}{50}\right) e^{-\frac{Nt_1 m}{50}} - e^{-\frac{Nt_1 d}{50}} & d \geq m. \end{cases} \tag{10}$$

This should lead to a slight underestimation of π_r. From here on the calculation is exact. An equation identical to (10) holds for $P(\ell_2 > m)$. Integrating the probabilities of the two sites to be in shared segments over all possible coalescence times, we have, for $d < m$,

$$\pi_r = \int_0^\infty \int_0^\infty \Phi(t_1, t_2) d\frac{Nt_1}{50} e^{-\frac{Nt_1 m}{50}} d\frac{Nt_2}{50} e^{-\frac{Nt_2 m}{50}} dt_1 dt_2. \tag{11}$$

For $d \geq m$,

$$\pi_{\rm r} = \int_0^\infty \int_0^\infty \Phi(t_1,t_2) \left[\left(1 + m\frac{Nt_1}{50}\right) e^{-\frac{Nt_1 m}{50}} - e^{-\frac{Nt_1 d}{50}} \right] \left[\left(1 + m\frac{Nt_2}{50}\right) e^{-\frac{Nt_2 m}{50}} - e^{-\frac{Nt_2 d}{50}} \right] dt_1 dt_2. \quad (12)$$

As in the main text, this can be rewritten naturally in terms of the Laplace transform of Φ,

$$\widehat{\Phi}(q_1,q_2) = \int_0^\infty \int_0^\infty e^{-q_1 t_1 - q_2 t_2} \Phi(t_1,t_2) dt_1 dt_2. \quad (13)$$

After some algebra, we find, for $d < m$,

$$\pi_{\rm r} = d^2 \left[\frac{\partial}{\partial m_1} \frac{\partial}{\partial m_2} \widehat{\Phi}\left(\frac{m_1 N}{50}, \frac{m_2 N}{50}\right) \right]_{\substack{m_1=m \\ m_2=m}}. \quad (14)$$

For $d \geq m$,

$$\begin{aligned} \pi_{\rm r} = {} & \widehat{\Phi}\left(\frac{mN}{50}, \frac{mN}{50}\right) - 2m \left[\frac{\partial}{\partial m_1} \widehat{\Phi}\left(\frac{m_1 N}{50}, \frac{mN}{50}\right) \right]_{m_1=m} + m^2 \left[\frac{\partial}{\partial m_1} \frac{\partial}{\partial m_2} \widehat{\Phi}\left(\frac{m_1 N}{50}, \frac{m_2 N}{50}\right) \right]_{\substack{m_1=m \\ m_2=m}} \\ & + \widehat{\Phi}\left(\frac{dN}{50}, \frac{dN}{50}\right) - 2\widehat{\Phi}\left(\frac{mN}{50}, \frac{dN}{50}\right) + 2m \left[\frac{\partial}{\partial m_1} \widehat{\Phi}\left(\frac{m_1 N}{50}, \frac{dN}{50}\right) \right]_{m_1=m}. \end{aligned} \quad (15)$$

We are therefore left only with finding $\widehat{\Phi}(q_1,q_2)$. This can be carried out almost as in the main text, except that we must take into account that there was recombination before coalescence, that is, the Markov chain jumped from the initial state 1 to state 2 and not to state 8. Therefore, the coalescence times at the two sites, t_1 and t_2, can be seen as a sum of t', the time it took to jump from state 1 to state 2, and the times it took from state 2 until coalescence events occurred in both sites. As we explained just before Eq. (4), the time it takes to jump from state 1 to state 2, given recombination, is distributed exponentially with rate $(1+\rho)$. Therefore,

$$\Phi(t_1,t_2)dt_1 dt_2 = \begin{cases} \int_0^{t_1} (1+\rho) e^{-(1+\rho)t'} P_{21}(t_1 - t') \delta(t_2 - t_1) dt' dt_1 dt_2 & t_1 = t_2, \\ \int_0^{t_1} (1+\rho) e^{-(1+\rho)t'} \left[P_{22}(t_1 - t') + P_{23}(t_1 - t') \right] e^{-(t_2 - t_1)} dt' dt_1 dt_2 & t_1 < t_2, \\ \int_0^{t_2} (1+\rho) e^{-(1+\rho)t'} \left[P_{22}(t_2 - t') + P_{23}(t_2 - t') \right] e^{-(t_1 - t_2)} dt' dt_1 dt_2 & t_2 < t_1. \end{cases} \quad (16)$$

In the last equation, $P_{2i}(t)$ is the probability of the chain to be at state i at time t, given that it started at state 2. The reasoning behind the last equation is as follows. In the case $t_1 = t_2$, to coalesce at both sites at time t_1, we need to wait time t' to jump to state 2, then be back in state 1 after another period of $(t_1 - t')$ (probability $P_{21}(t_1 - t')$), and then jump to state 8 (probability dt_1). To coalesce at site 1 (the left one) only at time t_1, we need to wait time t' to get to state 2, and then be at state 2 (or 3) at time $(t_1 - t')$ (probability $P_{22}(t_1 - t')$ or $P_{23}(t_1 - t')$) and jump to state 5 (or 7; probability dt_1). Then, coalescence at site 2 (the right one) at time $t_2 > t_1$ occurs with probability $e^{-(t_2 - t_1)} dt_2$. The case $t_1 > t_2$ is similarly explained. Taking the Laplace transform of the last equation,

$$\begin{aligned} \widehat{\Phi}(q_1,q_2) = {} & \int_0^\infty \int_0^\infty e^{-q_1 t_1 - q_2 t_2} \int_0^{t_1} (1+\rho) e^{-(1+\rho)t'} P_{21}(t_1 - t') \delta(t_2 - t_1) dt' dt_1 dt_2 \\ & + \int_0^\infty \int_{t_1}^\infty e^{-q_1 t_1 - q_2 t_2} \int_0^{t_1} (1+\rho) e^{-(1+\rho)t'} \left[P_{22}(t_1 - t') + P_{23}(t_1 - t') \right] e^{-(t_2 - t_1)} dt' dt_2 dt_1 \\ & + \int_0^\infty \int_{t_2}^\infty e^{-q_1 t_1 - q_2 t_2} \int_0^{t_2} (1+\rho) e^{-(1+\rho)t'} \left[P_{22}(t_2 - t') + P_{23}(t_2 - t') \right] e^{-(t_1 - t_2)} dt' dt_1 dt_2. \end{aligned} \quad (17)$$

The first term of the right-hand-side can be solved as follows,

$$\int_0^\infty \int_0^\infty e^{-q_1 t_1 - q_2 t_2} \int_0^{t_1} (1+\rho) e^{-(1+\rho)t'} P_{21}(t_1 - t') dt' dt_1 \delta(t_2 - t_1) dt_2 =$$
$$(1+\rho) \int_0^\infty e^{-(q_1+q_2)t_1} \left[\int_0^{t_1} e^{-(1+\rho)t'} P_{21}(t_1 - t') dt' \right] dt_1 =$$
$$\frac{1+\rho}{1+\rho+q_1+q_2} \widehat{P}_{21}(q_1+q_2). \tag{18}$$

The last line results from the special structure of the integrals in the second line: the internal integral is a convolution between $e^{-(1+\rho)t}$ and $P_{21}(t)$, and the external integral is the Laplace transform $t_1 \to (q_1+q_2)$ of the internal integral. Applying the convolution theorem (recalling that the Laplace transform of e^{-at} is $(a+q)^{-1}$), we arrive at the last line. The second and third terms of Eq. (17) require more algebra but are solved similarly, finally giving

$$\widehat{\Phi}(q_1,q_2) = \frac{1+\rho}{1+\rho+q_1+q_2} \left\{ \left(\frac{1}{1+q_1} + \frac{1}{1+q_2} \right) \left[\widehat{P}_{22}(q_1+q_2) + \widehat{P}_{23}(q_1+q_2) \right] + \widehat{P}_{21}(q_1+q_2) \right\}. \tag{19}$$

By that we are almost done, since as in the main text, the Laplace transform of the transition probabilities $\widehat{P}_{2i}(q)$ can be readily found using the continuous-time Markov chain relation

$$\widehat{P_{2i}}(q) = (qI - Q)^{-1}_{2i}, \tag{20}$$

where Q is the transition rate matrix of the chain. Substituting, using MATHEMATICA, Eq. (20) in Eq. (19) gives

$$\widehat{\Phi}(q_1,q_2) = \frac{(1+\rho)\{2(6+q)[3+q_1(4+q_1)+q_2(4+q_2)+3q_1q_2]+\rho(2+q)(13+3q)+\rho^2(2+q)\}}{(1+q_1)(1+q_2)(1+q+\rho)[2(1+q)(3+q)(6+q)+\rho(2+q)(13+3q)+\rho^2(2+q)]}, \tag{21}$$

where $q = q_1 + q_2$. We then substituted, again using MATHEMATICA, Eq. (21) in Eqs. (14) and (15) to obtain the final expression for $\pi_{\rm r}$. We verified numerically that $\lim_{d\to m^+} \pi_{\rm r} = \lim_{d\to m^-} \pi_{\rm r}$. Eq. (5) for $\pi_{\rm nr}$, Eq. (3) for $p_{\rm nr}$, and Eq. (2) for π_2 complete the derivation.

S1.4 An alternative derivation of $\widehat{\Phi}(q_1,q_2)$ using the Feynman-Kac formula

In this subsection, we show how $\widehat{\Phi}(q_1,q_2)$ (Eq. (11) in the main text and Eq. (21) here) can be derived using the Feynman-Kac formula as described by Fitzsimmons and Pitman [5]. We thank an anonymous reviewer for pointing out this approach.

Let us start with Eq. (11) in the main text. Assume the same continuous-time Markov chain as in the main text, and define a *functional* of the Markov chain as $A_v = \int_0^T v(X_t)dt$, where X_t is the state of the chain at time t, T is the "killing" time when the chain reaches an absorbing state (in our case, state no. 8), and $v(x)$ assigns a value to each state. With this notation, the Laplace transform $\widehat{\Phi}(q_1,q_2)$ (for the case analyzed in the main text, when there is no restriction on the first transition) can be written as

$$\widehat{\Phi}(q_1,q_2) = \int_0^\infty \int_0^\infty e^{-q_1 t_1 - q_2 t_2} \Phi(t_1,t_2) dt_1 dt_2 = \left\langle e^{A_v} \right\rangle, \tag{22}$$

with $v = -(q_1+q_2, q_1+q_2, q_1+q_2, q_1, q_2, q_1, q_2)^T$. This is true, because the left-site coalescence time t_1 is the total time spent by the chain in states 1,2,3,4, and 6, whereas the right-site coalescence time t_2 is the total time spent in 1,2,3,5, and 7.

According to the Feynman-Kac formula [5],

$$\widehat{\Phi}(q_1,q_2) = \left\langle e^{A_v} \right\rangle = \lambda(Q' + M_v)^{-1} Q' \mathbf{1}, \tag{23}$$

where $M_v = \mathrm{diag}(v)$, λ is the initial condition (in our case, $\lambda = (1,0,0,0,0,0,0)$, since the chain always starts at state 1), and $\mathbf{1} = (1,1,1,1,1,1,1)^T$. The matrix Q' is obtained from the transition rate matrix Q by removing the row and column corresponding to the absorbing state (state 8). Carrying out the necessary matrix multiplications and inversions, we obtain the exact same expression as in Eq. (11) in the main text.

In the case analyzed in Section S1.3 above (leading eventually to Eq. (21)), the chain is guaranteed to jump from state 1 to state 2 (but not to state 8) at rate $(1+\rho)$. This can be incorporated into the Feynman-Kac framework by extending the chain to include a "ghost" state 0, from which the only outward transition is to state 2, at rate $(1+\rho)$. No transitions are allowed into state 0, and it is the initial state of the chain. Since neither site has coalesced while in state 0, we can write $\widehat{\Phi}(q_1,q_2) = \left\langle e^{A_v} \right\rangle$ with $v = -(q_1+q_2, q_1+q_2, q_1+q_2, q_1+q_2, q_1, q_2, q_1, q_2)^T$. We then use $\left\langle e^{A_v} \right\rangle = \lambda(Q''+M_v)^{-1}Q''\mathbf{1}$, where $\lambda = (1,0,0,0,0,0,0,0)$ and Q'' is equal to Q', but with an additional row and an additional column for the new state 0:

$$Q'' = \begin{pmatrix} -1-\rho & 0 & 1+\rho & 0 & 0 & 0 & 0 & 0 \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & Q' & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \\ 0 & & & & & & & \end{pmatrix}. \tag{24}$$

Solving and simplifying gives Eq. (21).

S1.5 A linearly expanding population

In this subsection we calculate the mean and the variance of the total sharing for a linearly expanding population. Define the population size as $N(t) = N_0\lambda(t)$, where

$$\lambda(t) = \begin{cases} 1+\tilde{r}(t_0-t) & 0 \le t \le t_0, \\ 1 & t > t_0. \end{cases} \tag{25}$$

This corresponds to a population maintaining a constant size until $t = t_0$ generations ago; starting at $t = t_0$ and until present, the population grows linearly at rate $\tilde{r}$. The PDF of the coalescence times is

$$\Phi(t) = \frac{e^{-\int_0^t \frac{dt'}{\lambda(t')}}}{\lambda(t)}. \tag{26}$$

Substituting $\lambda(t)$ from Eq. (25), we have, for $t \le t_0$,

$$\begin{aligned} \Phi(t) &= \frac{1}{1+\tilde{r}(t_0-t)} \exp\left[-\int_0^t \frac{dt'}{1+\tilde{r}(t_0-t')}\right] \\ &= \frac{1}{1+\tilde{r}(t_0-t)} \exp\left\{\frac{1}{\tilde{r}} \ln\left[\frac{1+\tilde{r}(t_0-t)}{1+\tilde{r}t_0}\right]\right\} \\ &= (1+\tilde{r}t_0)^{-1/\tilde{r}}[1+\tilde{r}(t_0-t)]^{1/\tilde{r}-1}. \end{aligned} \tag{27}$$

For $t > t_0$,

$$\begin{aligned} \Phi(t) &= \exp\left[-\int_0^{t_0} \frac{dt'}{1+\tilde{r}(t_0-t')} - \int_{t_0}^t dt'\right] \\ &= \exp\left[\frac{1}{\tilde{r}} \ln\left(\frac{1}{1+\tilde{r}t_0}\right) - (t-t_0)\right] \\ &= (1+\tilde{r}t_0)^{-1/\tilde{r}} e^{-(t-t_0)}. \end{aligned} \tag{28}$$

In summary,

$$\Phi(t) = (1+\tilde{r}t_0)^{-1/\tilde{r}} \begin{cases} [1+\tilde{r}(t_0-t)]^{1/\tilde{r}-1} & 0 \le t \le t_0 \\ e^{-(t-t_0)} & t > t_0 \end{cases}. \tag{29}$$

We then use Eq. (17) from the main text for the mean total sharing,

$$\langle f_T \rangle = \int_0^\infty \Phi(t) \left(1 + \frac{mN_0 t}{50}\right) e^{-\frac{mN_0 t}{50}} dt, \tag{30}$$

and Eq. (19) from the main text for the variance of the total sharing,

$$\mathrm{Var}\left[f_T\right] \approx 2\int_{m/L}^1 (1-x)\left[\int_0^\infty \Phi(t) e^{-txN_0L/50} dt\right] dx. \tag{31}$$

The integral in $\langle f_T \rangle$ and the internal integral (over t) in $\mathrm{Var}\left[f_T\right]$ can be evaluated in terms of incomplete Gamma functions (not shown). For $\mathrm{Var}\left[f_T\right]$, the external integral must be evaluated numerically. [We also tried to change the order of the integration in Eq. (31), that is, to compute the integral over x first. However, in that case, while the integral over x was solvable, the integral over t was not.] We compare the results of Eqs. (30) and (31) to simulations in Figure **S1**. In the simulations, the ancestral population size was set to $N_a = 10000$, the expansion started $E_t = 500$ generations ago, and the final (current) population size varied in the range $N_c = [10500, 15000]$. In terms of the parameters of $\lambda(t)$, this corresponds to $N_0 = N_a = 10000$, $t_0 = 500/10000 = 0.05$, and $\tilde{r}$ between $(1.05-1)/0.05 = 1$ and $(1.5-1)/0.05 = 10$. The comparison shows reasonable agreement with deviation of up to about 10%.

S2 The distribution of the total sharing

This section provides some additional results and discussion on *The total sharing distribution and an error model* section in the main text, in which an approximation to the distribution of the total sharing was presented.

S2.1 A bound on the probability of no sharing

A bound on the probability of no sharing, $P(f_T = 0)$, can be obtained directly from the one-sided Chebyshev inequality,

$$P\left(f_T \le \langle f_T \rangle - a\right) \le \frac{\sigma^2_{f_T}}{\sigma^2_{f_T} + a^2}. \tag{32}$$

Substituting $a = \langle f_T \rangle$ and noting that $P(f_T \le 0) = P(f_T = 0)$ immediately gives

$$P(f_T = 0) \le \frac{\sigma^2_{f_T}}{\sigma^2_{f_T} + \langle f_T \rangle^2}. \tag{33}$$

In practice, however, this bound is not very tight, as can be seen in Figure **S3**.

S2.2 IBD calculations in the founder model

The total sharing distribution and an error model section in the main text presented results for the distribution of total sharing assuming it is a sum of a Poisson distributed number of segments. Early calculations of the distribution of the total sharing were performed in a different population model, where a group of unrelated individuals is assumed to have recently founded the population. The distribution of the total length of the IBD shared segments was calculated, under somewhat strong assumptions, using renewal theory [6, 7]. In their model, it was assumed that if a region is not shared IBD, it is fully heterozygous (because it is

derived from different founders). In reality, however, all segments descend from a common ancestor at some point in the past, but the common ancestor of some segments is so ancient that they are too short to be detected. Our coalescent-based approach takes just that into account, by considering as IBD only segments longer than a certain length threshold.

S2.3 Matching the Poisson and exponential parameters

The parameters of the Poisson approximation, Eq. (24) in the main text, can be obtained by matching the first two moments of the total sharing distribution. The mean and variance of the Poisson approximation are given by (see, e.g., the main text Eq. (25))

$$\begin{aligned}\langle L_T\rangle &= n_0(\ell_0+m) = L\langle f_T\rangle,\\ \mathrm{Var}\left[L_T\right] &= n_0[\ell_0^2+(\ell_0+m)^2] = L^2\sigma_{f_T}^2,\end{aligned} \tag{34}$$

where $\langle f_T\rangle$ is given in the main text Eq. (4) and σ_{f_T} is given by one of the previously calculated approximations, e.g., the main text Eq. (15). Solving for n_0 and ℓ_0 in terms of $\langle f_T\rangle$ and σ_{f_T} gives

$$\begin{aligned}\ell_0 &= \frac{L\sigma_{f_T}^2 - 2\langle f_T\rangle m + \alpha}{4\langle f_T\rangle},\\ n_0 &= \frac{L\sigma_{f_T}^2 + 2\langle f_T\rangle m - \alpha}{2m^2/L},\end{aligned} \tag{35}$$

where $\alpha = \left(4\langle f_T\rangle Lm\sigma_{f_T}^2 + L^2\sigma_{f_T}^4 - 4\langle f_T\rangle^2 m^2\right)^{1/2}$. In practice, we found that using Eq. (35) matched well the distribution $P(f_T)$ only when we underestimated σ_{f_T} by 20-30%, probably because of the absence of the broad tail in Eq. (24). Therefore, in Figures 4 in the main text and **S2** here we used the fitted values of n_0 and ℓ_0.

S3 An estimator of the population size

In this subsection, we derive Eq. (39) in the main text for the variance of an estimator of the population size that is based on the average sharing between all pairs in a cohort. For a cohort of size n, define $\overline{\overline{f_T}} = \sum_{i=1}^{n}\sum_{j>i}^{n} f_T^{(i,j)}/\binom{n}{2}$, or

$$\overline{\overline{f_T}} = \frac{f_T^{(1,2)} + f_T^{(1,3)} + \cdots f_T^{(1,n)} + f_T^{(2,3)} + \cdots + f_T^{(2,n)} + \cdots + f_T^{(n-1,n)}}{\binom{n}{2}}. \tag{36}$$

The estimator takes the form

$$\hat{N} = \frac{100}{m\overline{\overline{f_T}}} - \frac{75}{m}. \tag{37}$$

The SD of $\hat{N}$ can be approximated as in the main text,

$$\sigma_{\hat{N}} \approx \frac{100}{m}\frac{\sigma_{\overline{\overline{f_T}}}}{\left\langle\overline{\overline{f_T}}\right\rangle^2}. \tag{38}$$

In fact, this approximation is better justified here than in the main text, as the distribution of $\overline{\overline{f_T}}$ is much narrower than that of f_T. Using $\left\langle\overline{\overline{f_T}}\right\rangle = \langle f_T\rangle \approx 100/(mN)$ gives

$$\sigma_{\hat{N}} \approx \frac{mN^2\sigma_{\overline{\overline{f_T}}}}{100}. \tag{39}$$

We therefore need to calculate the variance of $\overline{\overline{f_T}}$, from which we will then obtain the standard deviation $\sigma_{\overline{\overline{f_T}}}$. The variance of $\overline{\overline{f_T}}$ can be written as

$$\mathrm{Var}\left[\overline{\overline{f_T}}\right] = \text{var term} + \text{cov term}, \tag{40}$$

where the var term corresponds to the variances of the individual terms in the sum in the definition of $\overline{\overline{f_T}}$ (Eq. (36)), and the cov term corresponds to the covariances of these terms. More concretely, using Eq. (36),

$$\text{var term} = \frac{\binom{n}{2}\sigma^2_{f_T}}{\binom{n}{2}^2} = \frac{\sigma^2_{f_T}}{\binom{n}{2}} \approx \frac{2\sigma^2_{f_T}}{n^2} \approx \frac{2\cdot 100}{n^2 NL}\ln\left(\frac{L}{m}\right), \tag{41}$$

where we used Eq. (15) in the main text for $\sigma^2_{f_T}$. The covariance term is

$$\text{cov term} = \frac{\sum_{(i,j),i\neq j}\sum_{(k,l)\neq(i,j),k\neq l}\mathrm{Cov}\left[f_T^{(i,j)}, f_T^{(k,l)}\right]}{\binom{n}{2}^2}. \tag{42}$$

Note that the set (i,j,k,l) must have at least three distinct indexes. In most combinations of (i,j,k,l), we will have all i,j,k,l different, for which we assume that the covariance $\mathrm{Cov}\left[f_T^{(i,j)}, f_T^{(k,l)}\right]$ is zero. We therefore have to consider only covariances of the form $\mathrm{Cov}\left[f_T^{(i,j)}, f_T^{(i,k)}\right]$ and $\mathrm{Cov}\left[f_T^{(i,j)}, f_T^{(j,k)}\right]$. Since for each pair (i,j) (from which we have $\binom{n}{2}$) there are $(n-2)$ possible ks, we have

$$\text{cov term} \approx \frac{\binom{n}{2}2(n-2)\mathrm{Cov}\left[f_T^{(1,2)}, f_T^{(1,3)}\right]}{\binom{n}{2}^2} \approx \frac{4\mathrm{Cov}\left[f_T^{(i,j)}, f_T^{(i,k)}\right]}{n} \approx \frac{4\cdot 10000}{nN^2mL}, \tag{43}$$

where we used Eq. (27) in the main text for $\mathrm{Cov}\left[f_T^{(1,2)}, f_T^{(1,3)}\right]$. In total, the variance of $\overline{\overline{f_T}}$ is

$$\mathrm{Var}\left[\overline{\overline{f_T}}\right] \approx \frac{2\cdot 100}{n^2NL}\ln\left(\frac{L}{m}\right) + \frac{4\cdot 10000}{nN^2mL} = \frac{400}{nNL}\left[\frac{\ln\left(\frac{L}{m}\right)}{2n} + \frac{100}{Nm}\right], \tag{44}$$

and

$$\sigma_{\overline{\overline{f_T}}} \approx \frac{20}{\sqrt{nNL}}\sqrt{\frac{\ln\left(\frac{L}{m}\right)}{2n} + \frac{100}{Nm}}. \tag{45}$$

Finally,

$$\sigma_{\hat{N}} \approx \frac{mN^2\sigma_{\overline{\overline{f_T}}}}{100} \approx \frac{mN^{3/2}}{5\sqrt{nL}}\sqrt{\frac{\ln\left(\frac{L}{m}\right)}{2n} + \frac{100}{Nm}}, \tag{46}$$

which is precisely Eq. (39) in the main text.

S4 An admixture pulse

In the main text, an approximate solution was given for the integral in Eq. (43). The full solution is:

$$\begin{aligned}
\mathrm{Var}\left[f_T\right] &\approx 2\int_{m/L}^{1}(1-x)\left[\int_0^{T_a} e^{-t-txNL/50}dt\right]dx + 2\alpha^2\int_{m/L}^{1}(1-x)\left[\int_{T_a}^{\infty} e^{-t-txNL/50}dt\right]dx \\
&= \frac{100}{L^2N^2T_a}\left\{50(1-\alpha^2)\left[\exp\left(-\frac{T_a(50+NL)}{50}\right) - \exp\left(-\frac{T_a(50+Nm)}{50}\right)\right] - NT_a(L-m)\right. \\
&\left. + T_a(50+NL)\ln\left(\frac{50+NL}{50+Nm}\right) + T_a(50+NL)(1-\alpha^2)\left[E_i\left(-\frac{T_a(50+Nm)}{50}\right) - E_i\left(-\frac{T_a(50+NL)}{50}\right)\right]\right\},
\end{aligned} \tag{47}$$

where $E_i(x)$ is the exponential integral function. To obtain the simplified equation (43) of the main text, we assumed $T_a \ll 1$ (or $G_a = NT_a \ll N$), $m \ll L$, $mNT_a \ll 50$, $mN \gg 50$, and $LNT_a \gg 50$, and used the series expansion of the exponential integral. For the parameters for which we plotted the simulation results, the simplified expression deviates in no more than 1% from the full expression.

Simulations for the case of pulse admixture were performed using GENOME as described in the main text, with the following population history. The initial (current) population size was set to N, followed by splitting to two populations, at generation G_a, of relative sizes $N\alpha/(1-\alpha)$ and N, such that a fraction α of the linages descends from the first population (we could not find a way to implement the gene flow in GENOME while keeping the population size fixed). At the next generation, the first population size was reduced back to N, and the second was increased to 10^6, to practically eliminate IBD sharing within the second population. At generation 10^4, the two populations were merged into a single population of size N, to enable all linages to coalesce. Simulation results are presented in Figure **S10**A. Each data point corresponds to 500 runs. The apparent noise for large α might be due to this somewhat unnatural admixture model implementation.

References

[1] P. D. Keightley. Rates and fitness consequences of new mutations in humans. *Genetics*, 190:295–304, 2012.

[2] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado, USA, 2009.

[3] A. Gusev, J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, and I. Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, 19:318–326, 2009.

[4] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, 23:183–201, 1983.

[5] P. J. Fitzsimmons and J. Pitman. Kac's moment formula and the Feynman-Kac formula for additive functionals of a Markov process. *Stoch. Proc. Appl.*, 79:117–134, 1999.

[6] P. Stam. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res.*, 35:131–155, 1980.

[7] N. H. Chapman and E. A. Thompson. A model for the length of tracts of identity by descent in finite random mating populations. *Theor. Pop. Biol.*, 64:141–150, 2003.

[8] Y. Shen, R. Song, and I. Pe'er. Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association. *Bioinformatics*, 27:1995–1997, 2011.