Recombination, variance in genetic relatedness, and selection against introgressed DNA

Carl Veller[1,2,*] Nathaniel B. Edelman[1] Pavitra Muralidhar[1,2] Martin A. Nowak[1,2,3]

Abstract

The genomic proportion that two relatives share identically by descent—their genetic relatedness—can vary depending on the patterns of recombination and segregation in their pedigree. Here, we calculate the precise connection between genome-wide genetic shuffling and variance in genetic relatedness. For the relationships of grandparent-grandoffspring and siblings, the variance in genetic relatedness is a simple decreasing function of $\bar{r}$, the average proportion of locus pairs that recombine in gametogenesis. These formulations explain several recent observations about variance in genetic relatedness. They further allow us to calculate the neutral variance of ancestry among F2s in a hybrid cross, enabling F2-based tests for various kinds of selection, such as Dobzhansky-Muller incompatibilities and hybrid vigor. Our calculations also allow us to characterize how recombination affects the rate at which selection eliminates deleterious introgressed DNA after hybridization—by modulating the variance of introgressed ancestry across individuals. Species with low aggregate recombination rates, like *Drosophila*, purge introgressed DNA more rapidly and more completely than species with high aggregate recombination rates, like humans. These conclusions also hold for different genomic regions. Within the genomes of several species, positive correlations have been observed between local recombination rate and introgressed ancestry. Our results imply that these correlations can be driven more by recombination's effect on the purging of deleterious introgressed alleles than its effect in unlinking neutral introgressed alleles from deleterious alleles. In general, our results demonstrate that the aggregate recombination process—as quantified by $\bar{r}$ and analogs—acts as a variable barrier to gene flow between species.

[1]Department of Organismic and Evolutionary Biology, Harvard University, Massachusetts, USA
[2]Program for Evolutionary Dynamics, Harvard University, Massachusetts, USA
[3]Department of Mathematics, Harvard University, Massachusetts, USA
[*]carl.veller@gmail.com

1 Introduction

Variance in the amount of DNA shared by relatives identically by descent (IBD)—variance in genetic relatedness—is an important quantity in genetics (Thompson 2013). It translates to variance in the phenotypic similarity of relatives, and is a vital component of relatives-based estimates of heritability and the genetic variance that underlies traits (Visscher et al. 2006, 2007), and an important consideration when estimating pedigree relatedness and the degree of inbreeding from genotype data (Kardos et al. 2015; Wang 2016). Variance in genetic relatedness has also been hypothesized to drive the evolution of karyotypes and recombination rates in some clades (Sherman 1979; Wilfert et al. 2007).

For most pedigree relationships, genetic relatedness can vary because of variable patterns of recombination and segregation within the pedigree. For example, it is possible that a mother segregates only crossoverless paternal chromatids to an egg, in which case the resulting offspring inherits one half of its genome from its maternal grandfather and none from its maternal grandmother. On the other hand, if the mother shuffles her maternal and paternal DNA thoroughly into the egg, the offspring will be approximately equally related (genetically) to its maternal grandparents.

In theoretical calculations of the variance of genetic relatedness, it has typically been assumed that recombination is uniform along chromosomes and that crossover interference is absent [e.g., Franklin (1977); Hill (1993b); Guo (1996); Visscher et al. (2006); Hill and Weir (2011)]. White and Hill (2019) have recently developed a procedure to estimate the variance of genetic relatedness from linkage maps without the assumption of uniform recombination rates. However, their method assumes uniform recombination rates in regions between markers (restricting the method to high-density linkage maps) and ignores crossover interference.

In this paper, we derive a general, assumption-free formulation for the variance of genetic relatedness in terms of aggregate genetic shuffling. We demonstrate that the variance of genetic relatedness is a simple, decreasing function of certain newly-developed metrics of genome-wide genetic shuffling (Veller et al. 2019). This formulation allows effects on the variance of genetic relatedness to be reinterpreted—often more intuitively—in terms of effects on aggregate genetic shuffling. For example, it has recently been shown that crossover interference decreases the variance of genetic relatedness (Caballero et al. 2019). This can be explained by the intuitive fact that crossover interference, by spreading crossovers out evenly along chromosomes, increases the amount of genetic shuffling that they cause (Gorlov and Gorlova 2001; Veller et al. 2019).

Formulating the variance of genetic relatedness in terms of aggregate recombination also allows us to characterize how recombination influences the retention of introgressed DNA after hybridization, a topic of much recent interest [e.g., Schumer et al. (2018); Martin et al. (2019); Edelman et al. (2019)]. When introgressed DNA is deleterious to the recipient species, the rate at which selection purges it from the population is proportional to the variance of the amount of deleterious introgressed DNA carried by different members of the population (Harris and Nielsen 2016). The amount of introgressed DNA carried by an individual with hybrid ancestry can be interpreted as that individual's genetic relatedness to its hybridizing ancestor from the foreign species. Recombination affects the variance of genetic relatedness as characterized in this paper, and thus affects variance across individuals in the amount of deleterious introgressed DNA they carry. This insight enables us to investigate the factors that influence how effective the aggregate recombination process is as a barrier to gene flow between species.

In the calculations below, we assume that there is no inbreeding. The number of loci in the genome, L, is assumed to be very large, and loci i and j are recombinant in a random gamete with probability r_{ij} (e.g., $r_{ij} = 1/2$ if i and j are on different chromosomes).

2 Variance in genetic relatedness

2.1 Relationships of direct descent

Pedigree relationships of direct descent (or 'lineal' relationships) involve a single lineage, from an ancestor to one of its descendants. We will focus here on the particular example of grandparent-grandoffspring—calculations of the variance of genetic relatedness for general relationships of direct descent are given in SI Section S1.

Grandparent-grandoffpsring. Let the random variable IBD_{grand} be the proportion of a grandoffspring's genome inherited from a specified grandparent. We wish to calculate $\text{Var}(IBD_{\text{grand}})$. To give the flavor of the calculations used in this paper, we present the full derivation here. Derivations for all other cases can be found in SI Sections S1 and S2. Our approach is similar to that of Hill (1993a) and Visscher et al. (2006), although we do not make any assumptions about the recombination process.

Consider the gamete produced by the grandoffspring's parent (on the specified grandparent's side of the pedigree). Let the random variable $\hat{P}_k$ take the value 1 if the allele at locus k in this gamete derives from the grandparent, and 0 otherwise (hats will denote gametic values in this paper). Then $\mathbb{E}[\hat{P}_k] = 1/2$ and $\text{Var}(\hat{P}_k) = 1/4$ for each k. For the gametic alleles at loci i and j both to derive from the specified grandparent requires (a) that loci i and j be non-recombinant (probability $1 - r_{ij}$) and, given this, (b) that the specified grandparent's alleles segregated to the successful gamete (probability 1/2). Therefore,

$$\begin{aligned}\text{Cov}(\hat{P}_i, \hat{P}_j) &= \mathbb{E}[\hat{P}_i\hat{P}_j] - \mathbb{E}[\hat{P}_i]\mathbb{E}[\hat{P}_j] \\ &= \text{Prob}(\hat{P}_i = 1 \text{ and } \hat{P}_j = 1) - \mathbb{E}[\hat{P}_i]\mathbb{E}[\hat{P}_j] \\ &= \frac{1}{2}(1 - r_{ij}) - \frac{1}{4} = \frac{1}{2}\left(\frac{1}{2} - r_{ij}\right). \end{aligned} \tag{1}$$

Let $\hat{P} = \frac{1}{L}\sum_{i=1}^{L} \hat{P}_i$ be the proportion of the gamete's genome that derives from the focal grandparent. Then $\mathbb{E}[\hat{P}] = \frac{1}{L}\sum_{i=1}^{L} \mathbb{E}[\hat{P}_i] = 1/2$ and

$$\begin{aligned}\text{Var}(\hat{P}) = \frac{1}{L^2}\text{Var}\left(\sum_{k=1}^{L} \hat{P}_k\right) &= \frac{1}{L^2}\sum_{k=1}^{L} \text{Var}(\hat{P}_k) + \frac{1}{L^2}\sum_{i\neq j} \text{Cov}(\hat{P}_i, \hat{P}_j) \\ &= \frac{1}{4L} + \frac{1}{L^2}\sum_{i\neq j}\frac{1}{2}\left(\frac{1}{2} - r_{ij}\right) \\ &\xrightarrow[L\to\infty]{} \frac{1}{2}\left(\frac{1}{2} - \bar{r}\right), \end{aligned} \tag{2}$$

where $\bar{r}$ is the probability that a randomly chosen locus pair recombines in gametogenesis (Veller et al. 2019). The limit follows from the fact that, when L is large, there are $\sim L^2$ locus pairs (i, j) such that $i \neq j$. A graphical demonstration of Eq. (2), based on the possible segregation patterns of a given meiosis in the parent, is shown in Fig. 1.

Finally, because half of the grandoffspring's genome comes from this gamete, $IBD_{\text{grand}} = \hat{P}/2$, so that $\mathbb{E}[IBD_{\text{grand}}] = \mathbb{E}[\hat{P}]/2 = 1/4$ is the coefficient of relationship, and

$$\text{Var}(IBD_{\text{grand}}) = \frac{1}{4}\text{Var}(\hat{P}) = \frac{1}{8}\left(\frac{1}{2} - \bar{r}\right). \tag{3}$$

Note that the formulation in Eq. (3) and other such formulations in this paper apply to the whole genome, or a single chromosome, or any particular genomic region. In the latter cases, $\bar{r}$ is the probability that a random pair of loci in the region of interest recombine in gametogenesis. In addition,

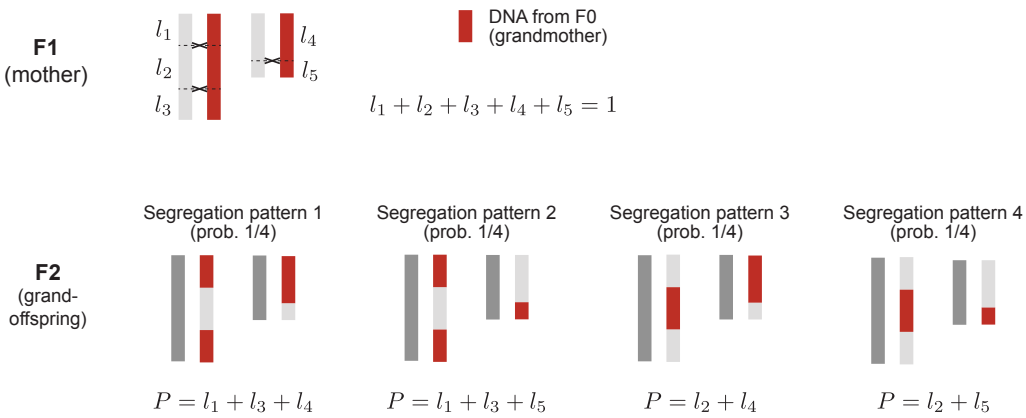

Figure 1: The variance of genetic relatedness between grandoffspring and grandparent, calculated from the possible segregation patterns of a single parental meiosis. In the figure, the positions of crossovers in a maternal meiosis (and the chromatids involved) are specified, but the segregation pattern in the resulting egg (and therefore offspring) is not. Averaging across the four segregation patterns, we find $\mathbb{E}[P] = (2l_1 + 2l_2 + 2l_3 + 2l_4 + 2l_5)/4 = 1/2$, and, from Eq. [1] in Veller et al. (2019), $\bar{r}^{\,♀} = (l_1 + l_3 + l_4)(l_2 + l_5) + (l_2 + l_4)(l_1 + l_3 + l_5)$. The variance of P associated with the possible segregation patterns is

$$\begin{aligned}\mathrm{Var}(P) = \mathbb{E}[P^2] - (\mathbb{E}[P])^2 &= \frac{1}{4}\left[(l_1 + l_3 + l_5)^2 + (l_1 + l_3 + l_4)^2 + (l_2 + l_4)^2 + (l_2 + l_5)^2\right] - \frac{1}{4} \\ &= \frac{1}{4} - \frac{1}{2}\left[(l_1 + l_3 + l_4)(l_2 + l_5) + (l_2 + l_4)(l_1 + l_3 + l_5)\right] = \frac{1}{2}\left(\frac{1}{2} - \bar{r}^{\,♀}\right),\end{aligned}$$

which is Eq. (3).

---

because the recombination process often differs between the sexes, the value of $\bar{r}$ can differ between spermatogenesis and oogenesis. In calculating the variance in genetic relatedness between a grandoffspring and one of its maternal grandparents, the value for oogenesis, $\bar{r}^{\,♀}$, would be used; the value for spermatogenesis, $\bar{r}^{\,♂}$, would be used for paternal grandparents.

$\bar{r}$ in Eq. (3) can be estimated from various kinds of data, including cytological data of crossover positions at meiosis I, sequence data from gametes, and linkage maps (Veller et al. 2019). For example, Veller et al. (2019) used cytological data from Lian et al. (2008) to estimate an autosomal value for human males of $\bar{r}^{\,♂} = 0.4873$. Substituting this value into Eq. (3) reveals that the variance of the (autosomal) genetic relatedness of a grandoffspring to its paternal grandparent is $\mathrm{Var}(IBD_{\mathrm{grand}}) = 1.6 \times 10^{-3}$, corresponding to a standard deviation of 0.04, or a coefficient of variation of 16%.

2.2 Indirect relationships

Indirect relationships involve multiple descendants of at least one individual in the pedigree. We will focus here on siblings and half-siblings—the calculation for general indirect relationships is given in SI Section S2.

Siblings. Let the random variable IBD_{sibs} be the proportion of two full-siblings' genomes that they share IBD, assuming their mother and father to be unrelated. Then $\mathbb{E}[IBD_{\mathrm{sibs}}] = 1/2$ is the coefficient of relationship, and

$$\mathrm{Var}(IBD_{\mathrm{sibs}}) = \frac{1}{8}\left(1 - \bar{r}^{\,♀}_{(2)} - \bar{r}^{\,♂}_{(2)}\right), \tag{4}$$

where $\bar{r}^{\,♀}_{(2)}$ is the probability that a randomly chosen locus pair recombines in an egg when the crossovers

of two of the mother's meioses are pooled into one meiosis, and $\bar{r}^{\♂}_{(2)}$ is the analogous quantity for the father (see Fig. 2 for an example of a pooled meiosis).

Half-siblings. Let the random variable $IBD_{\text{h-sibs}}$ be the proportion of two half-siblings' genomes that they share IBD, assuming that they have the same father but unrelated mothers. Then $\mathbb{E}[IBD_{\text{h-sibs}}] = 1/4$ is the coefficient of relationship, and

$$\text{Var}(IBD_{\text{h-sibs}}) = \frac{1}{2}\left(\frac{1}{2} - \bar{r}^{\♂}_{(2)}\right). \tag{5}$$

If the common parent were instead the mother, $\bar{r}^{\♀}_{(2)}$ would replace $\bar{r}^{\♂}_{(2)}$ in Eq. (5). A graphical demonstration of Eq. (5), based on the possible segregation patterns of two separate meioses in the parent, is given in Fig. 2.

Like $\bar{r}$, $\bar{r}_{(2)}$ can be estimated from various kinds of data, including cytological data of crossover positions at meiosis I and sequence data from gametes. Using cytological data for human male spermatocytes from Lian et al. (2008), we construct a pooled meiosis from every possible pair of spermatocytes. Calculating the value of $\bar{r}$ for each of these pooled meioses and averaging, we obtain $\bar{r}^{\♂}_{(2)} = 0.4912$. Thus, from Eq. (5), the genetic relatedness of half-sibs who share a father but have unrelated mothers has variance 1.1×10^{-3}, i.e., a standard deviation of 0.033, or a coefficient of variation of about 13%.

2.3 Application: Ancestry variance among F2s.

A common experimental design involves mating individuals from two populations or species (A and B) to form a hybrid 'F1' generation, and then mating the F1s to produce an F2 generation. Each individual in the F1 generation carries exactly one-half of its DNA from species A—i.e., there is no variance in ancestry among F1s—but there is variance in ancestry among F2s because of recombination and segregation in the F1s' meioses (Hill 1993a). Each F2 is produced by an egg from an F1 mother and a sperm from an F1 father. Let the random variables $\hat{P}^{\♀}$ and $\hat{P}^{\♂}$ be the proportion of species-A DNA in the egg and sperm, respectively, and let P be the proportion of species-A DNA in an F2's genome. Then $P = \hat{P}^{\♀}/2 + \hat{P}^{\♂}/2$, and, from Eq. (2), $\text{Var}\big(\hat{P}^{\♀}\big) = \frac{1}{2}\big(\frac{1}{2} - \bar{r}^{\♀}\big)$ and $\text{Var}\big(\hat{P}^{\♂}\big) = \frac{1}{2}\big(\frac{1}{2} - \bar{r}^{\♂}\big)$. Finally, because $\hat{P}^{\♀}$ and $\hat{P}^{\♂}$ are independent, the ancestry variance among F2s is

$$\text{Var}(P) = \frac{1}{4}\left(\text{Var}\big(\hat{P}^{\♀}\big) + \text{Var}\big(\hat{P}^{\♂}\big)\right) = \frac{1}{8}\left(1 - \bar{r}^{\♀} - \bar{r}^{\♂}\right). \tag{6}$$

This calculation assumes that no other forces are affecting ancestry among F2s. Such forces could include systematic selection among F2s in favor of alleles from one of the species, or meiotic drive in F1s, both of which would skew the distribution of ancestry among F2s towards one of the two species. A typical test for such forces would then involve comparing the mean ancestry against the neutral null expectation of 1/2. In this case, Eq. (6) gives the appropriate null variance for the purpose of statistical inference (the standard error of the test is $SE = \frac{1}{\sqrt{8}n}\sqrt{1 - \bar{r}^{\♀} - \bar{r}^{\♂}}$, where n is the number of F2s for which ancestry proportions have been measured).

There are alternative modes of selection under which the mean ancestry proportion among F2s remains 1/2 but the variance is skewed from the neutral expectation given in Eq. (6). For example, if selection acts on the basis of pairwise Dobzhansky-Muller incompatibilities, F2s with even ancestry are expected to be less fit than those with more skewed ancestry (because the number of incompatible pairs is proportional to $P(1-P)$, which is maximized at $P = 1/2$). If they are genotyped after selection has acted, the distribution of ancestry will have greater variance than predicted by Eq. (6).

Alternatively, if there is hybrid vigor, then the quantity relevant for selection among F2s is the proportion of the genome that is heterozygous. Because F2s with more even ancestry are likely to be

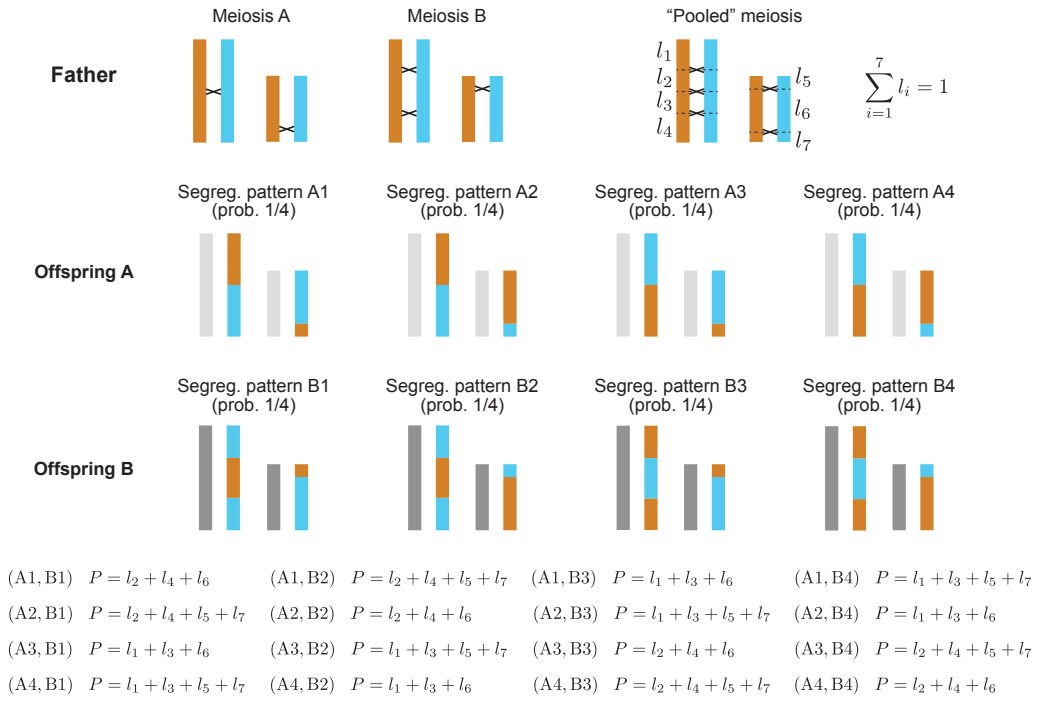

Figure 2: The variance of genetic relatedness between half-siblings, calculated from the possible segregation patterns of two meioses of their common parent. The positions of crossovers in two paternal meioses (and the chromatids involved) are specified, but the segregation patterns in the resulting sperm cells (and therefore the two offspring) are not. Averaging across the sixteen segregation patterns $(\mathrm{A}i, \mathrm{B}j)$, we find $\mathbb{E}[P] = (8l_1 + 8l_2 + 8l_3 + 8l_4 + 8l_5 + 8l_6 + 8l_7)/16 = 1/2$. Applying Eq. [1] in Veller et al. (2019) to the 'pooled meiosis' in which the crossovers from the two actual meioses have been combined, we find

$$\begin{aligned}\bar{r}^{\,\mathrm{♂}}_{(2)} &= \frac{1}{4}2(l_1 + l_3 + l_5 + l_7)(l_2 + l_4 + l_6) + \frac{1}{4}2(l_1 + l_3 + l_6)(l_2 + l_4 + l_5 + l_7) \\ &\quad + \frac{1}{4}2(l_2 + l_4 + l_5 + l_7)(l_1 + l_3 + l_5 + l_7) + \frac{1}{4}2(l_2 + l_4 + l_5 + l_7)(l_1 + l_3 + l_6) \\ &= (l_1 + l_3 + l_5 + l_7)(l_2 + l_4 + l_6) + (l_2 + l_4 + l_5 + l_7)(l_1 + l_3 + l_5 + l_7).\end{aligned}$$

The variance of P associated with the possible segregation patterns is

$$\begin{aligned}\mathrm{Var}(P) = \mathbb{E}[P^2] - (\mathbb{E}[P])^2 &= \frac{1}{4}\left[(l_1 + l_3 + l_5 + l_7)^2 + (l_2 + l_4 + l_6)^2 + (l_1 + l_3 + l_6)^2 + (l_2 + l_4 + l_5 + l_7)^2\right] - \frac{1}{4} \\ &= \frac{1}{4}\left[1 - 2(l_1 + l_3 + l_5 + l_7)(l_2 + l_4 + l_6)\right] + \frac{1}{4}\left[1 - 2(l_1 + l_3 + l_6)(l_2 + l_4 + l_5 + l_7)\right] - \frac{1}{4} \\ &= \frac{1}{4} - \frac{1}{2}\left[(l_1 + l_3 + l_5 + l_7)(l_2 + l_4 + l_6) + (l_1 + l_3 + l_6)(l_2 + l_4 + l_5 + l_7)\right] = \frac{1}{2}\left(\frac{1}{2} - \bar{r}^{\,\mathrm{♂}}_{(2)}\right),\end{aligned}$$

which is Eq. (4).

---

heterozygous at more sites, this mode of selection is expected to reduce ancestry variance among F2s below the level predicted by Eq. (6).

In testing for deviations of the variance of F2 ancestry from the null expectation, higher-order moments than the second are required for precise inference. These moments can be estimated computationally given sufficient information about the recombination process in males and females. Note, however, that higher moments will depend on meiotic features such as crossover interference along chromosomes and crossover covariation across chromosomes (Wang et al. 2019; see Discussion). To estimate higher moments from the recombination process taking into account these meiotic features,

one could use simulations of the beam-film model, a physical model of recombination that can be computationally calibrated to accurately reproduce crossover distributions (White et al. 2017).

Importantly, the tests described above can all be carried out for specific genomic regions by using region-specific values of $\bar{r}$ in Eq. (6).

In the case of hybrid vigor, we can be precise about the mean and variance of heterozygosity itself under the null hypothesis of no selection, and can further use these to derive an estimator for the strength of hybrid vigor. Let π be the proportion of loci at which randomly selected haploid genomes from species A and B have different alleles (i.e., the average level of heterozygosity among F1s), and let the random variable H be the proportion of loci that are heterozygous in an F2 zygote. The proportion of the zygote's genome homozygous by descent, F, is the proportion of loci with identical ancestry between the egg and sperm that produced the F2. $H = \pi(1-F)$, so that, under the neutral null, the average proportion of loci that are heterozygous in an F2 at the time of genotyping is $\mathbb{E}[H] = \pi/2$, and, using a calculation similar to Eq. (5), the variance is $\mathrm{Var}(H) = \frac{\pi^2}{2}\left(\frac{1}{2} - \bar{r}^{♀♂}_{(2)}\right)$, where $\bar{r}^{♀♂}_{(2)}$ is the $\bar{r}$ value that results from pooling the crossovers of a random female meiosis and a random male meiosis. $\mathrm{Var}(H)$ has been calculated by Franklin (1977) in terms of total map length, assuming a uniform recombination rate and no crossover interference.

Now let the random variable H' be the proportion of heterozygous loci in an F2 adult at the time of genotyping. If selection acts additively across loci so that an individual with a proportion h of heterozygous loci has relative viability $1 + Sh$, then it can be shown (SI Section S4) that

$$\mathbb{E}[H'] - \mathbb{E}[H] = \frac{S\mathrm{Var}(H)}{1 + S\mathbb{E}[H]}, \tag{7}$$

so that

$$\mathbb{E}[H'] = \frac{1}{2} + \frac{\pi^2 S}{2+S}\left(\frac{1}{2} - \bar{r}^{♀♂}_{(2)}\right). \tag{8}$$

From this expression, we can derive an F2-based estimator for S, the strength of hybrid vigor:

$$S = \frac{2\mathbb{E}[H'] - 1}{\pi^2\left(\frac{1}{2} - \bar{r}^{♀♂}_{(2)}\right) - \left(\mathbb{E}[H'] - \frac{1}{2}\right)}. \tag{9}$$

3 Selection against introgressed DNA

DNA introgressed from one species into another is often deleterious to the recipient species, either because the introgressed DNA is incompatible with the recipient species' genome or ecology, or because of higher genetic load in the donor species [reviewed by Martin and Jiggins (2017)].

Recombination can influence the retention of introgressed DNA because it allows neutral (and beneficial) introgressed alleles to recombine away from deleterious introgressed alleles before the deleterious alleles are eliminated (Brandvain et al. 2014; Schumer et al. 2018).

Recombination also affects the purging of the deleterious alleles themselves. The rate at which deleterious introgressed DNA is purged is determined by the variance of the amount of deleterious introgressed DNA carried by different members of the population (Harris and Nielsen 2016). The amount of introgressed DNA that an individual carries can be thought of as that individual's genetic relatedness to its hybridizing ancestor from the donor species. Therefore, because recombination affects the variance of this genetic relatedness, it affects the rate of purging of deleterious introgressed DNA. The calculations above allow us to characterize the role of recombination in this process.

We shall study a simple model in which selection acts additively against introgressed DNA: if a proportion p of an individual's genome is introgressed, its fitness is $1 - pS$. This is the additive version of the model in Barton (1983) and Barton and Bengtsson (1986), and corresponds to a situation where introgressed alleles at a large number of loci are deleterious in the recipient species, with fitness effects

additive at and across loci. In this model, loci are assumed to be uniformly spaced throughout the genome, although relaxing this assumption simply requires reinterpreting $\bar{r}$ and its analogs as averages taken across all locus pairs chosen from the set of loci at which selection is acting. For simplicity, we ignore sex chromosomes, though these often show distinctive signs of selection against introgressed DNA (Martin and Jiggins 2017).

Let P_t be the introgressed proportion of a random generation-t individual's genome (at zygote stage, before selection has acted). Then it can be shown (SI Section S5) that the amount of introgressed DNA purged by selection in the t-th generation is

$$\mathbb{E}\left[P_t\right] - \mathbb{E}\left[P_{t+1}\right] = \frac{S\mathrm{Var}\left(P_t\right)}{1 - S\mathbb{E}\left[P_t\right]}. \tag{10}$$

3.1 Initial purging of introgressed DNA

Selection in the first generation after hybridization. Similar to previous work [e.g., Harris and Nielsen (2016); Juric et al. (2016); Steinrücken et al. (2018)], we assume that hybridization occurs as a pulse in a single generation (F0), such that a fraction x of F1 offspring are hybrids. Then, because all F1 hybrids carry exactly one-half introgressed DNA, $\mathbb{E}[P_1] = x/2$ and $\mathrm{Var}(P_1) = x(1-x)/4$, so that, from Eq. (10),

$$\mathbb{E}\left[P_2\right] = \frac{x}{2} - \frac{x(1-x)S/4}{1 - xS/2} = \frac{x(1-S/2)}{2 - xS} =: \frac{y}{2}, \tag{11}$$

where y can be interpreted as the fraction of successful gametes that are produced by hybrids. The proportion of introgressed DNA removed by selection in the first generation is

$$\Delta_1 := \frac{\mathbb{E}[P_1] - \mathbb{E}[P_2]}{\mathbb{E}[P_1]} = \frac{x(1-x)S/4}{1 - xS/2} \Big/ \frac{x}{2} = \frac{(1-x)S}{2 - xS}. \tag{12}$$

Notice that all of the variance in the amount of introgressed DNA carried by F1 individuals is due to differences between hybrids and non-hybrids—there is no contribution from variance among hybrids, since they all carry exactly one-half introgressed DNA.

Selection in the second generation after hybridization. Some F2s will have hybrid parents; recombination in the meioses of these parents will affect the variance of the amount of introgressed DNA the F2s carry. To find $\mathrm{Var}(P_2)$, let $\hat{P}_1$ be the fraction of introgressed DNA in a successful gamete from generation 1 (i.e., after selection has acted). Clearly $\mathbb{E}[P_2] = \mathbb{E}[\hat{P}_1]$. The gamete's genome is inherited by an F2 individual from a particular F1 parent. If the F1 parent is a hybrid, the F2 offspring has a grandparent from the donor species, and we can interpret the proportion of introgressed DNA in the gamete, $\hat{P}_1$, as $\hat{P}$ from Section 2.1. Therefore, among gametes produced by F1 hybrids (a fraction y of all gametes), $\mathbb{E}[\hat{P}_1 \,|\, \mathrm{hybrid}] = 1/2$, and, from Eq. (2), $\mathrm{Var}(\hat{P}_1 \,|\, \mathrm{hybrid}) = \frac{1}{2}\left(\frac{1}{2} - \bar{r}\right)$. So

$$\begin{aligned}\mathbb{E}[\hat{P}_1^2 \,|\, \mathrm{hybrid}] &= \mathrm{Var}(\hat{P}_1 \,|\, \mathrm{hybrid}) + \Big(\mathbb{E}[\hat{P}_1 \,|\, \mathrm{hybrid}]\Big)^2 \\ &= \frac{1}{2}\left(\frac{1}{2} - \bar{r}\right) + \frac{1}{4} = \frac{1}{2}\left(1 - \bar{r}\right),\end{aligned} \tag{13}$$

where $\bar{r} = (\bar{r}^{\,♀} + \bar{r}^{\,♂})/2$ is the sex-averaged value. From this and the fact that $\mathbb{E}[\hat{P}_1] = \mathbb{E}[P_2] = y/2$,

$$\begin{aligned}\mathrm{Var}(\hat{P}_1) &= \mathbb{E}(\hat{P}_1^2) - \Big(\mathbb{E}[\hat{P}_1]\Big)^2 \\ &= y\mathbb{E}[\hat{P}_1^2 \,|\, \mathrm{hybrid}] - y^2/4 \\ &= y(1-\bar{r})/2 - y^2/4 \\ &= \frac{1}{4}y(1-y) + \frac{1}{2}y\left(\frac{1}{2} - \bar{r}\right).\end{aligned} \tag{14}$$

The first term in Eq. (14) is the contribution to total variance owing to the fact that some gametes are produced by hybrids and some are not, while the second term is the contribution from variance among gametes produced by hybrids.

Assuming random mating, an F2's genome is created by sampling two F1 gametes independently. So $P_2 = \hat{P}_1^{(1)}/2 + \hat{P}_1^{(2)}/2$, where the $\hat{P}_1^{(i)}$ are independent random variables with the same distribution as $\hat{P}_1$. Therefore,

$$\begin{aligned}\mathrm{Var}\left(P_2\right) &= \mathrm{Var}\left(\hat{P}_1^{(1)}/2 + \hat{P}_1^{(2)}/2\right) = \frac{1}{2}\mathrm{Var}(\hat{P}_1) \\ &= \frac{1}{8}y(1-y) + \frac{1}{4}y\left(\frac{1}{2} - \bar{r}\right),\end{aligned} \tag{15}$$

where the first term in Eq. (15) is the contribution to total variance owing to some individuals having hybrid ancestry and others not, while the second term is the contribution from variance among those with hybrid ancestry.

From Eqs. (15) and (10),

$$\mathbb{E}[P_3] = \frac{y}{2} - \frac{S\left[\frac{1}{8}y(1-y) + \frac{1}{4}y\left(\frac{1}{2} - \bar{r}\right)\right]}{1 - \frac{yS}{2}}, \tag{16}$$

so that the fraction of remaining introgressed DNA purged by selection in the second generation is

$$\begin{aligned}\Delta_2 &= \frac{\mathbb{E}[P_2] - \mathbb{E}[P_3]}{\mathbb{E}[P_2]} \\ &= \frac{S}{1 - yS/2}\left[\frac{1}{8}y(1-y) + \frac{1}{4}y\left(\frac{1}{2} - \bar{r}\right)\right] \Big/ \frac{y}{2} \\ &= \frac{\frac{1}{4}\left(1-y\right)S}{1 - yS/2} + \frac{\frac{1}{2}\left(\frac{1}{2} - \bar{r}\right)S}{1 - yS/2}.\end{aligned} \tag{17}$$

The first term in Eq. (17) is the effect of selection acting on variation between individuals with hybrid ancestry and those without, while the second term is the effect of selection acting on variation among individuals with hybrid ancestry. The importance of the second source of variation relative to the first is given by their ratio,

$$\frac{1 - 2\bar{r}}{1 - y}. \tag{18}$$

3.2 Example: Human vs. *Drosophila*.

To gain insight into the practical influence of the recombination process on the purging of deleterious introgressed DNA, we can compare the recombination processes of humans and *Drosophila melanogaster*. *D. melanogaster* has only two major autosomes and no crossing over in males. Aggregate genetic shuffling is therefore low. Using chromosome lengths from Release 6 of the *D. melanogaster* reference genome (Hoskins et al. 2015) and the female linkage map produced by Comeron et al. (2012), we calculate autosomal values of $\bar{r}^{\,♂} = 0.253$ and $\bar{r}^{\,♀} = 0.358$. The sex-averaged value is $\bar{r} = 0.305$. Humans, on the other hand, have 22 autosomes, causing aggregate genetic shuffling to be high. Using chromosome lengths from assembly GRCh38.p11 of the human reference genome and the linkage maps produced by Kong et al. (2010), we calculate autosomal values of $\bar{r}^{\,♂} = 0.485$ and $\bar{r}^{\,♀} = 0.491$, for a sex-averaged value of $\bar{r} = 0.488$. These calculations assume no crossover interference, to match our simulations below. In particular, map distances between non-adjacent loci were translated to recombination rates using Haldane's mapping function, which ignores crossover interference. (Using Kosambi's mapping function, which does take crossover interference into account, the values are

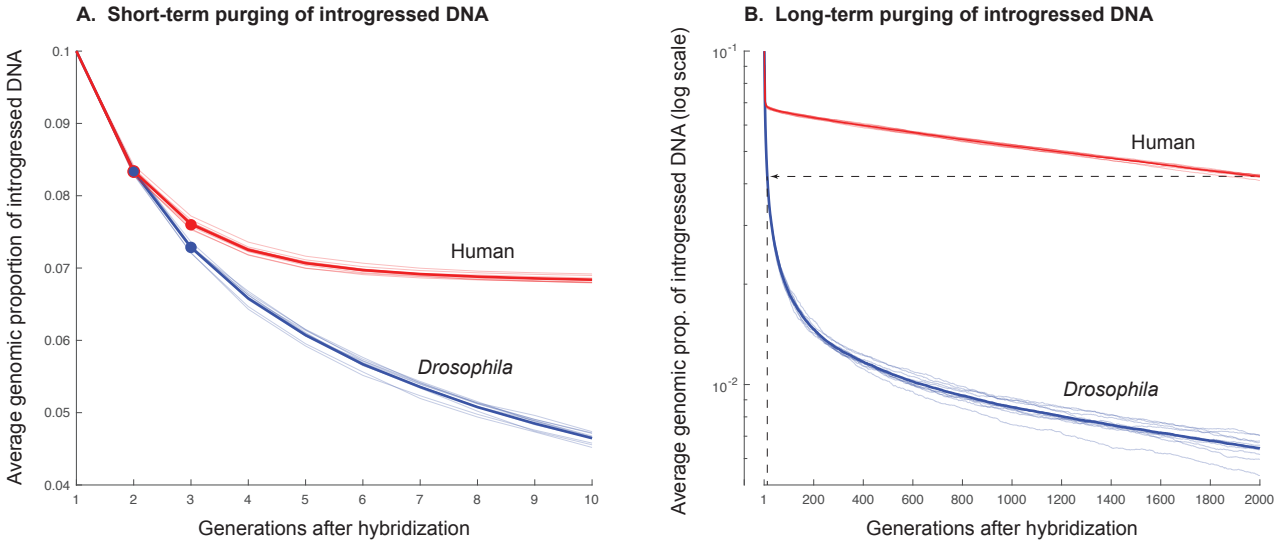

Figure 3: The rate at which introgressed DNA is purged following hybridization, given the recombination processes of humans and *D. melanogaster*. In **A**, the dots are analytical expectations calculated from Eq. (11) (generation 2) and Eq. (16) (generation 3). The bold lines are average trajectories calculated from 10 independent simulations of the model (faint trajectories). *Drosophila*'s recombination process causes much more introgressed ancestry to be purged in the early generations, because it is associated with a lower aggregate rate of genetic shuffling (lower value of $\bar{r}$ and analogs), driven largely by the small karyotype of *Drosophila* (2 major autosomes) relative to humans (22 autosomes). The dotted line in **B** shows that *Drosophila* purges as much introgressed DNA in 13 generations as humans do in 2,000 generations.

$\bar{r}^{\♂} = 0.253$, $\bar{r}^{\♀} = 0.376$, $\bar{r} = 0.314$ for *D. melanogaster*, and $\bar{r}^{\♂} = 0.487$, $\bar{r}^{\♀} = 0.493$, $\bar{r} = 0.490$ for humans).

While the calculations in Section 3.1 predict the rate of purging in the first few generations after hybridization, the interaction of selection and recombination in later generations becomes complicated, and so we turn to computer simulations, making use of the SLiM 3 simulation software (Haller and Messer 2019) (all code used in this paper can be found at github.com/nbedelman/IBD). This requires that we additionally specify the number of loci at which introgressed alleles are deleterious. Suppose that, after the introgression pulse, a fraction $x = 0.2$ of F1 zygotes are hybrids. This initial introgression fraction is consistent with that estimated by Harris and Nielsen (2016) for Neanderthal DNA introgressed into non-African humans, although the simulations of Harris and Nielsen (2016) begin with 10% individuals with complete Neanderthal ancestry, rather than 20% F1 hybrids (which causes the rate of purging in their simulations initally to be higher than in ours—see Discussion). The introgressed alleles at $L = 1{,}000$ loci are deleterious, with these loci evenly spaced throughout the (autosomal) genome. An individual carrying a fraction p of introgressed DNA has relative fitness $1 - 0.4p$, so that $S = 0.4$, with the deleterious fitness effect of each individual allele being $s = S/(2L) = 2 \times 10^{-4}$. These values of L and s are consistent with several estimates for Neanderthal-human introgression (Harris and Nielsen 2016; Juric et al. 2016), and are similar to estimates for other species as well [e.g., Aeschbacher et al. (2017)]. Recombination rates between adjacent loci are sex-specific, and are interpolated from the linkage maps mentioned above. We assume no crossover interference along chromosomes, and we further assume no crossover covariation across chromosomes (Wang et al. 2019). The population size in our simulations is $N = 10^5$.

Fig. 3 shows trajectories of the fraction of introgressed ancestry after a hybridization pulse in generation 0. Several features of these trajectories are noteworthy.

First, most of the purging of introgressed DNA occurs in the first few generations after hybridization. In the human population, more than half of the introgressed DNA ultimately purged by gener-

ation 2,000 was purged in the first 5 generations (7 for *Drosophila*). This effect has been observed in previous simulation studies (Harris and Nielsen 2016; Schumer et al. 2018; Petr et al. 2019).

Second, in these first few generations, introgressed DNA is purged more rapidly in *Drosophila* than in humans, leading to a much more profound initial depletion of introgressed ancestry. Our calculations above imply that this is due to the lower value of aggregate genetic shuffling in *Drosophila*. In both populations, the fraction of introgressed DNA decreases from 10% in the first generation to 8.3% in the second generation [Eq. (11); Fig. 3A]. This reduction is independent of the recombination process [Eq. (11)]. However, substituting into Eq. (16) the sex-averaged autosomal values of $\bar{r}$ calculated above, we find that the proportion of introgressed DNA decreases from 8.3% in the second generation to a third-generation value of 7.6% in humans and 7.3% in *Drosophila*, reductions of 8.8% and 12.5% respectively (Fig. 3A). The reduction in *Drosophila* is greater because of its lower value of $\bar{r}$. Using Eq. (18), we can compare the relative importance of the recombination-mediated source of variance for second-generation purging. In humans, the former is only 1/40 as important as variance due to some individuals having hybrid ancestry and some not, while in *Drosophila*, it is about 1/2 as important. Thus, recombination's effect on variance in the ancestry of second-generation individuals is about 20 times more important in *Drosophila* than in humans.

Third, purging eventually slows down to approximately the same rate in humans and *Drosophila* (Fig. 3B). Measurement reveals this apparently asymptotic rate of purging to be $\sim 2 \times 10^{-4}$ $(= s)$.

Overall, despite the fact that the rate of purging eventually converges to the same value in humans and *Drosophila*, the higher initial rate in *Drosophila* ensures that, ultimately, substantially more introgressed DNA is purged than in humans. Thus, after 2,000 generations, *Drosophila* retains less than 7% of the initial introgressed DNA, while the human population retains more than 40% (Fig. 3B). Put differently, the *Drosophila* population takes just 13 generations to purge the same amount of introgressed DNA that it takes the human population 2,000 generations to purge. Therefore, the recombination process of *Drosophila* acts as a much more effective barrier to gene flow.

3.3 A unified understanding of short-term and long-term purging

Recombination slows down the purging of introgressed DNA because it takes the initial large quantities in a few individuals and shuffles them into many individuals, reducing the variance across individuals in the amount of introgressed DNA that they carry (Harris and Nielsen 2016). From a different perspective, the effect of recombination is to chop up the initially large blocks of introgressed DNA into smaller and smaller blocks. Here, we define 'blocks' as sets of introgressed alleles co-transmitted from the same hybridizing ancestor. Thus, introgressed blocks in F1s are entire haploid genomes, while in F2s, they are patchworks across chromosomes. Therefore, blocks need not be contiguous stretches of introgressed DNA, but almost all blocks will in fact be so after a few generations.

We can distinguish two sources of variance across individuals in how much introgressed DNA they carry: (i) variance in the number of blocks carried; (ii) variance in the length of blocks. We are interested in the effect of recombination on these two sources of variance.

In the early generations after hybridization, the number of blocks that an individual carries depends almost entirely on its pedigree, and not on the particular patterns of recombination within that pedigree. For example, a second-generation individual will carry 0, 1, or 2 blocks if, respectively, 0, 1, or 2 of its parents were F1 hybrids. This is because, under almost all recombination processes, an F1 hybrid is nearly certain to transmit some introgressed DNA to an offspring. Similarly, an F2 with two blocks is almost certain to transmit two (smaller) blocks to a third-generation offspring. Therefore, while recombination plays a role here in chopping up blocks and distributing the smaller blocks among a greater number of progeny, it is a relatively invariant role. Instead, the variable role of the recombination process in these early generations is to generate block length variation. This is especially clear in the case where all F1 individuals are hybrids. Then there is essentially no block number variance in each of the early generations, and so the variance in introgressed ancestry that

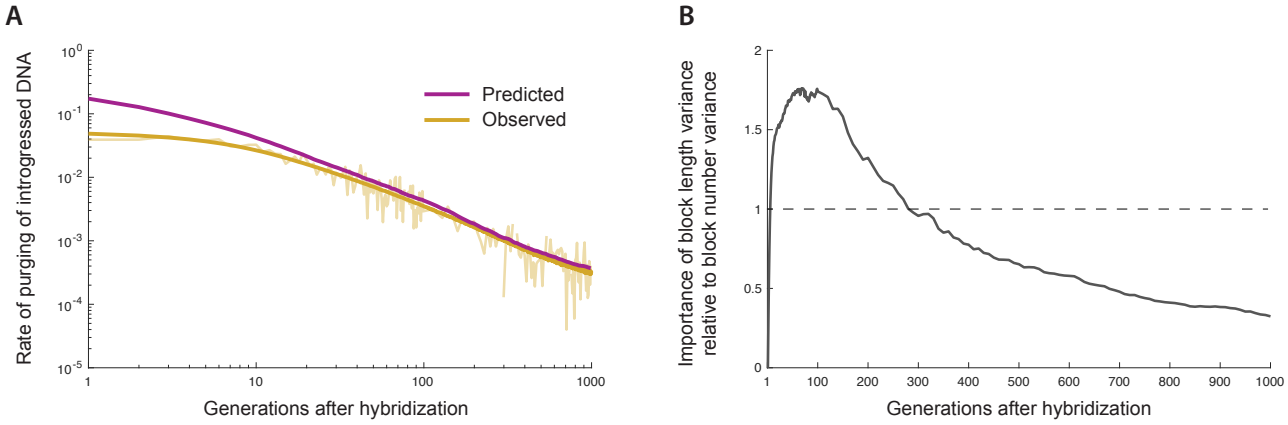

Figure 4: The distribution of introgressed block lengths determines the rate of purging of introgressed DNA. The recombination process is that of *D. melanogaster*, and we assume that all F1s are hybrids ($x = 1$). For computational reasons, the population size is $N = 10{,}000$, rather than the value of $N = 100{,}000$ used elsewhere in this paper. **A**. Observed rate of purging vs. the prediction of Eq. (19). The purple and faded gold trajectories are averages calculated from 150 replicate simulations in which block lengths were tracked, while the bold gold trajectory is an average calculated from 10,000 independent replicates in which block lengths were not tracked. Eq. (19) overestimates the rate of purging in the generations immediately after hybridization, as expected (see text), but becomes more accurate in later generations. This accuracy validates the model we have used to understand the purging of introgressed ancestry in terms of the distribution of block lengths (SI Section S6) for all but the earliest generations after hybridization. **B**. Relative importance of the two sources of variance in introgressed ancestry, calculated using Eq. (23). Trajectory is averaged across the 150 replicate simulations in which block lengths were tracked. Block length variance is important in the early generations, implicating the aggregate recombination process in the purging of introgressed DNA, while block number variance is more important in the later generations, implicating fine-scale recombination rates. Although the figure does not display this, block length variance is in fact substantially more important than block number variance in the earliest generations after hybridization, for reasons explained in the text. The discrepancy arises because the assumptions of the model used to derive Eq. (23) do not hold in the earliest generations (SI Section S6).

selection acts upon is provided entirely by the recombination process's ability to generate block length variance.

Eventually, however, blocks become sufficiently mixed among the population that they can be assumed to have been inherited independently from one another. In this case, as shown in SI Section S6, the distribution of block lengths determines the rate of purging of introgressed DNA according to

$$\Delta_t = \frac{\mathbb{E}[P_t] - \mathbb{E}[P_{t+1}]}{P_t} = \frac{S\overline{l_t^2}/\bar{l}_t}{1 - S\mathbb{E}[P_t]}, \tag{19}$$

where P_t is the proportion of introgressed ancestry of a random generation-t individual, S is the fitness disadvantage of an individual with entirely introgressed ancestry, and $\bar{l}_t$ and $\overline{l_t^2}$ are the averages of the block length and squared block length respectively. Fig. 4A shows the numerical accuracy of Eq. (19) for *Drosophila*'s recombination process, when all F1s are hybrids. When $S\mathbb{E}[P_t]$ is small, Eq. (19) can be approximated by

$$\Delta_t \approx S\overline{l_t^2}/\bar{l}_t. \tag{20}$$

If all blocks are the same length, then $0 = \mathrm{Var}(l_t) = \overline{l_t^2} - (\bar{l}_t)^2$, so that $\overline{l_t^2} = (\bar{l}_t)^2$, and the rate of purging is $\Delta_t = S\bar{l}_t$. This is the case when, eventually, recombination has dissociated all introgressed alleles from one another, so that every block is 1 locus long ($l_\infty = 1/[2L]$). The rate of purging is then

$$\Delta_\infty = Sl_\infty = S/(2L) = s, \tag{21}$$

the asymptotic value observed in our simulations above (Fig. 3B). [This asymptotic rate will be reached as long as the population is not so small that genetic drift dominates selection on individual alleles ($Ns \ll 1$).] Therefore, eventually, all of the ancestry variance across individuals is due to (single-locus) block number variance. This is in contrast to the early generations, where block length variance contributes substantially to overall ancestry variance. We can compare the contributions of these two sources of variance generally by decomposing Eq. (20) as follows:

$$\Delta_t \approx S\overline{l_t^2}/\bar{l}_t = S\left(\bar{l}_t + \frac{\mathrm{Var}(l_t)}{\bar{l}_t}\right). \tag{22}$$

The first term in the brackets is the component due block number variance, and the second term is the component due to block length variance (SI Section S6). The contribution of block length variance relative to block number variance is then

$$\frac{\mathrm{Var}(l_t)}{\left(\bar{l}_t\right)^2}, \tag{23}$$

which is simply the square of the coefficient of variation of block lengths. Fig. 4B plots this quantity over time in the case where all F1s are hybrids, and, together with the arguments above, reveals that block length variance is important for the purging of introgressed DNA in the early generations after hybridization. However, over time, it becomes less important, with block number variance eventually becoming the only source of ancestry variance across individuals.

This allows us to interpret the role of recombination in the purging of introgressed DNA as follows: In the first few generations after hybridization, recombination affects the rate at which introgressed DNA is purged because it chops the initial linkage blocks into smaller blocks of variable size. This implicates the aggregate recombination process—in particular, heterogeneity in chromosome size and the spatial distribution of crossovers—in the early purging of introgressed DNA [Eq. (17)]. In later generations, recombination affects the rate at which introgressed DNA is purged primarily because it affects block number variance, which is proportional to average block length [Eq. (22)], which in turn is inversely proportional to the total number of blocks. Therefore, the key effect of recombination in later generations is simply to chop blocks up into more blocks. Because a block with a crossover in it becomes two blocks no matter where that crossover is, this implicates the fine-scale recombination rate (cM/Mb) in later-generation purging.

In summary, the aggregate recombination process (as quantified by $\bar{r}$ and analogs) is most important in the early generations after hybridization (when most purging occurs), while the fine-scale recombination process is most important in the later generations. We can illustrate this interpretation by considering the impact of crossover distributions that differ only in the spatial location—and not the number—of crossovers. Here, crossover distributions associated with lower aggregate genetic shuffling (because of more terminal placement of crossovers) lead to faster purging of introgressed DNA in the early generations (when aggregate shuffling matters) but not in later generations (when the fine-scale recombination rate matters) (Fig. 5).

4 Discussion

Relatives vary in how much of their DNA they share identically by descent, because of variable patterns of recombination and segregation in their pedigrees. Here, we have calculated the variance in genetic relatedness as a function of parameters of the aggregate recombination process. In particular, we have found that the variances for different pedigree relationships are decreasing functions of members of a family of metrics of aggregate genetic shuffling.

For example, in the simple case of grandparent-grandoffspring, the variance of genetic relatedness is $(1/2-\bar{r})/8$ [Eq. (3)], where $\bar{r}$ is the average proportion of locus pairs that recombine in gametogenesis (Veller et al. 2019). Intuitively, when aggregate shuffling is higher in the parent's meiosis, the offspring's

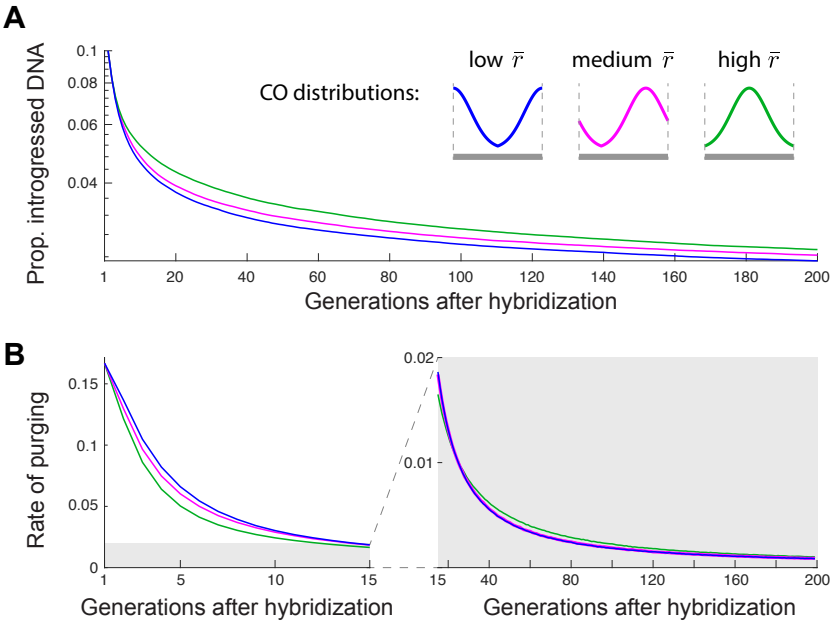

Figure 5: Purging of introgressed DNA for three spatial distributions of crossovers. The genome is a single chromosome, which experiences two crossovers per gamete on average. The three spatial distributions of crossovers have the same shape, but are shifted sideways from one other. In the green distribution, crossovers are concentrated centrally on the chromosome, leading to higher values of $\bar{r}$ and analogs. In the blue distribution, on the other hand, crossovers are terminally concentrated, leading to low values of $\bar{r}$ and analogs. The average fine-scale recombination rate is the same for each distribution, however, owing to an equal average number of crossovers per gamete. **A**. Purging of introgressed DNA is more rapid, and ultimately more complete, for the distributions associated with lower values of aggregate genetic shuffling. **B**. This is largely because of differences in the rates of purging in the generations shortly after hybridization, consistent with our interpretation that early purging is governed largely by aggregate genetic shuffling, while later purging is governed largely by the average fine-scale rate of recombination. Parameters are as for Fig. 3, except for the recombination process. Trajectories are averages across 10,000 replicate simulations.

inheritance of alleles is independent for a greater number of locus pairs, and so the offspring receives a more even allocation of grandmaternal and grandpaternal alleles (Thompson 2013).

4.1 Factors that influence variance in genetic relatedness

Recasting the variance in genetic relatedness in terms of aggregate genetic shuffling allows us to understand the impact of features of meiosis on the former in terms of their effect on the latter. A number of such features have recently been reported. Below, we show that their influence can be understood, perhaps more intuitively, by thinking in terms of aggregate genetic shuffling.

Sex differences in recombination. In many species, male and female meiosis differ both in the number and location of crossovers [reviewed by Lenormand and Dutheil (2005); Sardell and Kirkpatrick (2019)]. In male meiosis in humans, crossovers are fewer (by ∼50%) and more terminally localized than in female meiosis. Both factors decrease the total amount of genetic shuffling (Veller et al. 2019). This explains the observation of Caballero et al. (2019) that relatives who are related predominantly via males have a higher variance of genetic relatedness than relatives related predominantly via females.

Such effects will be especially pronounced in species where one sex has no crossing over in meiosis. For example, in *Drosophila*, there is crossing over in oogenesis but not in spermatogenesis. Substituting into Eq. (3) the values calculated in Section 3.2 for autosomal $\bar{r}$ in male and female *D. melanogaster*

($\bar{r}^{\♂} = 0.253$, $\bar{r}^{\♀} = 0.376$), we find that the variances of (autosomal) relatedness to a paternal and a maternal grandparent are 0.0308 and 0.0156, respectively.

Crossover interference. It has recently been shown, by computer simulation of various forms of crossover patterning along chromosomes, that crossover interference tends to decrease the variance of genetic relatedness between relatives (Caballero et al. 2019). Veller et al. (2019) demonstrated that interference among crossovers—by spreading them out more evenly along chromosomes—increases the amount of genetic shuffling that they cause (increasing $\bar{r}$ and analogs). This provides an intuitive explanation of the result of Caballero et al. (2019).

Recombination hotspots. White and Hill (2019) studied the effect of recombination hotspots on the variance of genetic relatedness between relatives, and concluded that the the effect of adding a recombination hotspot to a chromosome depends on the position of the hotspot. This can be understood in terms of the different effects of terminal and central crossovers noted above—i.e., it is an observation about the broad-scale distribution of crossovers. However, a separate question about hotspots concerns their effect on genetic relatedness if the broad-scale distribution of crossovers is held constant. Thinking in terms of aggregate genetic shuffling suggests that the effect of hotspots in this case will depend on the particular pedigree relationship. For example, in the presence of crossover interference, hotspots will have little effect on $\bar{r}$, since $\bar{r}$ is determined by the broad-scale distribution of crossovers in individual meioses; thus, hotspots will have little effect on genetic relatedness between grandparent and grandoffspring [Eq. (3)]. However, hotspots decrease the value of $\bar{r}_{(2)}$ because, in the pooling of crossovers from two independent meioses, hotspots will sometimes cause crossovers from the two meioses to land directly on each other, 'wasting' one of them. Thus, the existence of hotspots should increase the variance of genetic relatedness between siblings [Eq. (4)].

Crossover covariation. It has recently been shown across diverse eukaryotes that the number of crossovers per chromosome covaries positively across chromosomes within individual meiotic nuclei (Wang et al. 2019). This 'crossover covariation' increases the variance of the number of crossovers per gamete, which obviously will affect the distribution of genetic relatedness among relatives. However, crossover covariation does not change the probability that a particular pair of loci are recombinant in a given gamete, and therefore does not affect $\bar{r}$ or its analogs (since these are averages of functions of pairwise recombination rates—see SI Sections S1 and S2). Thus, crossover covariation does not affect the variance of genetic relatedness among relatives. However, it will affect higher-order moments of the distribution of genetic relatedness. For example, notice that the fourth-moment version of Eq. (2) involves terms of the form $\mathbb{E}[\hat{P}_a\hat{P}_b\hat{P}_c\hat{P}_d]$, which is the probability that the alleles at four distinct loci a, b, c, d in a gamete are inherited from the same grandparent. This requires that no pair of these loci are recombinant in the gamete. If a and b lie on one chromosome, and c and d on another, then crossover covariation makes this more likely.

4.2 Selection against introgressed DNA

The aggregate recombination process as a barrier to gene flow. Selection purges deleterious introgressed DNA at a rate proportional to the variance across the population in the amount of introgressed DNA carried [Eq. (10)]. Most of the purging of deleterious introgressed DNA happens in the first few generations after hybridization (Harris and Nielsen 2016; Schumer et al. 2018), when individuals with hybrid ancestry are direct descendants of recent ancestors from the donor species. Our results on ancestry variance in pedigree relationships of direct descent therefore reveal a role for the aggregate recombination process—as quantified by $\bar{r}$ and analogs—in modulating the amount of introgressed DNA purged in these critical early generations.

Thus, in simulations matched for all other parameters, a population with a *Drosophila*-like re-

combination process purges substantially more introgressed DNA shortly after hybridization than a population with a human-like recombination process (Fig. 3). This is because, with a small karyotype and no crossing over in males, *Drosophila* has a much lower aggregate rate of genetic shuffling than humans (*D. melanogaster* autosomal $\bar{r} = 0.314$; human autosomal $\bar{r} = 0.490$). Therefore, the aggregate recombination process of *Drosophila* acts as a more effective barrier to gene flow than the recombination process of humans.

Generally, aggregate genetic shuffling is dominated by independent assortment of chromosomes in meiosis (Crow 1988; Veller et al. 2019), so that species with more chromosomes will tend to have higher aggregate genetic shuffling. In addition, because each chromosome typically receives at least one crossover—but not many more—in meiosis, increases in chromosome number tend to be associated with increases in crossover number too (Stapley et al. 2017). Therefore, larger karyotypes are associated with less efficient purging of introgressed DNA—and are thus a weaker barrier to gene flow—both in the short run, owing to higher aggregate genetic shuffling, and in the long run, owing to more crossovers (Fig. S1).

Because it is the aggregate rate of genetic shuffling that modulates the initial rate at which introgressed DNA is purged, features of meiosis that alter the aggregate rate of genetic shuffling (as discussed in Section 4.1) also modulate how effective the recombination process is as a barrier to gene flow. For example, crossover interference increases $\bar{r}$ by spacing crossovers out evenly along chromosomes. These crossover dissect large introgressed blocks into more evenly sized smaller blocks, which reduces the average rate at which deleterious alleles in these blocks are purged [Eq. (22)].

In contrast to the early generations, the rate of purging in later generations is governed largely by the number of crossovers, i.e., the fine-scale recombination rate. Eventually, when recombination has dissociated all of the initial linkage relations between the deleterious introgressed alleles, selection acts upon them individually, and the rate of purging is equal to the average selective coefficient of the individual alleles [Eq. (21)].

The fate of neutral introgressed alleles. Neutral introgressed alleles can ultimately survive in the recipient population if they manage to recombine away from the deleterious introgressed alleles with which they are initially associated before those deleterious alleles are eliminated by selection. Bengtsson (1985) studied this process under the assumption that a neutral allele is on a separate chromosome to every deleterious allele, and defined the 'gene flow factor' (gff) as the probability that the neutral allele is maintained despite its initial association with deleterious alleles. Barton and Bengtsson (1986) calculated the gff in the case where the neutral allele lies on the same chromosome as the deleterious alleles.

Almost every neutral allele will initially lie between two adjacent deleterious alleles. As long as the neutral allele's initial linkage to either of these flanking deleterious alleles is maintained, its dynamics are as if it were a deleterious allele itself (Barton and Bengtsson 1986). When there are many loci at which introgressed alleles are deleterious [as we have assumed, and as supported empirically (Juric et al. 2016; Aeschbacher et al. 2017; Steinrücken et al. 2018)], each neutral allele is only a small recombination distance away from its flanking deleterious alleles, and therefore takes many generations to be freed of its association with them. So, for many generations after hybridization, the frequency dynamics of neutral introgressed alleles are very similar to those of deleterious introgressed alleles (Fig. 6). By implication, the factors that govern the rate of purging of deleterious introgressed DNA also govern the retention of neutral introgressed DNA (and thus the gff). In particular, the retention of neutral introgressed DNA is influenced by factors that affect the aggregate rate of genetic shuffling, such as chromosome number, heterogeneity in chromosome size, and the spatial distribution of crossovers (including crossover interference).

To get a numerical sense of for how many generations neutral introgressed alleles behave like deleterious introgressed alleles, consider a setup with 1,000 deleterious alleles spread across 20 chromosomes,

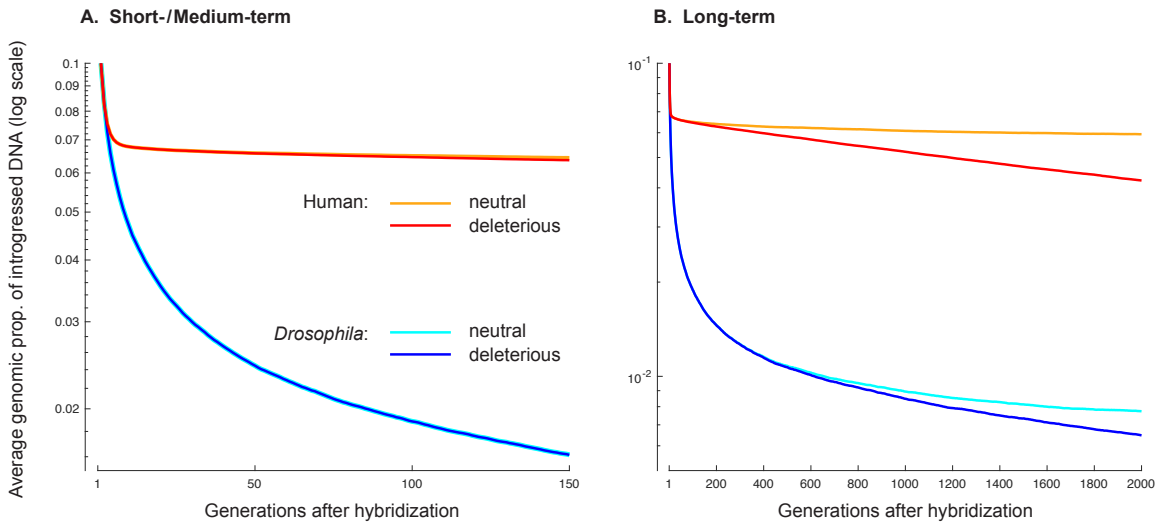

Figure 6: The purging of neutral vs. deleterious introgressed alleles under the recombination processes of humans and *D. melanogaster*. The model is the same as in Fig. 3, with 1,000 evenly spaced loci at which the introgressed alleles are deleterious, but now, midway between every adjacent pair of these loci, there is a locus at which the introgressed allele is neutral. **A**. For several hundred generations after hybridization, the average frequency of the neutral introgressed alleles closely follows that of the deleterious alleles, because most of the neutral alleles have not yet had sufficient time to recombine away from both of their nearest flanking deleterious alleles. In the figure, the lines for neutral introgressed ancestry have been widened slightly to make them more visible. **B**. After a sufficient number of generations, many of the neutral alleles have recombined away from the deleterious alleles they were initially in linkage with, and the average rate of purging of neutral introgressed ancestry slows relative to that of deleterious introgressed ancestry. Note that the placement of the neutral loci exactly midway between adjacent pairs of deleterious loci minimizes the average time required for the neutral alleles to recombine away from their flanking deleterious counterparts. If we were to include neutral loci closer to one flanking deleterious locus than the other, the trajectory of neutral introgressed ancestry would track that of deleterious introgressed ancestry even more closely than shown here. The trajectories displayed are averages from 10 replicate simulations.

with an average of one crossover per chromosome per gamete (resembling the case of Neanderthal-human introgression). Then adjacent deleterious loci recombine at rate $r \sim 1/50$, and so the number of generations required for a neutral introgressed allele situated midway between two deleterious alleles to recombine away from both of them is about $3/r \sim 150$ generations (and even longer for neutral alleles closer to one adjacent deleterious allele than to the other). If, instead, there are only 2 chromosomes (*Drosophila*), the analogous number of generations is ∼1,500 generations. Thus, the frequency dynamics of neutral introgressed alleles are expected to be similar to those of deleterious alleles for many generations after hybridization. Fig. 6 illustrates this in the case of the human and *D. melanogaster* recombination processes.

Introgressed DNA is purged more efficiently in regions of low recombination. A number of papers have recently reported a positive correlation, within the genomes of several species, between local recombination rate and the proportion of introgressed ancestry (Brandvain et al. 2014; Schumer et al. 2018; Martin et al. 2019; Edelman et al. 2019). This has been interpreted as due to the effect of (fine-scale) recombination in unlinking neutral (or beneficial) introgressed alleles from their deleterious counterparts.

However, our calculations reveal a further cause for the observed correlation between local recombination rate and introgressed ancestry: recombination mediates the rate at which the deleterious alleles themselves are purged. In regions with low recombination, deleterious introgressed alleles are

maintained in blocks of greater and more variable length, and are therefore purged more rapidly than in regions with high recombination [Eq. (22); Fig. S2]. Thus, in regions of low recombination, not only do linked neutral alleles suffer from the slower rate at which they recombine away from linked deleterious alleles, but they also have less time to do so before the linked deleterious alleles are eliminated. These factors reinforce each other in generating the observed correlation between recombination rate and retention of introgressed ancestry.

The above arguments implicate both aggregate and fine-scale recombination rates in the differences in retention of introgressed DNA observed across genomic regions. Thus, a chromosome with a terminal distribution of crossovers will tend to retain less introgressed DNA (including neutral introgressed DNA) than a chromosome elsewhere in the genome that has a more central distribution of crossovers, even if both chromosomes receive the same number of crossovers on average (Fig. S2A). On the other hand, a chromosome that receives fewer crossovers than another will tend to retain less introgressed ancestry (including neutral introgressed ancestry), even if the spatial distribution of crossover locations is the same for the two chromosomes (Fig. S2B).

These results, together with the arguments above about the number of generations it takes for neutral introgressed alleles to recombine away from the deleterious alleles with which they are initially associated, suggest that observed positive correlations between regional recombination rate and introgressed ancestry might be driven more by regional differences in the purging of deleterious alleles than by the unlinking of neutral alleles from deleterious alleles (e.g., Fig. S2B).

Introgression generates selection on the recombination process. We have found that the recombination process affects the rate at which deleterious introgressed DNA is purged following hybridization. This suggests, conversely, that introgression can exert an evolutionary pressure on the recombination process.

Introgression-mediated selection on modifiers of local recombination rates is straightforward. A modifier allele that reduces its local recombination rate prevents deleterious introgressed alleles from recombining onto its background, and is thus favored. For example, a segregating inversion keeps together a haplotype of non-introgressed alleles, and is therefore favored over the haplotype whose orientation is the same as that in the donor species and which therefore admits deleterious introgressed alleles by recombination (Kirkpatrick and Barton 2006).

Our results also point to how selection acts on global modifiers of the recombination process in the face of introgression. A modifier allele that reduces the aggregate recombination rate ($\bar{r}$ and analogs) increases the variance among its descendants in how much introgressed DNA they carry. This allows selection to purge introgressed DNA more efficiently among descendants of the modifier allele, causing the allele to end up in fitter genotypes and thus to be positively selected. This logic is similar, but the conclusion opposite, to the usual case where global modifiers that increase the recombination rate are favored by selection because they increase fitness variance among their descendants (Barton 1995; Burt 2000; Barton and Otto 2005). The conclusions are opposite because, in the usual case, the interaction of selection and random drift generates, on average, negative linkage disequilibria between deleterious alleles (Barton and Otto 2005), whereas in our case, the deleterious alleles are introgressed into the recipient population in perfect positive linkage disequilibrium.

Therefore, introgression generates selection on both local and global modifiers to reduce the recombination rate. Local modifiers of recombination include structural rearrangements (Kirkpatrick 2010), alterations to the binding sites of recombination-specifying proteins (Paigen and Petkov 2018; Grey et al. 2018), and mutations that affect local chromatin structure in meiotic prophase [e.g., Stack et al. (2017)]. On global modification, note that, even though reducing chromosome number is the most effective way to reduce aggregate recombination (Veller et al. 2019), introgression is not expected to select for reduced chromosome number, owing to fertility problems generally experienced by chromosome-number heterozygotes and hybrids (White 1978). Therefore, global modification of

the recombination process is restricted to modification of the number and spatial distribution of crossovers. Nevertheless, our expanding knowledge of the molecular biology of meiosis and recombination [reviewed in Hunter (2015); Zickler and Kleckner (2015)] suggests that global modifiers are probably very common. They include, for example, mutations to key meiosis proteins, such as those that determine the lengths of chromosome axes in meiotic prophase [e.g., Novak et al. (2008); Hong et al. (2019)], those that control the interference process along chromosome axes [e.g., Zhang et al. (2014)], and those that specify recombination hotspots (Paigen and Petkov 2018; Grey et al. 2018).

Limitations of our analysis. We have focused primarily on the qualitative impact of differences in the aggregate recombination process on the rate of purging of introgressed DNA. However, for reference, our simulations have been calibrated roughly to match parameters recently estimated for Neanderthal-human introgression (Harris and Nielsen 2016; Juric et al. 2016). Therefore, it is interesting to consider the limitations of our model in this context.

First, though, note that, while we have assumed the initial introgression fraction of 10% initially to be carried by F1 hybrids, Harris and Nielsen (2016) modelled the initial 10% fraction in the Neanderthal-human case as being present in fully Neanderthal individuals. In their model, strong selection against these Neanderthals in the first generation leads to a substantial reduction in Neanderthal ancestry by the time F1s are produced (from 10% to ∼6%), explaining why, in their simulations, the introgressed fraction is eventually substantially lower than in ours (compare our Fig. 3 with their Fig. 4; also see Fig. S3).

We have assumed that the loci at which introgressed alleles are deleterious are uniformly spaced throughout the genome. In reality, we expect these loci to be more common in functional regions such as genes (Sankararaman et al. 2014) and gene-regulatory elements (Telis et al. 2019), and depleted in repetitive and/or non-functional regions. Consistent with this latter point, Langley et al. (2019) have recently discovered large (multi-Mb) haplotypes segregating in humans that span centromeres and appear to be of archaic introgressed origin. Uneven spacing of the loci at which introgressed alleles are deleterious can be accommodated in our calculations by redefining $\bar{r}$ and analogs as averages across pairs of these loci (and blocks can be defined in terms of the number of loci they span). Therefore, our conclusions about the importance of the aggregate recombination process for the purging of introgressed DNA are unaffected.

For tractability, we have assumed that each introgressed allele has the same deleterious effect. In reality, effect sizes will vary across alleles. For example, if introgressed alleles are deleterious because of a higher genetic load in the donor species owing to its smaller effective population size (Juric et al. 2016; Harris and Nielsen 2016), then the effect size distribution for introgressed alleles will depend on the donor species' effective population size and distribution of fitness effects. In the early generations after hybridization (when most purging occurs), deleterious alleles are contained in large linkage blocks, selection against which depends on the sum of the effect sizes of the many alleles they contain. Therefore, in these early generations, variable effect sizes are not expected to substantially alter the rate of purging (Fig. S4). Later on, however, the rate of purging converges to the average allelic effect size, which will be smaller with variable effect sizes than with fixed effect sizes, because in the former case, the average allelic effect decreases as large-effect alleles are purged more rapidly. Therefore, holding the initial average allelic effect size constant, the eventual rate of purging will be smaller in the case of variable effect sizes (Fig. S4).

The selection scheme we have modelled, with additive fitness effects across loci, best resembles the case where introgressed alleles are simply deleterious in the recipient species (e.g., because of higher load in the donor species). This is a plausible type of selection against introgressed alleles (Juric et al. 2016; Harris and Nielsen 2016), with evidence favoring it in the case of Neanderthal-human introgression (Steinrücken et al. 2018). An alternative is that introgressed alleles are deleterious because of pairwise (or higher-order) epistatic interactions with alleles fixed or segregating in the recipient

species (Dobzhansky 1937; Muller 1942). Patterns of depletion of introgressed DNA in hybridizing swordtail fishes, for example, are best explained by selection against Dobzhansky-Muller incompatibilities (Schumer et al. 2018). Modelling selection against incompatibilities is slightly more complicated than the approach we have taken. For example, in the simplest model of pairwise incompatibilities, we might expect an individual's fitness reduction to be proportional to $P(1-P)$, where P is the proportion of introgressed DNA it carries. In our model, this individual's fitness is proportional to P. The effects of these selection schemes will thus differ in general, but will be approximately the same when introgressed ancestry is rare (because then $P(1-P) \approx P$).

Finally, in our calculations and simulations, we have assumed random mating among individuals with different degrees of introgressed ancestry. However, individuals could also mate non-randomly based on their degrees of introgressed ancestry, owing to mating preferences for conspecifics in both the recipient and the donor species. Assortative mating of this kind is often viewed as a pre-zygotic barrier to gene flow—individuals with high introgressed ancestry are disfavored by the majority of potential mates in the recipient population, which could result in their having fewer matings and thus reduced fitness. Thinking in terms of ancestry variance reveals another way in which assortative mating acts as a barrier to gene flow, even when individuals with high introgressed ancestry achieve the same number of matings as those with low introgressed ancestry. Assortative mating generates a positive correlation of introgressed ancestry across the two haploid genomes of individuals. This increases the variance across individuals in how much introgressed DNA they carry, and therefore increases the rate at which introgressed DNA is purged by selection. Thus, a preference for mating with conspecifics can act as both a pre- and post-zygotic barrier to gene flow between species.

Acknowledgements

We are grateful to Erin Calfee, Graham Coop, Jim Mallet, Molly Schumer, and Sivan Yair for helpful discussions.

References

Aeschbacher, S., Selby, J. P., Willis, J. H., and Coop, G. (2017). Population-genomic inference of the strength and timing of selection against gene flow. *Proceedings of the National Academy of Sciences*, 114(27):7061–7066.

Barton, N. H. (1983). Multilocus clines. *Evolution*, 37(3):454–471.

Barton, N. H. (1995). A general model for the evolution of recombination. *Genetics Research*, 65(2):123–144.

Barton, N. H. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Barton, N. H. and Otto, S. P. (2005). Evolution of recombination due to random drift. *Genetics*, 169(4):2353–2370.

Bengtsson, B. O. (1985). The flow of genes through a genetic barrier. In Greenwood, P. J., Harvey, P. H., and Slatkin, M., editors, *Evolution: Essays in honor of John Maynard Smith*, pages 31–42. Cambridge University Press, Cambridge.

Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G., and Sweigart, A. L. (2014). Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics*, 10(6):e1004410.

Burt, A. (2000). Perspective: Sex, recombination, and the efficacy of selection—was Weismann right? *Evolution*, 54(2):337–351.

Caballero, M., Seidman, D. N., Sannerud, J., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Carmi, S., and Williams, A. L. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *bioRxiv*. https://doi.org/10.1101/527655.

Comeron, J. M., Ratnappan, R., and Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, 8(10):e1002905.

Crow, J. F. (1988). The importance of recombination. In Michod, R. E. and Levin, B. R., editors, *The evolution of sex: An examination of current ideas*, pages 56–73. Sinauer, Sunderland.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University Press, New York.

Edelman, N. B., Frandsen, P., Miyagi, M., Clavijo, B. J., Davey, J., Dikow, R., Accinelli, G. G., Van Belleghem, S. M., Patterson, N. J., Neafsey, D. E., et al. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science*, 366:594–599.

Franklin, I. R. (1977). The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology*, 11(1):60–80.

Gorlov, I. P. and Gorlova, O. Y. (2001). Cost–benefit analysis of recombination and its application for understanding of chiasma interference. *Journal of Theoretical Biology*, 213(1):1–8.

Grey, C., Baudat, F., and de Massy, B. (2018). PRDM9, a driver of the genetic map. *PLoS Genetics*, 14(8):e1007479.

Guo, S.-W. (1996). Variation in genetic identity among relatives. *Human Heredity*, 46(2):61–70.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, 36(3):632–637.

Harris, K. and Nielsen, R. (2016). The genetic cost of Neanderthal introgression. *Genetics*, 203(2):881–891.

Hill, W. G. (1993a). Variation in genetic composition in backcrossing programs. *Journal of Heredity*, 84(3):212–213.

Hill, W. G. (1993b). Variation in genetic identity within kinships. *Heredity*, 71(6):652–653.

Hill, W. G. and Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(1):47–64.

Hong, S., Joo, J. H., Yun, H., Kleckner, N., and Kim, K. P. (2019). Recruitment of Rec8, Pds5 and Rad61/Wapl to meiotic homolog pairing, recombination, axis formation and S-phase. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz903.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., et al. (2015). The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Research*, 25(3):445–458.

Hunter, N. (2015). Meiotic recombination: the essence of heredity. *Cold Spring Harbor Perspectives in Biology*, 7:a016618.

Juric, I., Aeschbacher, S., and Coop, G. (2016). The strength of selection against Neanderthal introgression. *PLoS Genetics*, 12(11):e1006340.

Kardos, M., Luikart, G., and Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, 115(1):63.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, 8(9):e1000501.

Kirkpatrick, M. and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1):419–434.

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103.

Langley, S. A., Miga, K. H., Karpen, G. H., and Langley, C. H. (2019). Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *eLife*, 8:e42989.

Lenormand, T. and Dutheil, J. (2005). Recombination difference between sexes: a role for haploid selection. *PLoS Biology*, 3(3):e63.

Lian, J., Yin, Y., Oliver-Bonet, M., Liehr, T., Ko, E., Turek, P., Sun, F., and Martin, R. H. (2008). Variation in crossover interference levels on individual chromosomes from human males. *Human Molecular Genetics*, 17(17):2583–2594.

Martin, S. H., Davey, J. W., Salazar, C., and Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology*, 17(2):e2006288.

Martin, S. H. and Jiggins, C. D. (2017). Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, 47:69–74.

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. *Biological Symposia*, 6:71–125.

Novak, I., Wang, H., Revenkova, E., Jessberger, R., Scherthan, H., and Höög, C. (2008). Cohesin Smc1β determines meiotic chromatin axis loop organization. *Journal of Cell Biology*, 180(1):83–90.

Paigen, K. and Petkov, P. M. (2018). PRDM9 and its role in genetic recombination. *Trends in Genetics*, 34(4):291–300.

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences*, 116(5):1639–1644.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357.

Sardell, J. M. and Kirkpatrick, M. (2019). Sex differences in the recombination landscape. *American Naturalist*. https://doi.org/10.1086/704943.

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C., Sankararaman, S., Andolfatto, P., Rosenthal, G. G., et al. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, 360(6389):656–660.

Sherman, P. W. (1979). Insect chromosome numbers and eusociality. *The American Naturalist*, 113(6):925–935.

Stack, S. M., Shearer, L. A., Lohmiller, L., and Anderson, L. K. (2017). Meiotic crossing over in maize knob heterochromatin. *Genetics*, 205(3):1101–1112.

Stapley, J., Feulner, P. G., Johnston, S. E., Santure, A. W., and Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736):20160455.

Steinrücken, M., Spence, J. P., Kamm, J. A., Wieczorek, E., and Song, Y. S. (2018). Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology*, 27(19):3873–3888.

Telis, N., Aguilar, R., and Harris, K. (2019). Selection against archaic DNA in human regulatory regions. *bioRxiv*. https://doi.org/10.1101/708230.

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326.

Veller, C., Kleckner, N., and Nowak, M. A. (2019). A rigorous measure of genome-wide genetic shuffling that takes into account crossover positions and Mendel's second law. *Proceedings of the National Academy of Sciences*, 116(5):1659–1668.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W. G., Hottenga, J.-J., Willemsen, G., Boomsma, D. I., et al. (2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *The American Journal of Human Genetics*, 81(5):1104–1110.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2(3):e41.

Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, 107:4–13.

Wang, S., Veller, C., Sun, F., Ruiz-Herrera, A., Shang, Y., Liu, H., Zickler, D., Chen, Z., Kleckner, N., and Zhang, L. (2019). Per-nucleus crossover covariation and implications for evolution. *Cell*, 177:326–338.

White, I. M. S. and Hill, W. G. (2019). Effect of heterogeneity in recombination rate on variation in realised relationship. *Heredity*. https://doi.org/10.1038/s41437-019-0241-z.

White, M. A., Wang, S., Zhang, L., and Kleckner, N. (2017). Quantitative modeling and automated analysis of meiotic recombination. *Methods in Molecular Biology*, 1471. https://doi.org/10.1007/978-1-4939-6340-9_18.

White, M. J. D. (1978). *Modes of Speciation*. W. H. Freeman & Co., San Francisco.

Wilfert, L., Gadau, J., and Schmid-Hempel, P. (2007). Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity*, 98(4):189–197.

Zhang, L., Wang, S., Yin, S., Hong, S., Kim, K. P., and Kleckner, N. (2014). Topoisomerase II mediates meiotic crossover interference. *Nature*, 511(7511):551–556.

Zickler, D. and Kleckner, N. (2015). Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor Perspectives in Biology*, 7:a016626.

S1 General case for direct descent

Label the starting generation 0, so that the offspring generation is 1, the grand-offspring generation 2, etc. For an individual in generation 0, one of its descendants in generation t, and a locus k, let $P_k^{(t)}$ be a random variable that takes the value 1 if an allele carried by the generation-t descendant at locus k was inherited from the generation-0 individual, and takes the value 0 otherwise. Clearly,

$$\text{Prob}\left(P_k^{(t)} = 1\right) = \frac{1}{2^{t-1}}. \tag{S.1}$$

and because $P_k^{(t)}$ can take on only the values 0 or 1, $\mathbb{E}[P_k^{(t)}] = \text{Prob}\left(P_k^{(t)} = 1\right) = 1/2^{t-1}$. Then

$$\text{Var}\left(P_k^{(t)}\right) = \mathbb{E}\left[\left(P_k^{(t)}\right)^2\right] - \left(\mathbb{E}\left[P_k^{(t)}\right]\right)^2 = \text{Prob}\left(P_k^{(t)} = 1\right) - \left(\frac{1}{2^{t-1}}\right)^2 = \frac{1}{2^{t-1}} - \left(\frac{1}{2^{t-1}}\right)^2. \tag{S.2}$$

Case 1: No sex differences in recombination. Consider two loci, i and j. For the alleles at these loci in the generation-t descendant both to have been inherited from the generation-0 ancestor requires that the loci never have been recombinant in any of the gametes linking the generation-t descendant and the specified generation-0 ancestor (probability $1 - r_{ij}$ for each relevant gamete, starting with that produced by the generation-1 descendant and ending with that produced by the generation-$[t-1]$ descendant) and, conditional on this, that the appropriate alleles co-segregated to the gamete in each relevant meiosis (probability 1/2 each time). Therefore,

$$\text{Prob}\left(P_i^{(t)} = P_j^{(t)} = 1\right) = \frac{1}{2^{t-1}}(1 - r_{ij})^{t-1}, \tag{S.3}$$

so that

$$\begin{aligned}
\text{Cov}\left(P_i^{(t)}, P_j^{(t)}\right) &= \mathbb{E}\left[P_i^{(t)} P_j^{(t)}\right] - \mathbb{E}\left[P_i^{(t)}\right]\mathbb{E}\left[P_j^{(t)}\right] \\
&= \text{Prob}\left(P_i^{(t)} = P_j^{(t)} = 1\right) - \left(\frac{1}{2^{t-1}}\right)^2 \\
&= \frac{1}{2^{t-1}}(1 - r_{ij})^{t-1} - \left(\frac{1}{2^{t-1}}\right)^2 \\
&= \frac{1}{2^{t-1}}\left((1 - r_{ij})^{t-1} - \frac{1}{2^{t-1}}\right).
\end{aligned} \tag{S.4}$$

Finally, assume that there are L loci in total, with L very large, and let $P^{(t)}$ be the proportion of the the generation-t descendant's genome inherited from the generation-0 ancestor: $P^{(t)} = \frac{1}{L}\sum_{k=1}^{L} P_k^{(t)}$. Then

$$\mathbb{E}[P^{(t)}] = \frac{1}{L}\sum_{k=1}^{L}\mathbb{E}[P_k^{(t)}] = \frac{1}{L}\sum_{k=1}^{L}\frac{1}{2^{t-1}} = \frac{1}{2^{t-1}}, \tag{S.5}$$

and

$$\begin{aligned}
\mathrm{Var}\left(P^{(t)}\right) &= \mathrm{Var}\left(\frac{1}{L}\sum_{k=1}^{L} P_k^{(t)}\right) = \frac{1}{L^2}\sum_{k=1}^{L}\mathrm{Var}\left(P_k^{(t)}\right) + \frac{1}{L^2}\sum_{i\neq j}\mathrm{Cov}\left(P_i^{(t)}, P_j^{(t)}\right) \\
&= \frac{1}{L^2}L\left(\frac{1}{2^{t-1}} - \left(\frac{1}{2^{t-1}}\right)^2\right) + \frac{1}{L^2}\sum_{i\neq j}\frac{1}{2^{t-1}}\left((1-r_{ij})^{t-1} - \frac{1}{2^{t-1}}\right) \\
&= \frac{1}{L}\left(\frac{1}{2^{t-1}} - \left(\frac{1}{2^{t-1}}\right)^2\right) + \frac{L(L-1)}{L^2}\frac{1}{L(L-1)}\sum_{i\neq j}\frac{1}{2^{t-1}}\left((1-r_{ij})^{t-1} - \frac{1}{2^{t-1}}\right) \\
&\xrightarrow[L\to\infty]{} \frac{1}{2^{t-1}}\left(\overline{(1-r)^{t-1}} - \frac{1}{2^{t-1}}\right) \qquad \text{(S.6)} \\
&= \frac{1}{2^{t-1}}\left(1 - \frac{1}{2^{t-1}} + \sum_{\tau=1}^{t-1}(-1)^{\tau}\binom{t-1}{\tau}\overline{r^{\tau}}\right) \qquad \text{(S.7)}
\end{aligned}$$

where a bar represents the average taken with respect to all locus pairs, and $\binom{t-1}{\tau} = \frac{(t-1)!}{\tau!(t-1-\tau)!}$. The limit follows from the fact that $1/L \to 0$, $L(L-1)/L^2 \to 1$, and the number of pairs (i,j) such that $i \neq j$ is $L(L-1)$.

In the special case of the descendant being a grand-offspring ($t = 2$), Eq. (S.6) becomes

$$\mathrm{Var}(IBD_{\mathrm{grand}}) = \mathrm{Var}\left(P^{(2)}\right) \xrightarrow[L\to\infty]{} \frac{1}{2}\left(\overline{(1-r)} - \frac{1}{2}\right) = \frac{1}{2}\left(\frac{1}{2} - \bar{r}\right), \qquad \text{(S.8)}$$

which is Eq. (3) in the Main Text.

Case 2: Sex differences in recombination. Let $r_{ij}^{♀}$ and $r_{ij}^{♂}$ be the sex-specific recombination rates between loci i and j. If, among the $t-1$ individuals in the lineage between the generation-0 ancestor and the focal generation-t descendant, there are f females and $m = t-1-f$ males, then

$$\mathrm{Cov}\left(P_i^{(t)}, P_j^{(t)}\right) = \frac{1}{2^{t-1}}\left((1-r_{ij}^{♀})^f(1-r_{ij}^{♂})^m - \frac{1}{2^{t-1}}\right), \qquad \text{(S.9)}$$

so that, by a similar calculation to Eq. (S.6) above,

$$\mathrm{Var}\left(P^{(t)}\right) = \frac{1}{2^{t-1}}\left(\overline{(1-r^{♀})^f(1-r^{♂})^m} - \frac{1}{2^{t-1}}\right). \qquad \text{(S.10)}$$

If the number of females in the lineage is not known, it can be taken to be binomially distributed with parameter 1/2, in which case the average in Eq. (S.10) is calculated across all locus pairs and all possible numbers of females $f = 0, 1, \ldots, t-1$ (with associated probabilities $\binom{t-1}{f}/2^{t-1}$).

S2 General case for indirect relationships

Consider an individual (generation 0) and two of its descendants (generation t_1 and t_2) who have no more recent common ancestor than the generation-0 individual. The two generation-1 ancestors of the focal descendants (which could be the focal descendants themselves if t_1 and/or t_2 is 1) are half-sibs. Let $P_k^{(t_1,t_2)}$ be a random variable that takes on the value 1 if both focal descendants carry, at locus k, an allele inherited from their common generation 0 ancestor. Assuming Mendelian segregation,

$$\text{Prob}\left(P_k^{(t_1,t_2)} = 1\right) = \frac{1}{2^{t_1+t_2-1}}, \tag{S.11}$$

so that $\mathbb{E}\left[P_k^{(t_1,t_2)}\right] = \text{Prob}\left(P_k^{(t_1,t_2)} = 1\right) = 1/2^{t_1+t_2-1}$ and

$$\text{Var}\left(P_k^{(t_1,t_2)}\right) = \mathbb{E}\left[\left(P_k^{(t_1,t_2)}\right)^2\right] - \left(\mathbb{E}\left[P_k^{(t_1,t_2)}\right]\right)^2 = 1/2^{t_1+t_2-1} - \left(1/2^{t_1+t_2-1}\right)^2.$$

Now consider two loci, i and j. For the alleles at both loci in both descendants to have been inherited from their common ancestor in generation 0 (i.e., for the individuals to be IBD at these two loci) requires (i) that the two generation-1 ancestors be IBD at the two loci, which, because they are half-sibs, occurs with probability $[(1-r_{ij})^2 + r_{ij}^2]/2 = 1/2 - r_{ij}(1-r_{ij})$, (ii) that the two loci not be recombinant in any subsequent gamete leading to the focal generation-t_1 and generation-t_2 descendants, which occurs with probability $(1-r_{ij})^{t_1+t_2-2}$, and (iii) that, given (i) and (ii), the ancestor's alleles always segregate into the gametes leading to the focal descendants, which occurs with probability $1/2^{t_1+t_2-2}$. Therefore,

$$\text{Prob}\left(P_i^{(t_1,t_2)} = P_j^{(t_1,t_2)} = 1\right) = \left(\frac{1}{2} - r_{ij}(1-r_{ij})\right)(1-r_{ij})^{t_1+t_2-2}\frac{1}{2^{t_1+t_2-2}}, \tag{S.12}$$

so that

$$\begin{aligned}
\text{Cov}\left(P_i^{(t_1,t_2)}, P_j^{(t_1,t_2)}\right) &= \mathbb{E}\left[P_i^{(t_1,t_2)}P_j^{(t_1,t_2)}\right] - \mathbb{E}\left[P_i^{(t_1,t_2)}\right]\mathbb{E}\left[P_j^{(t_1,t_2)}\right] \\
&= \text{Prob}\left(P_i^{(t_1,t_2)} = P_j^{(t_1,t_2)} = 1\right) - \left(\frac{1}{2^{t_1+t_2-1}}\right)^2 \\
&= \left(\frac{1}{2} - r_{ij}(1-r_{ij})\right)(1-r_{ij})^{t_1+t_2-2}\frac{1}{2^{t_1+t_2-2}} - \left(\frac{1}{2^{t_1+t_2-1}}\right)^2 \\
&= \frac{1}{2^{t_1+t_2-1}}\left([1-2r_{ij}(1-r_{ij})](1-r_{ij})^{t_1+t_2-2} - \frac{1}{2^{t_1+t_2-1}}\right).
\end{aligned} \tag{S.13}$$

Now let $P^{(t_1,t_2)}$ be the fraction of the genome that both the focal descendants have inherited from their common generation-0 ancestor: $P^{(t_1,t_2)} = \frac{1}{L}\sum_{k=1}^{L} P_k^{(t_1,t_2)}$. Then

$$\mathbb{E}\left[P^{(t_1,t_2)}\right] = \frac{1}{L}\sum_{k=1}^{L}\mathbb{E}\left[P_k^{(t_1,t_2)}\right] = \frac{1}{L}\sum_{k=1}^{L}\frac{1}{2^{t_1+t_2-1}} = \frac{1}{2^{t_1+t_2-1}}, \tag{S.14}$$

while

$$
\begin{aligned}
\mathrm{Var}\left(P^{(t_1,t_2)}\right) &= \mathrm{Var}\left(\frac{1}{L}\sum_{k=1}^{L} P_k^{(t_1,t_2)}\right) \\
&= \frac{1}{L^2}\sum_{k=1}^{L}\mathrm{Var}\left(P_k^{(t_1,t_2)}\right) + \frac{1}{L^2}\sum_{i\neq j}\mathrm{Cov}\left(P_i^{(t_1,t_2)}, P_j^{(t_1,t_2)}\right) \\
&= \frac{1}{L^2}L\frac{1}{2^{t_1+t_2-1}}\left(1-\frac{1}{2^{t_1+t_2-1}}\right) \\
&\quad + \frac{1}{L^2}\sum_{i\neq j}\frac{1}{2^{t_1+t_2-1}}\left(\left[1-2r_{ij}(1-r_{ij})\right](1-r_{ij})^{t_1+t_2-2} - \frac{1}{2^{t_1+t_2-1}}\right) \\
&\xrightarrow[L\to\infty]{} \frac{1}{2^{t_1+t_2-1}}\left(\overline{\left[1-2r_{ij}(1-r_{ij})\right](1-r_{ij})^{t_1+t_2-2}} - \frac{1}{2^{t_1+t_2-1}}\right). \qquad \text{(S.15)}
\end{aligned}
$$

In the special case of the focal descendants being half-sibs ($t_1 = t_2 = 1$), Eq. (S.15) becomes

$$
\mathrm{Var}(IBD_{\text{h-sibs}}) = \mathrm{Var}\left(P^{(1,1)}\right) \xrightarrow[L\to\infty]{} \frac{1}{2}\left(\frac{1}{2} - 2\overline{r_{ij}(1-r_{ij})}\right) = \frac{1}{2}\left(\frac{1}{2} - \bar{r}_{(2)}\right), \qquad \text{(S.16)}
$$

which is Eq. (5) in the Main Text. Here, $2r_{ij}(1-r_{ij})$ is the probability that i and j are recombinant in exactly one of two gametes, and $\bar{r}_{(2)}$ is the average value of $\bar{r}$ calculated from the pooled crossovers of two independent meioses.

S3 Variance in ancestry among F2s

An F1 generation is created by hybridizing individuals from species A with individuals from species B. The F1s are then mated randomly with each other to produce an F2 generation. We assume that there is no selection, so that the distribution of genotypes among assayed F2s is the same as among F2 zygotes. Let the random variable $\hat{P}_k$ take the value 1 if the allele at locus i in an F1's gamete is from species A, and 0 if it is instead from species B. $\mathbb{E}[\hat{P}_k] = 1/2$ and $\mathrm{Var}(\hat{P}_k) = 1/4$.

For the alleles at loci i and j in an F1's gamete both to come from species A requires (a) that these loci not be recombinant in the gamete (probability $1 - r_{ij}$) and (b) that, assuming (a), the species-A alleles at these loci segregated to the gamete (probability 1/2). Therefore

$$\mathrm{Cov}(\hat{P}_i, \hat{P}_j) = \mathbb{E}[\hat{P}_i\hat{P}_j] - \mathbb{E}[\hat{P}_i]\mathbb{E}[\hat{P}_j] = \mathrm{Prob}(\hat{P}_i = \hat{P}_j = 1) - \frac{1}{4} = \frac{1}{2}\left(\frac{1}{2} - r_{ij}\right). \tag{S.17}$$

The genomic proportion of an F1's gamete that is derived from species A is $\hat{P} = \frac{1}{L}\sum_{k=1}^{L}\hat{P}_k$, where L is the number of loci and is assumed to be large. $\mathbb{E}[\hat{P}] = \frac{1}{L}\sum_{k=1}^{L}\mathbb{E}[\hat{P}_k] = 1/2$, and

$$\begin{aligned}
\mathrm{Var}(\hat{P}) = \mathrm{Var}\left(\frac{1}{L}\sum_{k=1}^{L}\hat{P}_k\right) &= \frac{1}{L^2}\left(\sum_{k=1}^{L}\mathrm{Var}(\hat{P}_k) + \sum_{i\neq j}\mathrm{Cov}(\hat{P}_i, \hat{P}_j)\right) \\
&= \frac{1}{L^2}\left(\frac{L}{4} + \sum_{i\neq j}\frac{1}{2}(1 - r_{ij})\right) = \frac{1}{4L} + \frac{1}{L^2}\sum_{i\neq j}\frac{1}{2}\left(\frac{1}{2} - r_{ij}\right) \\
&\xrightarrow[L\to\infty]{} \frac{1}{2}\left(\frac{1}{2} - \bar{r}\right).
\end{aligned} \tag{S.18}$$

If the F1 in question is female, $\bar{r}^{\,♀}$ is used to derive $\mathrm{Var}(\hat{P}^{♀})$; if male, $\bar{r}^{\,♂}$ is used to derive $\mathrm{Var}(\hat{P}^{♂})$.

Let P be the proportion of an F2's genome that is derived from species A. $P = \frac{1}{2}\hat{P}^{♀} + \frac{1}{2}\hat{P}^{♂}$, and so

$$\mathrm{Var}(P) = \frac{1}{4}\mathrm{Var}(\hat{P}^{♀}) + \frac{1}{4}\mathrm{Var}(\hat{P}^{♂}) = \frac{1}{8}(1 - \bar{r}^{\,♀} - \bar{r}^{\,♂}), \tag{S.19}$$

which is Eq. (6) in the Main Text.

S4 Relationship between variance in heterozygosity and change in heterozygosity under hybrid vigor

Let the random variable H be the proportion of loci that are heterozygous in a zygote. Suppose that selection acts according to a scheme of hybrid vigor, where an individual with proportion h of loci heterozygous has relative viability $1 + hS$, where viability here refers to the probability of surviving from zygote stage to the stage at which genotyping of F2s occurs. The random variable H' is the proportion of loci that are heterozygous at the stage of genotyping. Let the probability density functions for H and H' be $f(h)$ and $g(h)$ respectively. Then

$$g(h) = \frac{f(h)(1+hS)}{\int_0^1 f(\eta)(1+\eta S)d\eta} = \frac{f(h)(1+hS)}{\int_0^1 f(\eta)d\eta + S\int_0^1 \eta f(\eta)d\eta} = \frac{f(p)(1+hS)}{1+S\mathbb{E}[H]}. \tag{S.20}$$

From this,

$$\begin{aligned}
\mathbb{E}[H'] - \mathbb{E}[H] &= \int_0^1 \eta g(\eta)d\eta - \mathbb{E}[H] \\
&= \int_0^1 \eta \frac{f(\eta)(1-\eta S)}{1+S\mathbb{E}[H]}d\eta - \mathbb{E}[H] = \frac{1}{1+S\mathbb{E}[H]}\left[\int_0^1 \eta f(\eta)d\eta + S\int_0^1 \eta^2 f(\eta)d\eta\right] - \mathbb{E}[H] \\
&= \mathbb{E}[H] - \frac{\mathbb{E}[H]+S\mathbb{E}[H^2]}{1+S\mathbb{E}[H]} = \frac{\mathbb{E}[H]+S\left(\mathbb{E}[H]\right)^2 - \mathbb{E}[H]+S\mathbb{E}[H^2]}{1+S\mathbb{E}[H]} \\
&= \frac{S\left[\mathbb{E}[H^2]-\left(\mathbb{E}[H]\right)^2\right]}{1+S\mathbb{E}[H]} = \frac{S\mathrm{Var}(H)}{1+S\mathbb{E}[H]},
\end{aligned} \tag{S.21}$$

which is Eq. (7) in the Main Text.

S5 Relationship between variance in the proportion of introgressed DNA and the rate at which it is purged

The calculation is similar to that in Section S4. Let the random variable P_t be the fraction of introgressed DNA carried by a member of the generation-t population, and let the random variable $\hat{P}_t$ be the fraction of introgressed DNA in a successful gamete from generation t (i.e., after selection has acted). Suppose that the probability density function for P_t is $f(p)$. Then the probability density function for $\hat{P}_t$ is

$$g(p) = \frac{f(p)(1-pS)}{\int_0^1 f(\rho)(1-\rho S)d\rho} = \frac{f(p)(1-pS)}{\int_0^1 f(\rho)d\rho - S\int_0^1 \rho f(\rho)d\rho} = \frac{f(p)(1-pS)}{1-S\mathbb{E}[P_t]}. \tag{S.22}$$

From this,

$$\begin{aligned}
\mathbb{E}[P_t] - \mathbb{E}[P_{t+1}] &= \mathbb{E}[P_t] - \mathbb{E}[\hat{P}_t] = \mathbb{E}[P_t] - \int_0^1 \rho g(\rho)d\rho \\
&= \mathbb{E}[P_t] - \int_0^1 \rho \frac{f(\rho)(1-\rho S)}{1-S\mathbb{E}[P_t]}d\rho = \mathbb{E}[P_t] - \frac{1}{1-S\mathbb{E}[P_t]}\left[\int_0^1 \rho f(\rho)d\rho - S\int_0^1 \rho^2 f(\rho)d\rho\right] \\
&= \mathbb{E}[P_t] - \frac{\mathbb{E}[P_t] - S\mathbb{E}[P_t^2]}{1-S\mathbb{E}[P_t]} = \frac{\mathbb{E}[P_t] - S\left(\mathbb{E}[P_t]\right)^2 - \mathbb{E}[P_t] + S\mathbb{E}[P_t^2]}{1-S\mathbb{E}[P_t]} \\
&= \frac{S\left[\mathbb{E}[P_t^2] - \left(\mathbb{E}[P_t]\right)^2\right]}{1-S\mathbb{E}[P_t]} = \frac{S\mathrm{Var}(P_t)}{1-S\mathbb{E}[P_t]},
\end{aligned} \tag{S.23}$$

which is Eq. (10) in the Main Text.

S6 The rate of purging of introgressed DNA as a function of the distribution of block lengths

In a population of size N, the proportion of introgressed ancestry is x, which we assume to be small. The introgressed DNA is in n 'blocks' of average length $\bar{l}$, measured as a fraction of total diploid genome length, so that $n = Nx/\bar{l}$. These blocks are defined as sets of introgressed alleles with identical inheritance pedigrees going back to the hybridization pulse, and, when sufficiently numerous, can be assumed to be distributed randomly among the population. Under this assumption, the probability that an individual gets a particular block is $1/N$, and so the number of blocks an individual gets, B, is binomially distributed with parameters n and $p = 1/N$. So $\mathbb{E}[B] = np = n/N = x/\bar{l}$ and $\mathrm{Var}(B) = np(1-p) = \frac{n}{N}\left(1 - \frac{1}{N}\right) \approx x/\bar{l}$ when N is large.

Let the random variable P be the fraction of an individual's genome that is introgressed. Then $P = \sum_{b=1}^{B} l_b$, where l_b is the length of the b-th block assigned to the individual. $\mathbb{E}[P|B] = \sum_{b=1}^{B} \mathbb{E}[l_b] = B\bar{l}$ and, because the l_b are independent, $\mathrm{Var}(P|B) = \sum_{b=1}^{B} \mathrm{Var}(l_b) = B\mathrm{Var}(l)$. From the law of total variance,

$$\begin{aligned}
\mathrm{Var}(P) &= \mathrm{Var}_B(\mathbb{E}[P|B]) + \mathbb{E}_B[\mathrm{Var}(P|B)] \\
&= \mathrm{Var}(B\bar{l}) + \mathbb{E}[B\mathrm{Var}(l)] \\
&= \left(\bar{l}\right)^2 \mathrm{Var}(B) + \mathrm{Var}(l)\mathbb{E}[B] \\
&\approx x\left(\bar{l} + \frac{\mathrm{Var}(l)}{\bar{l}}\right).
\end{aligned} \tag{S.24}$$

The first term in Eq. (S.24) is the contribution to variance in P from variance in the number of blocks carried by different individuals, while the second term is the contribution from variance in the lengths of different blocks.

From Eq. (S.23), the proportion of introgressed ancestry that is purged from one generation to the next is

$$\frac{x - x'}{x} = \frac{S\mathrm{Var}(P)}{x(1 - Sx)},$$

where S is the fitness reduction of individuals with 100% introgressed ancestry. Substituting in the variance calculation above,

$$\frac{x - x'}{x} \approx \frac{S\left(\bar{l} + \frac{\mathrm{Var}(l)}{\bar{l}}\right)}{1 - Sx} \approx S\left(\bar{l} + \frac{\mathrm{Var}(l)}{\bar{l}}\right). \tag{S.25}$$

So the rate of purging depends only on the average and variance of the block lengths. Eq. (S.25) can be written in simpler form:

$$\frac{x - x'}{x} \approx S\left(\bar{l} + \frac{\mathrm{Var}(l)}{\bar{l}}\right) = S\frac{\left(\bar{l}\right)^2 + \mathrm{Var}(l)}{\bar{l}} = S\overline{l^2}/\bar{l}, \tag{S.26}$$

where $\overline{l^2} = \mathbb{E}\left[l^2\right]$ is the uncentered second moment of the distribution of block lengths.

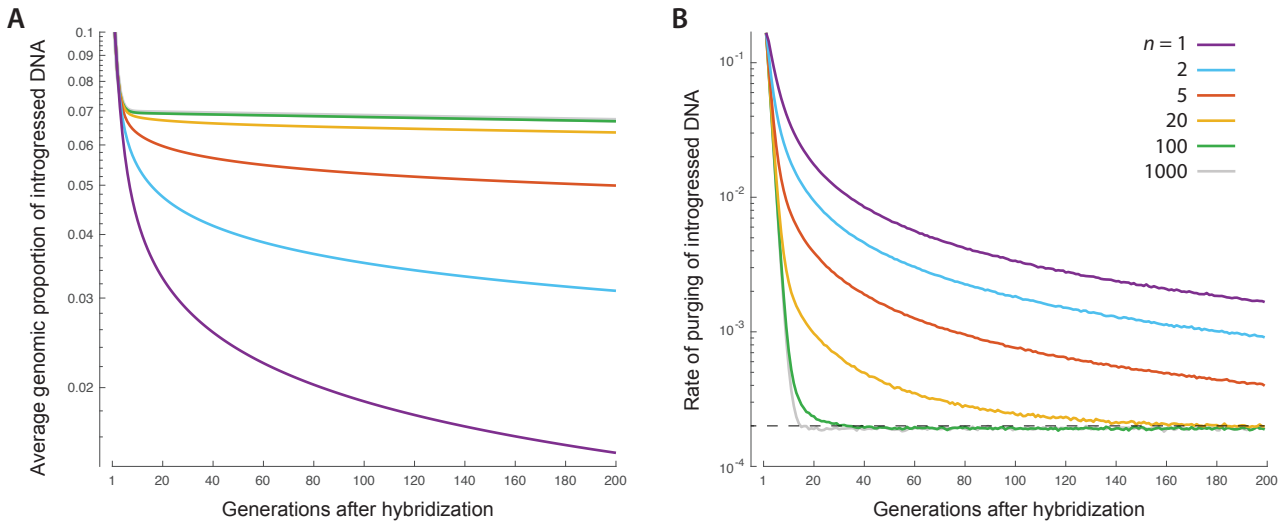

Figure S1: The effect of chromosome number on the purging of introgressed DNA. The model simulated here involves n chromosomes of equal size, with 1,000 loci allotted equally among the chromosomes and spaced evenly along them. There is, on average, one crossover per chromosome per gamete, with crossover positions uniformly distributed along the chromosome. When n is larger, the purging of introgressed DNA is substantially slowed, most obviously in the short run (owing to higher aggregate recombination—$\bar{r}$ and analogs—caused largely by independent assortment of a greater number chromosomes) but also thereafter (owing to a higher average fine-scale recombination rate, caused by a greater number of crossovers). Note, however, that the rate of purging in each case will eventually converge to the average allelic effect, $s = 2 \times 10^{-4}$ (dotted line in **B**).

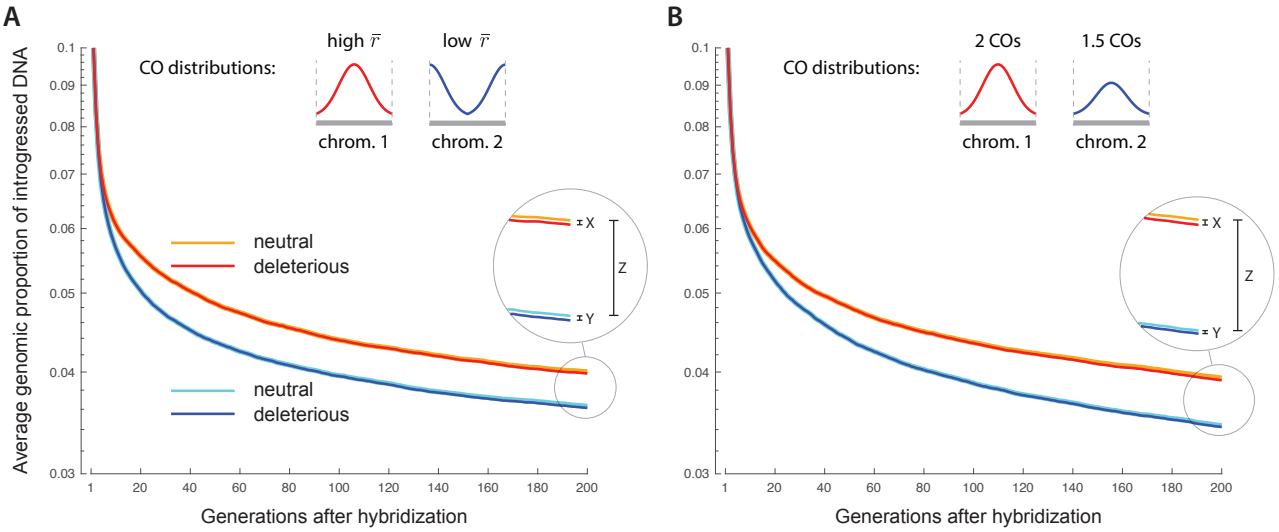

Figure S2: Genomic regions with low aggregate recombination purge more introgressed DNA. **A**. Chromosome 1 and chromosome 2 experience the same average number of crossovers per meiosis (2), but these crossovers tend to be more terminally localized on chromosome 2, causing chromosome 2 to have a lower aggregate rate of recombination ($\bar{r}$ and analogs) than chromosome 1. Because of this, the rate of purging of introgressed DNA is initially higher on chromosome 2 than on chromosome 1. Because chromosome 1 and chromosome 2 have equal average fine-scale recombination rates, their rates of purging become similar fairly quickly after the hyrbidization pulse. Nonetheless, the rate differences in the crucial early generations result in chromosome 2 ultimately carrying substantially less introgressed ancestry than chromosome 1. **B**. Chromosomes 1 and 2 have the same spatial distribution of crossovers, but chromosome 2 experiences 25% fewer crossovers than chromosome 1. This causes both the aggregate and fine-scale recombination rates of chromosome 2 to be lower, and so the rate of purging of introgressed DNA is higher for chromosome 2, both in the early generations after hybridization, and later on. In both **A** and **B**, the frequency trajectories of neutral introgressed alleles interspersed between deleterious alleles closely resemble the trajectories of the deleterious alleles, for reasons explained in the Main Text. Therefore, the effect of recombination on differences in neutral introgressed ancestry across the chromosomes is driven almost entirely by the effect of recombination on the purging of deleterious introgressed alleles (quantity Z), rather than recombination's effect in unlinking neutral introgressed alleles from their deleterious flanking alleles (quantity $X - Y$). This is true even in **B**, where chromosome 1 has a higher average fine-scale recombination rate than chromosome 2, causing neutral alleles on chromosome 1 to recombine away from linked deleterious alleles at an unambiguously faster rate than on chromosome 2 (resulting in $X > Y$). Because the effect on deleterious introgressed alleles depends largely on purging in the early generations, this implicates the aggregate recombination process in the purging of introgressed ancestry—deleterious and neutral.

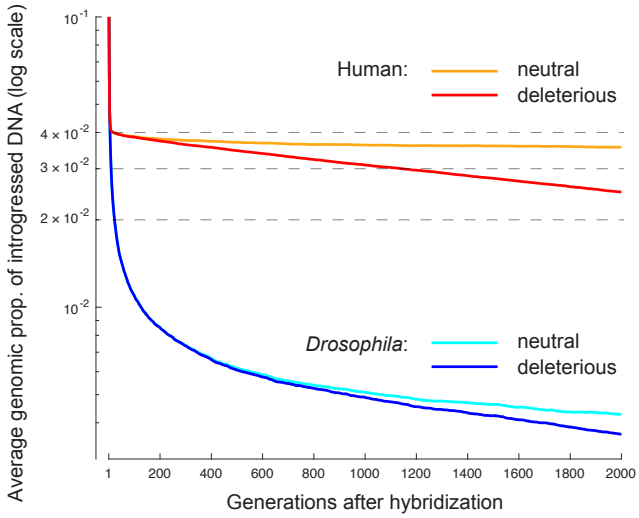

Figure S3: The purging of deleterious and neutral introgressed DNA in our model, assuming the initial setup of Harris and Nielsen (2016), where 10% of the population carries only donor-species DNA and 90% of the population carries only recipient-species DNA. This is in contrast to the setup we have used elsewhere in this paper, where 20% of the initial population are F1 hybrids between the donor and recipient species, and 80% carry only recipient-species DNA. The initial fraction of introgressed ancestry is 10% in both cases. The key difference for the purging of introgressed DNA is that, under the setup of Harris and Nielsen (2016), a large fraction of introgressed DNA (∼40%) is purged in the first generation—because of selection against the fully donor-species individuals—before F1s are formed. With this initial setup, under the human recombination process, we recover frequency trajectories of introgressed ancestry in our model that resemble the frequency trajectory in Fig. 4 of Harris and Nielsen (2016). These trajectories result in an eventual level of introgressed ancestry that resembles the level of Neanderthal ancestry in modern non-African humans. The elimination of introgressed DNA is still much more profound under the recombination process of *Drosophila melanogaster* with the initial setup of Harris and Nielsen (2016).

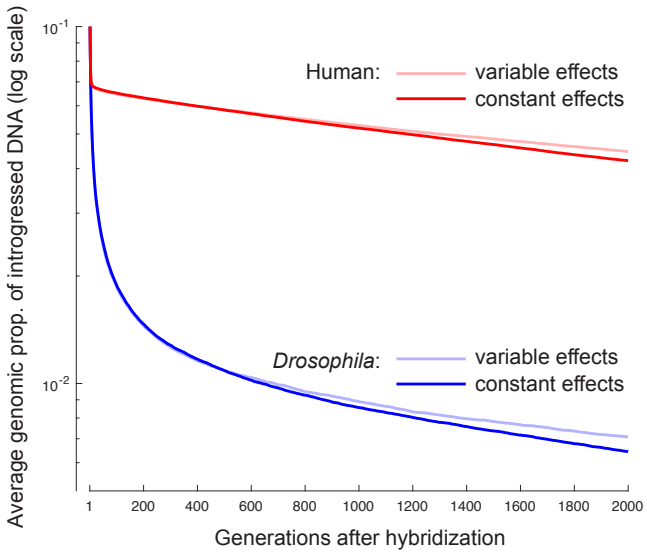

Figure S4: Purging of deleterious introgressed DNA when the deleterious alleles have constant effect sizes (bold lines) vs. variable effect sizes (faded lines). In the variable case, the effect size of the introgressed allele at each locus is drawn independently from an exponential distribution whose mean is equal to the effect size in the constant effect size case (2×10^{-4}). The rate of purging is similar between the two cases when block lengths are still large, because the mean allelic effect size is then most important. Later, the rate becomes slower in the variable effect size case, because large-effect alleles have been preferentially purged so that the average effect size has declined below its initial value (which always remains the average value in the constant effect size case). Nonetheless, the impact of allowing variable effect sizes is small.