

University of Groningen

Haplotype block dynamics in hybrid populations

Janzen, Thijs; Nolte, Arne W.; Traulsen, Arne

DOI:
[10.1101/058107](https://doi.org/10.1101/058107)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Early version, also known as pre-print

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Janzen, T., Nolte, A. W., & Traulsen, A. (2016). *Haplotype block dynamics in hybrid populations*. (BioRxiv).
BioRxiv. <https://doi.org/10.1101/058107>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Haplotype block dynamics in hybrid populations

Thijs Janzen^{*1}, Arne W. Nolte^{*,†} and Arne Traulsen^{*}

^{*}Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Plön, Germany, [†]Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9–11, 26111, Oldenburg, Germany

ABSTRACT When species originate through hybridization, the genomes of the ancestral species are blended together. Over time genomic blocks that originate from either one of the ancestral species accumulate in the hybrid genome through genetic recombination. Modeling the accumulation of ancestry blocks can elucidate processes and patterns of genomic admixture. However, previous models have ignored ancestry block dynamics for chromosomes that consist of a discrete, finite number of chromosomal elements. Here we present an analytical treatment of the dynamics of the mean number of blocks over time, for continuous and discrete chromosomes, in finite and infinite populations. We describe the mean number of haplotype blocks as a universal function dependent on population size, the number of genomic elements per chromosome, the number of recombination events, and the initial relative frequency of the ancestral species.

KEYWORDS hybridization, haplotype blocks, recombination, junctions

2 **S**peciation through hybridization has long been recognized as
3 a potential driver in the formation of new species in plants
4 (Grant 1981). More recently, it has also received attention as a
5 process that may lead to speciation in animals (Abbott *et al.* 2013).
6 It has been pointed out that genetic admixture between differ-
7 entiated lineages should only be considered hybrid speciation
8 when the joint contribution of both parental species is instrumen-
9 tal in the rise of the new species, for example by creating direct
10 barriers to reproduction with the parental species or by facilitat-
11 ing ecological isolation of the emerging hybrid lineage (Mallet
12 2007; Nolte and Tautz 2010; Abbott *et al.* 2013). Hybrid speciation
13 involves that parental genetic variance that reduces the fitness of
14 the emerging lineage is purged or selected for if it helps to adapt
15 to a new niche (Buerkle *et al.* 2000; Barton 2001). Although these
16 studies predicted a lag phase during which a hybrid lineage has
17 to go through an evolutionary optimization, empirical studies
18 suggest that hybrid speciation can occur rapidly, possibly within
19 hundreds of generations (Nolte *et al.* 2005; Buerkle and Rieseberg
20 2008; Abbott *et al.* 2013). Hence, systems and methods to gain
21 better insight in the timeframes required for hybrid speciation
22 are needed. Conventional molecular clock estimates are usually
23 too coarse to be applied to cases of rapid speciation, but lineages
24 of hybrid origin hold the potential to estimate rather short time-

25 frames from the ancestry structure of admixed genomes (Buerkle
26 and Rieseberg 2008; Liang and Nielsen 2014). Newly formed
27 hybrids contain contiguous genomic blocks that originate from
28 either one of the ancestral species and decay from generation to
29 generation through genetic recombination. Understanding how
30 ancestral genomic blocks decay over time can inform us about
31 how genomes of hybrid lineages evolved.

32 Fisher already recognized that the mix of genetic material
33 after a hybridization event is organized within contiguous hap-
34 lotype blocks. The dynamics of the delineation between these
35 blocks, ‘junctions’, can be traced through time, and he formu-
36 lated the expected number of junctions given the number of
37 generations passed since the onset of hybridization (Fisher 1949,
38 1954). Fisher developed the theory of junctions for full sibmating,
39 and the theory of junctions was quickly extended towards self-
40 fertilization (Bennett 1953), alternate parent-offspring mating
41 (Fisher 1959; Gale 1964), random mating (Stam 1980) and recom-
42 binant inbred lines (Martin and Hospital 2011). In order to derive
43 expected numbers of junctions and variation in the number of
44 junctions, Fisher had to assume that the size of the genetic blocks
45 delineated by these junctions was exponentially distributed. Us-
46 ing simulations, Chapman and Thompson (2003) showed that
47 this assumption was inaccurate, and that large blocks tended
48 to be overrepresented compared to an exponential block size
49 distribution. Furthermore, Chapman and Thompson extended
50 the theory of junctions towards populations growing in size at a
51 constant rate, and towards subdividing populations (Chapman
52 and Thompson 2002, 2003).

doi: 10.1534/genetics.XXX.XXXXX

Manuscript compiled: Thursday 9th June, 2016%

¹Thijs Janzen, Department of Evolutionary Theory, Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Plön, Germany, E-mail: janzen@evolbio.mpg.de

53 Within the theory of junctions, the chromosome is assumed
54 to be continuous, and to be infinitely divisible. Given that a
55 chromosome consists of an array of base pairs, such an assump-
56 tion provides an accurate approximation only if the number of
57 base pairs is extremely large. However, ancestry data is usually
58 not acquired on the level of base pairs, but rather using a lim-
59 ited number of markers (microsatellites, single nucleotide poly-
60 morphisms (SNPs), single feature polymorphisms (SFPs)) per
61 chromosome. Unfortunately, the number of markers required
62 to detect all haplotype blocks needs to be high; in order to de-
63 tect 90% of all apparent haplotype blocks at least 10 times more
64 markers than blocks are required (MacLeod *et al.* 2005). One
65 way to circumvent such high marker numbers is by comparing
66 simulated haplotype block densities with observed haplotype
67 block densities and apply a post-hoc correction to the molecular
68 data (Buerkle and Rieseberg 2008). Both using extremely high
69 marker densities and performing post-hoc corrections are ad hoc
70 solutions, and both these approaches lack a solid theoretical un-
71 derpinning connecting current standing theory for the number
72 of blocks in continuous chromosomes, with theory describing
73 the number of blocks in chromosomes described by discrete
74 numbers of markers.

75 Furthermore, the theory of junctions has focused severely
76 on the idealized situation in which the genome of all F1 hybrid
77 individuals contains equal proportions of the genetic material
78 of the parental species. In nature, the overall ancestry contri-
79 bution of parental species to the founding hybrid swarm may
80 differ (Edmands *et al.* 2005; Nolte and Tautz 2010; Stemshorn
81 *et al.* 2011). Deviations from an even ratio can cause the genetic
82 material in the F1 hybrids to sway in favor of one of the ancestral
83 species. Strong deviations from equality can increase the impact
84 of drift, as fixation of genetic material in a population becomes
85 more likely, especially for small population sizes. The impact
86 of deviations from an even ancestry contribution of both ances-
87 tral species and the interaction with population size remains
88 understudied so far.

89 Here we present a universal haplotype block theory describ-
90 ing the mean number of haplotype blocks in the population. Our
91 universal haplotype block theory includes both continuous and
92 discrete chromosomes, it takes into account the ancestry distri-
93 bution of the parental species in the founding hybrid swarm,
94 and it includes drift due to finite population size. We compare
95 our theory with previously obtained results by MacLeod *et al.*
96 (2005) and Buerkle and Rieseberg (2008), and confirm validity of
97 our theory using individual based models.

98 Our paper is structured as follows: first we derive the mean
99 number of haplotype blocks in a continuous chromosome, for
100 infinite and finite population size. Then, we proceed to derive
101 dynamics of the mean number of haplotype blocks for discrete
102 chromosomes consisting of a finite number of recombination
103 sites. We then infer universal haplotype block dynamics, by
104 combining properties of our previous derivations for the mean
105 number of blocks in discrete and continuous chromosomes. Us-
106 ing individual based simulations, we then demonstrate the val-
107 idity of our derivations, and extend our derivations towards
108 discrete chromosomes in finite populations, and towards multi-
109 ple recombination events per meiosis.

110 1. Analytical model

111 A. The expected number of haplotype blocks in a continuous 112 chromosome

113 First, we derive the expected number of blocks depending on
114 the time since the onset of hybridization. We assume infinite
115 population size, random mating, and an continuous chromo-
116 some, e.g. there are an infinite number of recombination sites
117 along the chromosome. We assume that only a single crossover
118 event occurs per chromosome per meiosis, which corresponds
119 to the assumption that chromosomes are 1 Morgan long. The re-
120 combination rate is assumed to be uniformly distributed across
121 the chromosome. Both chromosomes are interchangeable, and
122 we do not keep track of the identity of chromosomes. Each
123 individual is diploid and chromosomes are inherited independ-
124 ently, which allows us to track haplotype blocks within only
125 one chromosome pair, rather than all pairs simultaneously.

126 We start by formulating a recurrence equation based on the
127 expected change, or the change in mean number of blocks per
128 generation. Given a recombination site picked randomly across
129 the length of the chromosome, the genomic material on either
130 chromosome can either be identical or different. If the genomic
131 material is identical, no change in the number of blocks occurs.
132 If the genomic material is different, a new block is formed (see
133 Figure 1). The probability of observing the same type of genomic
134 material on both chromosomes is proportional to the frequency
135 of genomic material of that type in the population, which in turn
136 is dependent on the frequency of the corresponding ancestral
137 species in the ancestral population. We denote the frequency
138 of type P genomic material from ancestral species 1 as p , and
139 the frequency of genomic material of the other type Q (from the
140 other ancestral species) as q , where $p = 1 - q$. The probability of
141 having the same type of genomic material on both chromosomes
142 at the recombination site is then $p^2 + q^2$, in which case no change
143 in the number of blocks is observed. With probability $2pq$ the
144 type of genomic material on both chromosomes differs and an
145 increase in number of blocks is observed. We obtain

$$146 \bar{n}_{t+1} = \bar{n}_t + 2pq. \quad (1)$$

147 Here \bar{n}_t is the average number of blocks at time t . The solution
of Equation (1) is given by

$$148 \bar{n}_t = n_0 + 2pqt. \quad (2)$$

149 The number of blocks increases linear in time. The probability of
150 having a different type of genomic material, $2pq$, is the heterozy-
151 gosity H , and we can write Equation (2) in terms of heterozy-
152 gosity, $\bar{n}_t = n_0 + Ht$. Taking into account that the number of
153 junctions J at time t is $J_t = n_t - 1$ and assuming a non-constant
154 heterozygosity H , we recover the previously obtained result
155 (Chapman and Thompson 2002; MacLeod *et al.* 2005; Buerkle
and Rieseberg 2008)

$$156 J_t = \sum_{i=0}^{t-1} H_i, \quad (3)$$

157 where J_t is the number of junctions J at time t . Because the
158 population size is infinite, in our case the average heterozygosity
159 does not change from H_0 , and $J_t = H_0t$.

160 If the population size is not infinite, but finite with size N_j
161 at time j , and we allow for selfing, the average heterozygosity
changes over time (Crow and Kimura 1970) as

$$162 H_i = H_0 \prod_{j=0}^{i-1} \left(1 - \frac{1}{2N_j} \right). \quad (4)$$

162 Assuming constant population size, $N_j = N$ for all j , the
 163 expected number of junctions is given by

$$J_t = \sum_{i=0}^{t-1} H_i = \sum_{i=0}^{t-1} H_0 \left(1 - \frac{1}{2N}\right)^{i-1}. \quad (5)$$

164 The expected number of blocks, given an initial proportion p of
 165 species 1 is given by:

$$\bar{n}_t = 1 + 2pq \sum_{j=0}^{t-1} \left(1 - \frac{1}{2N}\right)^{j-1}. \quad (6)$$

166 In the limit of $N \rightarrow \infty$ we recover Equation (2). For $t \rightarrow \infty$, we
 167 have (MacLeod *et al.* 2005)

$$\bar{n}_\infty = 1 + 4pqN. \quad (7)$$

168 Thus, for finite N , the number of blocks converges to a finite
 169 number determined only by the population size and the initial
 170 frequencies of genomic material, which in turn depends on the
 171 frequency of the ancestral species during the first admixture
 172 event.

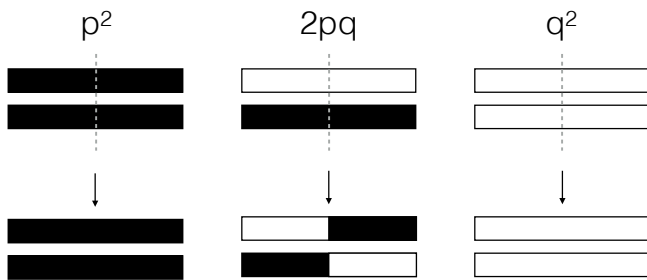


Figure 1 Change in number of haplotype blocks depending on the genomic match between blocks. Top row depicts the two parental chromosomes, bottom row depicts the two chromosomes produced after recombination has taken place at the grey dotted line during meiosis. Genome types are indicated using either black color (type P), or white color (type Q). With probability $p^2 + q^2$ no change in the number of blocks is observed. With probability $2pq$ we observe an increase in the number of blocks, where $p = 1 - q$ is the fraction of genomic material of type 1.

173 **B. A finite number of recombination spots: a discrete chromosome**
 174

175 In the previous section we have assumed that recombination
 176 never occurs twice at the same spot. In reality, a chromosome
 177 can not be indefinitely divided into smaller parts. We therefore
 178 proceed to study the change in number of blocks in a chromo-
 179 some consisting of L different chromosomal segments, where
 180 each segment represents a minimal genomic element that can not
 181 be broken down further, for instance a single nucleotide, a gene,
 182 a specific codon, or a genomic area delineated by two genetic
 183 markers. Considering a chromosome of L genomic segments,
 184 there are $L - 1$ possible crossover spots (junctions). Given that
 185 there are n blocks on the chromosome, there are $n - 1$ points on
 186 the chromosome where one block ends, and a new block begins
 187 (Fisher called these external junctions), excluding the tips of the
 188 chromosome. The probability that a recombination event takes

189 place at exactly such a point, given that there are n_t blocks at
 190 time t is:

$$\alpha = \frac{n_t - 1}{L - 1} \quad (8)$$

191 Conditioning on the type of the first chromosome, and only
 192 looking at the first of the two produced chromosomes (all other
 193 produced chromosomes are identical or the exact mirror image
 194 of this chromosome) we can distinguish four possible events,
 195 taking into account the location of the recombination spot on
 196 both chromosomes (Figure 2):

- 197 A) recombination takes place on an existing junction on both
 198 chromosomes (probability α^2)
- 199 B) recombination takes place on an existing junction on one
 200 chromosome, and within a block on the other chromosome
 201 (probability $\alpha(1 - \alpha)$)
- 202 C) recombination takes place on within a block on one chromo-
 203 some, and on an existing junction on the other chromosome
 204 (probability $(1 - \alpha)\alpha$)
- 205 D) recombination takes place within a block on both chromo-
 206 somes (probability $(1 - \alpha)^2$)

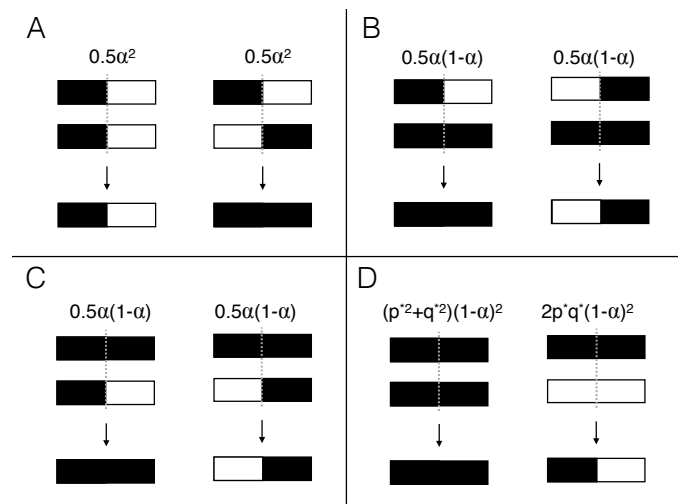


Figure 2 Change in number of haplotype blocks depending on the genomic match between blocks. Full chromosomes are shown here, but the same rationale applies to subsets of a chromosome. Top rows within each panel indicate the two parental chromosomes, bottom row indicates one of two possible resulting chromosomes after meiosis, where recombination takes place at the dotted grey line. Genomic material of type 1 is indicated in black, genomic material of type 2 is indicated in white. With probability α^2 recombination takes place on an existing junction on both chromosomes **A**, with probability $\alpha(1 - \alpha)$ recombination takes place on an existing junction on one chromosome, and within a block on the other chromosome **B**, with probability $(1 - \alpha)\alpha$ recombination takes place on within a block on one chromosome, and on an existing junction on the other chromosome **C**, with probability $(1 - \alpha)^2$ recombination takes place within a block on both chromosomes **D**.

A) When a crossover event takes place on an existing junction on both chromosomes, there is either no change in the number

of blocks (when the two junctions are identical), or a decrease in the number of blocks (when the two junctions are of opposing type). The probability of either event happening is $\frac{1}{2}$, yielding an average change in the number of blocks when crossover takes place on an existing junction on both chromosomes of $-\frac{1}{2}$.

B) When a crossover event takes place on an existing junction on one chromosome, and within a block on the other chromosome, there are two possibilities: either the block on the other chromosome is of the same type as the genomic material before the existing junction, or it is of the other type. If it is of the same type, the existing junction disappears, and the number of blocks decreases by one. If it is of the other type, the existing junction remains and the number of blocks does not change. The probability of either event happening is $\frac{1}{2}$, and hence we expect the total number of blocks on average to change by $-\frac{1}{2}$.

C) When a crossover event takes place within a block on the first chromosome, and on an existing junction on the other chromosome (the inverse of the previous situation), the outcome is exactly the opposite. If the genetic material after the junction on the second chromosome is of the same type as the block on the first chromosome, no new junction is formed and the number of blocks stays the same. If the genetic material after the junction on the second chromosome is of a different type than that of the block on the first chromosome, a new junction is formed and the number of blocks increases by one. The probability of either event happening is $\frac{1}{2}$, and hence we expect the total number of blocks on average to change by $\frac{1}{2}$.

D) When recombination takes place within a block on both chromosomes, matters proceed as described for the continuous chromosome: with probability $p^2 + q^2$ we observe no change in the number of blocks, and with probability $2pq$ we observe an increase. But, since we are dealing with a finite number of junction positions along the chromosome, the frequency of junction spots of a genomic type is no longer directly related to p . If there would be no blocks, i.e. if the genomic material would be distributed in an uncorrelated way, we know that $p(L-1)$ junction spots are of type P , that is, they are within a block of type P . Similarly, $q(L-1)$ junction spots are within a block of type Q . As new blocks are formed, the number of junction spots that are still within a block decreases. With the formation of a new block, on average both a junction within a block of type P and a junction within a block of type Q are lost, such that on average, after the formation of a new junction, the number of junctions of type P decreases by $\frac{1}{2}(n-1)$. Thus the number of junctions within a block of type P is $p(L-1) - \frac{1}{2}(n-1)$. Similarly, the number of junctions within a block of type Q is $q(L-1) - \frac{1}{2}(n-1)$. The probability then of selecting an internal junction of type P is the number of internal junctions of type P divided by the total number of junctions. Let us denote the probability of selecting an internal junction of type P by p^* , which is then given by

$$p_t^* = \frac{p(L-1) - \frac{1}{2}(\bar{n}_t - 1)}{p(L-1) - \frac{1}{2}(\bar{n}_t - 1) + q(L-1) - \frac{1}{2}(\bar{n}_t - 1)} = \frac{p(L-1) - \frac{1}{2}(\bar{n}_t - 1)}{L - \bar{n}_t}. \quad (9)$$

And the probability q_t^* is

$$q_t^* = 1 - p_t^* = \frac{q(L-1) - \frac{1}{2}(\bar{n}_t - 1)}{L - \bar{n}_t}. \quad (10)$$

With probability $2p_t^*q_t^*$ we observe an increase in number of blocks. Combining the scenarios (A)-(D) we can formulate the total expected change in number of blocks

$$\begin{aligned} \bar{n}_{t+1} &= \bar{n}_t + 2p_t^*q_t^*(1-\alpha)^2 + \frac{1}{2}\alpha(1-\alpha) - \frac{1}{2}(\alpha(1-\alpha)) - \frac{1}{2}\alpha^2 \\ &= \bar{n}_t + 2p_t^*q_t^*(1-\alpha)^2 - \frac{1}{2}\alpha^2. \end{aligned} \quad (11)$$

In terms of p , q , and L , Equation (11) can be written as

$$\bar{n}_{t+1} = 2pq + \frac{1}{L-1} + \frac{L-2}{L-1}\bar{n}_t \quad (12)$$

The solution of the recursion Equation (12) is given by

$$\bar{n}_t = \bar{n}_0 \left(\frac{L-2}{L-1} \right)^t + (1+2pq(L-1)) \left(1 - \left(\frac{L-2}{L-1} \right)^t \right). \quad (13)$$

The exponential decay terms ensures that we have convergence at $t \rightarrow \infty$, where we obtain

$$\bar{n}_\infty = 1 + 2pq(L-1). \quad (14)$$

A Taylor expansion at $t = 0$ shows that initially, the number of blocks increases linearly

$$\bar{n} \approx \bar{n}_0 + (\bar{n}_0 - \bar{n}_\infty) \ln \left(\frac{L-2}{L-1} \right) t. \quad (15)$$

C. Multiple recombination events

So far, we have assumed that during meiosis only a single crossover event occurs. Although this might often apply, multiple crossover events occur frequently. First, we consider the case of two crossover events. Assuming that the position of the two crossovers is independent, that there is no interference between the two crossovers, and that the two crossovers do not take place at the same position, we can extend our recurrence equations as follows.

For an infinite population, with a discrete chromosome, the first position is still chosen as in Equation (11). To obtain the probability for the second position of selecting a junction that lies within two dissimilar blocks, we have to correct p_t^* and q_t^* (whereas previously there were $L-1$ spots, there are now $L-2$), and we obtain

$$\begin{aligned} \bar{n}_{t+1} &= \bar{n}_t + \left(2p_{1(t)}^*q_{1(t)}^*(1-\alpha_1)^2 - \frac{1}{2}\alpha_1^2 \right) + \\ &\quad \left(2p_{2(t)}^*q_{2(t)}^*(1-\alpha_2)^2 - \frac{1}{2}\alpha_2^2 \right). \end{aligned} \quad (16)$$

Where:

$$p_{1(t)}^* = \frac{p(L-1) - \frac{1}{2}(\bar{n}_t - 1)}{L - \bar{n}_t} \quad (17)$$

$$q_{1(t)}^* = \frac{q(L-1) - \frac{1}{2}(\bar{n}_t - 1)}{L - \bar{n}_t} \quad (18)$$

$$p_{2(t)}^* = \frac{p(L-2) - \frac{1}{2}(\bar{n}_t - 1)}{L - \bar{n}_t - 1} \quad (19)$$

$$q_{2(t)}^* = \frac{q(L-2) - \frac{1}{2}(\bar{n}_t - 1)}{L - \bar{n}_t - 1} \quad (20)$$

$$\alpha_1 = \frac{(\bar{n}_t - 1)}{L - 1} \quad (21)$$

$$\alpha_2 = \frac{(\bar{n}_t - 1)}{L - 2}. \quad (22)$$

Equation (16) has the solution:

$$\bar{n}_t = \bar{n}_0 \left(\frac{L^2 - 5L + 5}{(L-2)(L-1)} \right)^t + \left(1 + 4pq \frac{(L-2)(L-1)}{2L-3} \right) \left(1 - \left(\frac{L^2 - 5L + 5}{(L-2)(L-1)} \right)^t \right) \quad (23)$$

Again we have convergence at $t \rightarrow \infty$, where we obtain:

$$\bar{n}_\infty = 1 + 4pq \frac{(L-1)(L-2)}{2L-3} \approx 1 + 2pq(L-1), \quad (24)$$

where the approximation holds for large L . The difference in the maximum number of blocks between one and two recombination events is

$$\begin{aligned} \bar{n}_{\infty 2} - \bar{n}_{\infty 1} &= 1 + 4pq \frac{(L-1)(L-2)}{2L-3} - (1 + 2pq(L-1)) \\ &= -2pq \frac{L-1}{2L-3} \approx -pq, \end{aligned} \quad (25)$$

where the approximation again holds for large L . An increase in the number of recombination events thus decreases the maximum number of blocks.

Extending Equation (16) towards M recombination events is similar,

$$\bar{n}_{t+1} = \bar{n}_t + \sum_{i=1}^M \left(2p_i^* q_i^* (1 - \alpha_i)^2 - \frac{1}{2} \alpha_i^2 \right) \quad (26)$$

with

$$p_i = \frac{p(L-i) - \frac{1}{2}(\bar{n}_t - 1)}{p(L-i) - \frac{1}{2}(\bar{n}_t - 1) + q(L-i) - \frac{1}{2}(\bar{n}_t - 1)} \quad (27)$$

$$q_i = \frac{q(L-i) - \frac{1}{2}(\bar{n}_t - 1)}{p(L-i) - \frac{1}{2}(\bar{n}_t - 1) + q(L-i) - \frac{1}{2}(\bar{n}_t - 1)} \quad (28)$$

$$\alpha_i = \frac{(\bar{n}_t - 1)}{L-i}. \quad (29)$$

For not too small L , and not too large M , taking into account Equation (25), we can approximate \bar{n}_∞ by:

$$\bar{n}_\infty = 1 + 2pq(L-1) - 2pq(M-1) \frac{L-1}{2L-3} \quad (30)$$

$$\approx 1 + 2pq(L-1) - pq(M-1) \quad (31)$$

Using numerical iteration of Equation (26) for $M = [2, 3, 4, 5]$ and $L = [2M + 1, 2M + 2, \dots, 200]$, and comparing the maximum number of blocks with the approximation of Equation (31) shows that Equation (31) is a good approximation (error $< 0.1\%$) for $L \geq 10M - 3$.

2. Universal haplotype dynamics

The general pattern of the accumulation of blocks over time is highly similar across different scenarios: after an initial period of a strong increase in the number of blocks, the number of blocks starts to increase more slowly, and approaches a maximum. Generally, we can infer that the maximum number of blocks is dependent on p , N and L , as do the time dynamics required to reach this maximum. Hence, we can describe the number of blocks relative to the number of blocks at $t = \infty$, which we define here as K . During meiosis, the average number of blocks

in the most ideal case can increase with γ blocks. There might be limitations, dependent on p , N , L or the number of already existing blocks. As such we make the ansatz:

$$\frac{d\bar{n}}{dt} = \gamma - \lambda \bar{n} \quad (32)$$

where γ is the maximum growth rate and λ encompasses all factors limiting the formation of new blocks. This includes, but is not limited to, factors induced by a finite population size N by a finite chromosome size L . Defining $\tau = \lambda t$, we can rescale time in Equation (32) such that

$$\frac{d\bar{n}}{d\tau} = \frac{\gamma}{\lambda} - \bar{n} \quad (33)$$

Measuring the number of blocks in terms of their equilibrium values, $\bar{n} = \lambda \frac{\bar{n}}{\gamma}$, we obtain

$$\frac{d\bar{n}}{d\tau} = 1 - \bar{n} \quad (34)$$

The solution of Equation (34) is of the form:

$$\bar{n}(\tau) = 1 - e^{-\tau} \quad (35)$$

Given that at $\tau = 0$, a chromosome, by definition, consists of one block, the number of blocks at $\tau = 0$ has to be equal to $1/K$: $\bar{n}(0) = 1/K$. Thus, in order to accurately describe haplotype block dynamics, we expect the dynamics to be of the form

$$\bar{n}(t) = (1 - e^{-\tau}) + \frac{1}{K} e^{-\tau} \quad (36)$$

We can find the scaling of time by solving

$$\bar{n}(1) - \bar{n}(0) = 1 - e^{-\beta t} + \frac{1}{K} e^{-\beta t} - \frac{1}{K} \quad (37)$$

By definition $\bar{n}(0)$ is always $\frac{1}{K}$, which leads to

$$\bar{n}(1) = 1 - e^{-\beta} + \frac{1}{K} e^{-\beta} \quad (38)$$

This allows us to calculate β from K and $\bar{n}(1)$,

$$\beta = \ln \left(\frac{K-1}{K-\bar{n}(1)} \right). \quad (39)$$

Returning to our original notation, the haplotype block dynamics is given by:

$$\bar{n}(t) = K \left(1 - \left(\frac{K-1}{K-\bar{n}(1)} \right)^{-t} \right) + \left(\frac{K-1}{K-\bar{n}(1)} \right)^{-t} \quad (40)$$

For $t \rightarrow \infty$, this converges to K , and for $t = 0$ is equal to 1 . Equation (40) provides us with a general scalable equation where all haplotype block dynamics are described in terms of K and $\bar{n}(1)$.

A. Implementing universal dynamics

To implement the universal equation (40), we only require to know K and $\bar{n}(1)$. For an infinite population, with a continuous chromosome, we have previously derived that the average number of blocks at time t is given by: $\bar{n}_t = 1 + 2pqt$ (equation 2), and thus $\bar{n}_{t=1} = 1 + 2pq$. The maximum number of blocks K is equal to ∞ .

For a finite population with a continuous chromosome, we have shown that $K = 1 + 4pqN$ and that the number of blocks at

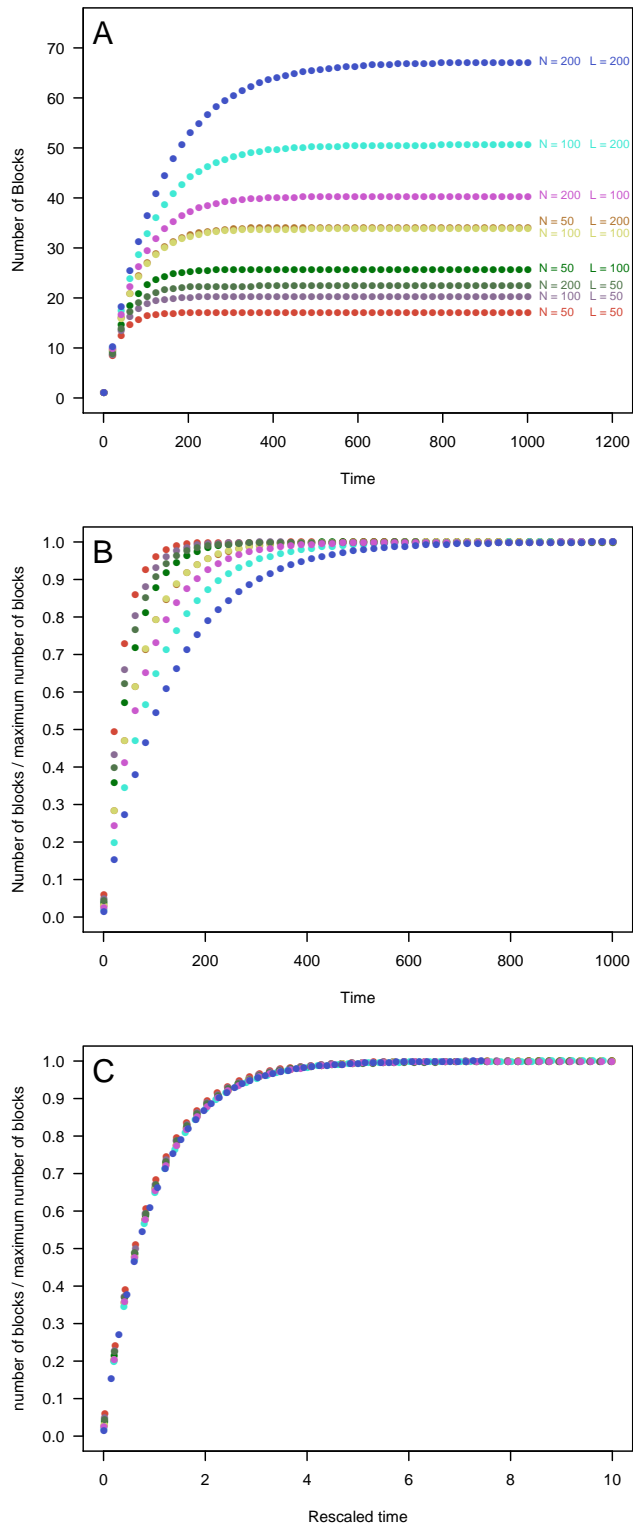


Figure 3 Graphical example of the construction of universal haplotype block dynamics using results from individual based simulations. **A:** mean number of haplotype blocks for $p = 0.5$, $N = [50, 100, 200]$ and $L = [50, 100, 200]$, number of replicates = 10,000. **B:** The mean number of blocks for the same parameter combinations, after rescaling the number of blocks relative to the maximum number of blocks K . **C:** The rescaled number of blocks vs rescaled time, by rescaling time according to β in Equation (39). After rescaling both the number of blocks according to K , and time according to β , all curves for different values of N and L reduce to a single, universal, curve, which follows Equation (40).

time t is given by $\bar{n}_t = 1 + 2pq \sum_{j=0}^{t-1} (1 - \frac{1}{2N})^{t-1}$ (Equation (6)), and hence $\bar{n}_{t=1} = 1 + 2pq$.

For an infinite population with a discrete chromosome, we have shown that $K = 2pq(L - 1) + 1$ and that the number of blocks at time t is given by $\bar{n}_t = \bar{n}_0 \left(\frac{L-2}{L-1}\right)^t + (2pq(L - 1) + 1) \left(1 - \left(\frac{L-2}{L-1}\right)^t\right)$ (Equation (13)), and hence:

$$\begin{aligned} \bar{n}_{t=1} &= \frac{L-2}{L-1} + (1 + 2pq(L-1)) \left(1 - \frac{L-2}{L-1}\right) \\ &= \frac{L-2}{L-1} + 1 + 2pq(L-1) - (1 + 2pq(L-1)) \frac{L-2}{L-1} \\ &= \frac{L-2}{L-1} + 1 + 2pq(L-1) - 2pq(L-2) - \frac{L-2}{L-1} \\ &= 1 + 2pq(L-1) - 2pq(L-2) \\ &= 1 + 2pq. \end{aligned} \quad (41)$$

We find that regardless whether the chromosome is continuous or discrete, and regardless of whether the population is finite or infinite, $\bar{n}(1) = 1 + 2pq$, which makes intuitive sense: in the first generation, none of the factors that limit recombination as a result of finite population size, or finite chromosome size come into play. When the population is finite, the formation of new blocks is limited by recombination taking place at a recombination spot where in a previous generation recombination has already taken place. In the first generation, all chromosomes are non-recombined, and finite population effects have no effect yet. When the chromosome is discrete, the formation of new blocks is limited by recombination taking place on a site that has previously recombined. In the first generation, no previous recombination events have happened yet, and recombination is thus not limited (yet).

3. Individual based simulations

To verify our analytical framework, and extend the framework towards discrete chromosomes in finite populations, we test our findings using an individual based model. We model the hybrid population as a Wright-Fisher process, extended with recombination:

- Non-overlapping generations
- Constant population size N
- Random mating
- Diploid
- Uniform recombination rate across the genome
- M recombination events per meiosis

Each individual has 2 chromosomes of length L , which are a sequence of 0 and 1's, where 1 represents an allele from an ancestral parent of type P and a 0 represents an allele from an ancestral parent of type Q . The model works as follows. In the first time step, N individuals are generated, where each individual can have either two parents of type P (with probability p^2), two parents of type Q (with probability $(1-p)^2$) or one parent of type P and one parent of type Q (with probability $2p(1-p)$).

In every consecutive time step, N new individuals are produced, where each individual is the product of a reproduction event between two individuals (including selfing) from the previous generation. Parental individuals are drawn with replacement, such that one individual could reproduce multiple times, but will on average reproduce one time. We assume that in a mating event both parents produce a large number of haploid

356 gametes from which two gametes (one from each parent) are
 357 chosen to form the new offspring. During production of the
 358 gametes, M recombination sites are chosen. The location of the
 359 recombination sites follows a uniform distribution between 0
 360 and L .

361 A. Continuous Chromosome

362 We model each chromosome as a continuous line, and only keep
 363 track of junctions delineating the end of a block. For each junction,
 364 we record the position along the chromosome (a number
 365 between 0 and 1) and whether the transition is $0 \rightarrow 1$ or $1 \rightarrow 0$.
 366 Over time, the number of blocks reaches a maximum value, but
 367 only if the population is finite (Figure 4). In the first few gener-
 368 ations the accumulation of blocks follows that of an infinite
 369 population (dotted line in Figure 4), but rapidly simulation re-
 370 sults start deviating from the infinite population dynamics. The
 371 maximum number of blocks is roughly obtained within $10N$
 372 generations. Furthermore, when the amount of genetic materi-
 373 al from either of the ancestral species is strongly skewed (e.g.
 374 $p = 0.9$), the maximum number of blocks is lower, and is reached
 375 within a shorter timespan.

376 B. Discrete Chromosome

377 To mimic a chromosome consisting of a discrete number of ge-
 378 nomic elements, we model the chromosome as a bitstring, where
 379 a 0 indicates a chromosomal segment of type 0, and a 1 indicates
 380 a chromosomal segment of type 1. To approximate an infinite
 381 population we use a population of size 100,000. Mean number of
 382 blocks in the stochastic simulations closely follow our analytical
 383 estimates (Figure 5), for all chromosome lengths considered here.
 384 The maximum number of blocks is reached in roughly $10L$ gener-
 385 ations. Again we observe that for strongly skewed ancestral
 386 proportions ($p = 0.9$), the maximum number of blocks is lower,
 387 and the maximum is reached within a shorter timeframe.

388 C. Discrete Chromosome in a Finite Population

389 We have shown in section 2 that we can describe haplotype
 390 dynamics for a population of any N and a chromosome of
 391 any L , as long as either N or L , or both are infinite. To de-
 392 scribe haplotype dynamics for any given N and L , we only
 393 need to know K and $\bar{n}(1)$. Since $\bar{n}(1) = 1 + 2pq$ for any N
 394 and L , this leaves us with disentangling the relationship be-
 395 tween N , L and K . We know that for a given L , with vary-
 396 ing N , K will approach $2pq(L - 1) + 1$ when $N \rightarrow \infty$ (Figure
 397 6A). The exact functional relationship however, remains un-
 398 known. Similarly for given N , with $L \rightarrow \infty$, K will approach
 399 $4pqN + 1$ (Figure 6 B). But again, the exact functional rela-
 400 tionship remains unknown. We therefore turn to individual based
 401 simulations to generate values of K for different combinations
 402 of L and N . Then, using non-linear least-squares fitting we for-
 403 mulate an approximate functional relationship between N , L ,
 404 and K . Nevertheless, such a formulation will remain a crude
 405 approximation only and should be interpreted as such. In order
 406 to do so, we simulate 10,000 populations for all combinations
 407 of N and L for values $[10, 20, 30, \dots, 200, 250, 300, \dots, 500]$ and
 408 $p = [0.5, 0.6, 0.7, 0.8, 0.9]$. We then use the mean of the maximum
 409 number of haplotype blocks across these 10,000 replicates to
 410 reconstruct our functional relationship. Given a value of L , we
 411 guess the relationship between N and K to be approximately of
 412 the form: $K_N = (2pq(L - 1) + 1)N / (aN + b)$, where a and b are
 413 to be determined and are possibly dependent on L and p . Note
 414 that we could have also chosen another form, such as $(2pq(L -$

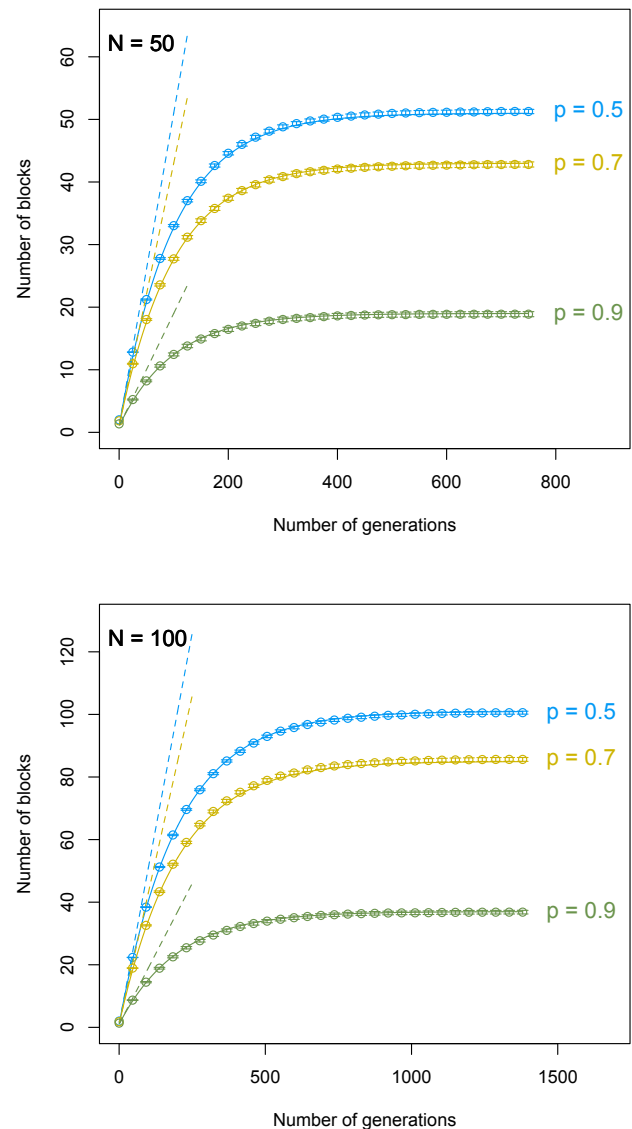


Figure 4 Number of haplotype blocks over time for stochastic simulations assuming a continuous chromosome and a population size of 50, 100 individuals (circles), the analytical prediction for an infinite population size (dotted line), or the analytical prediction for a finite population size (Equation (40), $K = 4pqN + 1$), solid line. Error bars indicate the standard error of the mean across 1,000 replicates. Shown are results for different initial frequencies of the two parental species p , such that the initial heterozygosity $H_0 = 2p(1 - p)$.

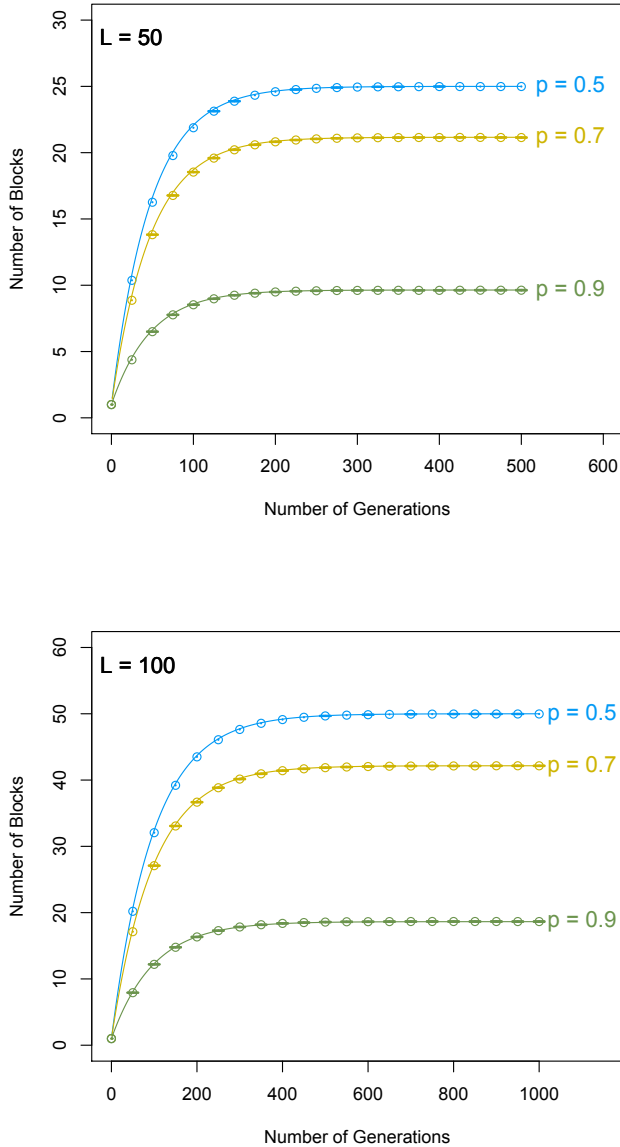


Figure 5 Number of haplotype blocks over time for either stochastic simulations assuming a continuous chromosome and a population size of 100,000 (circles), or the analytical prediction according to Equation (40), $K = 2pq(L - 1) + 1$ (solid line). Error bars indicate the standard error of the mean across 100 replicates. Shown are results for different initial frequencies of the two parental species p , such that the initial heterozygosity $H_0 = 2p(1 - p)$.

415 $1) + 1)(1 - 1/(aN)^b)$ or $(2pq(L - 1) + 1)(1 - \exp(-aN + b))$,
 416 however, we found $K_N = (2pq(L - 1) + 1)N/(aN + b)$, to be
 417 easiest to fit to the data. Using non-linear least squares esti-
 418 mation we find that for $L = 100$ and $p = 0.5$, $a = 1.014$ and
 419 $b = 47.83$ (Figure 6 A). We can repeat this process for differ-
 420 ent values of L (but keeping $p = 0.5$) and find that a is always
 421 close to 1 and that b is always close to $2pq(L - 1) + 1$. Hence,
 422 it makes more sense to find our approximation in the form:
 423 $K_N = c(2pq(L - 1) + 1)N/(aN + (2pq(L - 1) + 1))$ (fitting
 424 more than 2 parameters at the same time tends to lead to in-
 425 accurate results). We find, by fitting to varying values of N
 426 and p , that a and c both tend to $4pq$. With that, we obtain the
 427 approximation for K :

$$K = \frac{4pqN(1 + 2pq(L - 1))}{4pqN + 1 + 2pq(L - 1)}. \quad (42)$$

428 We recover both limits of K for $N \rightarrow \infty$ and $L \rightarrow \infty$, and formu-
 429 late our general approximation as:

$$K = \frac{(1 + 4pqN)(1 + 2pq(L - 1))}{2 + 4pqN + 2pq(L - 1)} \quad (43)$$

430 Comparing values of K expected following Equation (43) with
 431 the observed mean estimates from the simulations confirms that
 432 our approximation provides estimates that are close to the mean
 433 estimates (Figure 6 C, Observed vs Expected, intercept = -0.0727,
 434 slope = 0.9987, $R^2 = 0.9999$). Furthermore, equation (43) reduces
 435 to Equation (7) when $L \gg N$, and reduces to Equation (14)
 436 when $N \gg L$. As such, Equation (43), albeit an approximation
 437 for K , encompasses all combinations of N and L . To extend
 438 Equation (43) towards M recombinations per meiosis, e.g. to
 439 chromosomes of any maplength in Morgan, we can suffice with
 440 a more sparse grid. Here we only need to show (numerically)
 441 that we can correct K for M recombinations, and substitute the
 442 corresponding equations into Equation (43). It has been shown
 443 previously that K , for $L \rightarrow \infty$ is given by: $K = 4pqMN + 1$
 444 (Buerkle and Rieseberg 2008) and K for $N \rightarrow \infty$ is given by:
 445 $K = 2pq(L - 1) + 1 - pq(M - 1)$ (Equation (31)). Substituting
 446 these expressions for K into Equation (43) we obtain

$$K = \frac{(1 + 4pqMN)(1 + 2pq(L - 1) - (M - 1)pq)}{2 + 4pqMN + 2pq(L - 1) - (M - 1)pq}. \quad (44)$$

447 We simulate values of K for combinations of N and L in [50,
 448 100, 150, 200], p in [0.5, 0.9] and M in [2,3,4]. We find that
 449 results from individual based simulations are very close to our
 450 predicted equation (Figure 6, intercept = -0.1352, slope = 1.0010,
 451 $R^2 = 0.9999$). Equation (44), just like Equation (43), reduces to
 452 $K = 4pqMN + 1$ for $L \gg N$, and reduces to $K = 2pqM(L - 1) +$
 453 $1 - pq(M - 1)$ for $N \gg L$.

454 As a further test of the accuracy of Equation (44), we repeat
 455 simulating values of K , but for more typical maplengths found
 456 in molecular data $M = [0.5, 0.75, 1, 1.25, 1.5, 1.75, 2]$ (using
 457 the same sparse grid for N , L and p). We interpret fractional
 458 numbers of recombinations as a mean rate, such that an average
 459 of 1.25 recombinations implies that in 25% of all meiosis events
 460 there are 2 recombinations, and in 75% of all meiosis events
 461 there is 1 recombination. Again we find that obtained mean
 462 estimates for K are close to those predicted using Equation (44)
 463 (intercept = -0.2618, slope = 1.0027, $R^2 = 0.9999$, not shown in
 464 Figure 6).
 465

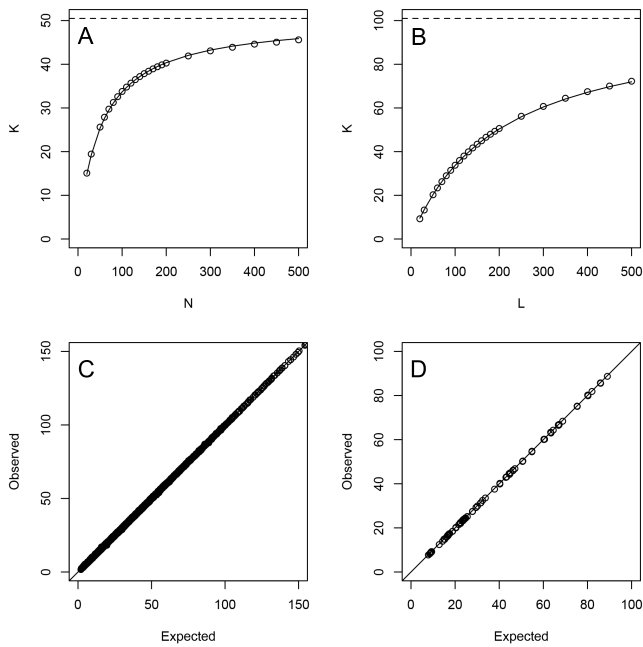


Figure 6 Numerical derivation of K for a finite population and discrete chromosome. **A:** Example relationship between K and N , for $L = 100$ and $p = 0.5$. Points are the mean of 10,000 individual based simulations. The solid line is the expected value of K following Equation (43). The dotted line represents the upper limit of K for $N = \infty$. **B:** Example relationship between K and L , for $N = 100$ and $p = 0.5$. Points are the mean of 10,000 individual based simulations. The solid line is the expected value of K following Equation (43). The dotted line represents the upper limit of K for $L = \infty$. **C:** Observed vs expected values for K for one recombination per meiosis for all combinations of N and L in [10, 20, 30 ... 200, 250, 300 ... 500], and p in [0.5, 0.6, 0.7, 0.8, 0.9] (3380 combinations, 10,000 replicates per combination). Expected values were calculated using Equation (43). $R^2 = 0.9999$, slope = 0.999. **D:** Observed vs expected values for K for M recombinations per meiosis, for all combinations of N and L in [50, 100, 150, 200], p in [0.5, 0.9] and M in [2,3,4] (96 combinations, 10,000 replicates per combination). Expected values for K were calculated using Equation (44). $R^2 = 0.9999$, slope = 1.0010.

4. Discussion

We have shown here how the number of haplotype blocks changes over time after a hybridization event. We have obtained analytical expressions for the change over time in haplotype blocks for discrete chromosomes, in infinite populations or in finite populations (including random drift) and for different frequencies of both ancestral species. Furthermore, we have developed a unifying framework that describes haplotype block dynamics for both discrete and continuous chromosomes of arbitrary map length, for any population size and for any initial frequency of both ancestral species.

We have extended standing theory for haplotype block dynamics for chromosomes with an infinite number of recombination spots, towards chromosomes that consist of discrete genetic elements with a finite number of recombination spots. Molecular data describing these elements typically presents itself in the form of genetic markers (microsatellites, SNPs, (SFP)) that

can be traced back to either of the two parental species. The genomic stretches delineated by these markers then form the genetic elements we have considered here, although in principle our framework also applies on the nucleotide level. We have assumed that the genetic elements are of equal size, and are uniformly distributed. Data based on molecular markers often deviates from this assumption, with markers non-uniformly distributed across the genome and the genomic stretches delineated by these markers differing in size. Deviations from a uniform distribution of markers across the chromosome could potentially lead to an underestimation of the number of blocks, as small blocks are more likely to remain undetected.

The universal haplotype block dynamics framework we have presented here provides a neutral expectation for the number of blocks over time. There is a wide range of processes that can cause deviations from our framework, which fall into two distinct categories: either processes acting upon the underlying genomic content, or processes that affect population dynamics. An important process acting upon the underlying genomic content is selection. If a genomic region is under selection, we expect to find a high frequency of genomic material from the beneficial parent. This reduces recombination events in this region, as recombined individuals are selected against (Kimura 1956; Lewontin and Hull 1967). As a result, we expect this region to be homozygous more often, and to have a lower potential for the formation of new blocks. We thus expect selection for either of the parental types to slow down the formation of haplotype blocks. Future work could focus on the minimal level of selection to offset neutral haplotype block dynamics or whether deviations from neutral haplotype block dynamics can be used to identify genomic areas that are under selection.

Alternatively, the formation of new blocks could speed up when the combination of alleles from both parents provides a selective advantage. The resulting overdominance favours a heterozygous genotype (Maruyama and Nei 1981), in turn favouring the formation of new blocks (equations (2), (6) and (13)). Likewise, positive epistasis among linked loci could favour a combination of alleles of different parents. As a result, recombination between these loci would be selected for, speeding up block formation. However, such increases in speed would only be seen as short bursts, because after establishment or fixation of favourable alleles, block formation is no longer affected.

A non-uniform recombination rate across the chromosome could cause deviations from neutral haplotype block dynamics. Empirical work has shown that recombination rates are often not equal across the chromosome, but are increased towards the peripheral ends of the chromosome (Lukaszewski and Curtis 1993; Pan et al. 2012; Roesti et al. 2012). As long as the chromosome is continuous, the exact shape of the recombination landscape has no effect on haplotype block dynamics. For a discrete chromosome however, haplotype block dynamics change (see Appendix A for a demonstration of increased recombination towards the peripheral ends). If some sites have an increased probability of recombination compared to others, recombination is more likely to occur at a site that has already experienced a recombination event before. As a result, the formation of new haplotype blocks is slower than expected under uniform recombination. The maximum number of blocks that can be reached remains unaffected. Similarly, the presence of recombination hotspots, areas in the chromosome with an elevated recombination rate (Gerton et al. 2000; Myers et al. 2005; Smagulova et al. 2011), also slows down the accumulation of blocks (see Ap-

pendix B for a demonstration of the effect of hotspots). Similar to increased recombination rates towards the peripheral ends, hotspots skew the recombination rate distribution to such an extent that recombination is much more likely to take place at a site that has been previously recombined, in which case no new block is formed.

Apart from processes that act upon the genomic content, population level processes are also expected to affect haplotype block dynamics. Firstly, deviations from having a constant-population size over time are expected to cause deviations from our haplotype block dynamics framework. A natural extension of our work would for instance be to include either exponentially or logistically growing populations in order to mimic real life dynamics more closely. In exponentially growing populations, the average heterozygosity does not change (Crow and Kimura 1970), which results in dynamics that resemble an infinite population. Similarly, for logistically growing populations, during the initial growth phase, haplotype block dynamics are expected to closely resemble block dynamics in an infinite population. We do have to take into account that even though the population is growing exponentially, drift effects could interfere and cause deviations from infinite population dynamics (Hallatschek *et al.* 2007). How drift and the rate of growth interact and influence haplotype block dynamics remains the subject of future study. Furthermore, in a growing population, the effect of selection is enhanced (Otto and Whitlock 1997), suggesting important interactions between selection, drift and population dynamics.

Secondly, population subdivision, founder effects, a bottleneck or a permanent decrease in population size could speed up fixation of haplotype blocks in the population through drift. Because haplotype blocks become fixed, the average heterozygosity decreases faster than expected, and the accumulation of new blocks is slowed down. Furthermore, the maximum number of blocks decreases as well (following Equation 43). Depending on the speed and timing of the decrease, individuals in the final population potentially display a larger number of blocks than expected from the current population size, retaining blocks fixed in the population before the population decreased in size.

Thirdly, secondary introgression, where admixture with the parental population after founding the hybrid population takes place, will affect haplotype block dynamics. Secondary introgression introduces new parental chromosomes that have not yet recombined and leads to an apparent reduction of the number of blocks (Pool and Nielsen 2009). The apparent reduction of the mean number of blocks effectively 'turns back time'. A secondary introgression event introduces haplotype blocks that are disproportionately large, compared to the standing haplotype block size distribution. As such, the haplotype block size distribution can potentially complement the mean number of blocks for inferring processes influencing genomic admixture after hybridization (Pool and Nielsen 2009).

Apart from the before mentioned processes, we expect that there are more processes that can affect the accumulation of blocks, including, but not limited to, sib-mating, interference between recombination events, mutation, segregation distortion and heterochiasmy. Except for overdominance or positive epistasis, all processes mentioned above slow down the accumulation of haplotype blocks. This suggests that the universal framework for haplotype block dynamics that we have presented here provides an upper limit to haplotype block dynamics.

Similar to punctuated admixture events that we have consid-

ered here, repeated or continuous gene flow between populations can result in haplotype block structures, where the genomic material of the blocks can be traced back to distinct populations (Payseur and Rieseberg 2016), but where secondary migrants introduce disproportionately large haplotype blocks into the population (Harris and Nielsen 2013; Liang and Nielsen 2014). However, in between migration or phases of increased admixture, the introduced genomic material breaks down into blocks, following similar dynamics as in our framework. Gravel has explored the impact of past migration events on the block size distribution, and was able to use the block size distribution to infer past migration events of human populations from genome data (Gravel 2012). Further studies have extended his approach, and increased the accuracy of inference, and extended his approach towards inferring effective population size and population substructuring (Palamara *et al.* 2012; Harris and Nielsen 2013; Hellenthal *et al.* 2014; Sedghifar *et al.* 2016). These studies rely on simulations in combination with likelihood methods to infer migration events from empirical data. Our framework complements that approach, contributes to a more complete understanding of the processes driving haplotype block dynamics.

We have shown here how the genomic material of two parental species mixes over time after a hybridization event. With the current advances in genomic methods, it is now possible, and affordable, to screen species for recurring haplotype blocks of other, closely related, species. Our framework can then be used inversely, by inferring the time of the hybridization event. Given that there are many processes that can potentially slow down the accumulation of haplotype blocks, inferring the time of hybridization using our universal haplotype block dynamics framework provides the lower time limit, e.g. the minimum age of the hybrid. Caution should be taken however, as our work shows that haplotype block dynamics tend to stabilize relatively quickly (on an evolutionary timescale), where typically the number of blocks reaches a maximum limit in the order of $10N$ or $10L$ generations. As a result, haplotype block patterns are especially useful for recent hybridization events.

5. Data availability

Computer code used for the individual based simulations has been made available on GitHub and can be found on: <https://github.com/thijsjanzen/Haplotype-Block-Dynamics>

Literature Cited

- Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Möst, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Väinölä, J. B. W. Wolf, and D. Zinner, 2013 Hybridization and speciation. *Journal of Evolutionary Biology* **26**: 229–246.
- Arbeithuber, B., A. J. Betancourt, T. Ebner, and I. Tiemann-Boege, 2015 Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences* **112**: 2109–2114.
- Barton, N., 2001 The role of hybridization in evolution. *Molecular Ecology* **10**: 551–568.
- Bennett, J., 1953 Junctions in inbreeding. *Genetica* **26**: 392–406.

- 666 Buerkle, C. A., R. J. Morris, M. A. Asmussen, and L. H. Rieseberg, 2000 The likelihood of homoploid hybrid speciation. *Heredity* **84**: 441–451.
- 667
- 668
- 669 Buerkle, C. A. and L. H. Rieseberg, 2008 The rate of genome stabilization in homoploid hybrid species. *Evolution* **62**: 266–275.
- 670
- 671
- 672 Chapman, N. and E. Thompson, 2003 A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical Population Biology* **64**: 141–150.
- 673
- 674
- 675 Chapman, N. H. and E. A. Thompson, 2002 The effect of population history on the lengths of ancestral chromosome segments. *Genetics* **162**: 449–458.
- 676
- 677
- 678 Crow, J. F. and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.
- 679
- 680 Edmands, S., H. Feaman, J. Harrison, and C. Timmerman, 2005 Genetic consequences of many generations of hybridization between divergent copepod populations. *Journal of Heredity* **96**: 114–123.
- 681
- 682
- 683
- 684 Fisher, R. A., 1949 *The Theory of Inbreeding*. Oliver and Boyd.
- 685
- 686 Fisher, R. A., 1954 A fuller theory of "junctions" in inbreeding. *Heredity* **8**: 187–197.
- 687
- 688 Fisher, R. A., 1959 An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity* **13**: 179–186.
- 689
- 690
- 691 Gale, J., 1964 Some applications of the theory of junctions. *Biometrics* pp. 85–117.
- 692
- 693
- 694 Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes, 2000 Global mapping of meiotic recombination hotspots and coldspots in the yeast *saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences* **97**: 11383–11390.
- 695
- 696
- 697 Grant, V., 1981 *Plant speciation*. Columbia University Press.
- 698
- 699
- 700 Gravel, S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607–619.
- 701
- 702
- 703
- 704 Hallatschek, O., P. Hersen, S. Ramanathan, and D. R. Nelson, 2007 Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences* **104**: 19926–19930.
- 705
- 706
- 707
- 708
- 709
- 710
- 711
- 712
- 713
- 714
- 715
- 716
- 717
- 718
- 719
- 720
- 721
- 722
- 723
- 724
- 725
- 726
- 727
- 728
- 729
- 730
- 731
- 732
- 733
- 734
- 735
- 736
- 737
- 738
- 739
- 740
- 741
- 742
- 743
- 744
- 745
- 746
- 747
- 748
- 749
- 750
- 751
- 752
- 753
- 754
- 755
- 756
- 757
- 758
- 759
- 760
- 761
- 762
- 763
- 764
- 765
- 766
- 767
- 768
- 769
- 770
- 771
- 772
- 773
- 774
- 775
- 776
- 777
- 778
- 779
- 780
- 781
- 782
- 783
- 784
- 785
- 786
- 787
- Martin, O. C. and F. Hospital, 2011 Distribution of parental genome blocks in recombinant inbred lines. *Genetics* **189**: 645–654.
- Maruyama, T. and M. Nei, 1981 Genetic variability maintained by mutation and overdominant selection in finite populations. *Genetics* **98**: 441–459.
- McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, 2004 The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581–584.
- Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321–324.
- Nolte, A. W., J. Freyhof, K. C. Stemshorn, and D. Tautz, 2005 An invasive lineage of sculpins, *cottus* sp. (pisces, teleostei) in the rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proceedings of the Royal Society B* **272**: 2379–2387.
- Nolte, A. W. and D. Tautz, 2010 Understanding the onset of hybrid speciation. *Trends in Genetics* **26**: 54–58.
- Otto, S. P. and M. C. Whitlock, 1997 The probability of fixation in populations of changing size. *Genetics* **146**: 723–733.
- Palamara, P. F., T. Lencz, A. Darvasi, and I. Pe'er, 2012 Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics* **91**: 809–822.
- Pan, Q., F. Ali, X. Yang, J. Li, and J. Yan, 2012 Exploring the genetic characteristics of two recombinant inbred line populations via high-density snp markers in maize. *PLoS ONE* **7**.
- Payseur, B. A. and L. H. Rieseberg, 2016 A genomic perspective on hybridization and speciation. *Molecular Ecology*.
- Pool, J. E. and R. Nielsen, 2009 Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**: 711–719.
- Roesti, M., A. P. Hendry, W. Salzburger, and D. Berner, 2012 Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology* **21**: 2852–2862.
- Sedghifar, A., Y. Brandvain, and P. Ralph, 2016 Beyond clines: lineages and haplotype blocks in hybrid zones. *Molecular Ecology* pp. n/a–n/a.
- Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. M. Hooper, A. I. Strand, Q. Li, B. Raney, C. N. Balakrishnan, S. C. Griffith, G. McVean, and M. Przeworski, 2015 Stable recombination hotspots in birds. *Science* **350**: 928–932.
- Smagulova, F., I. V. Gregoretta, K. Brick, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova, 2011 Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472**: 375–378.
- Stam, P., 1980 The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research* **35**: 131–155.
- Stemshorn, K. C., F. A. Reed, A. W. Nolte, and D. Tautz, 2011 Rapid formation of distinct hybrid lineages after secondary contact of two fish species (*cottus* sp.). *Molecular Ecology* **20**: 1475–1491.

A. Appendix

A. Non-uniform recombination rate

In the main text, we have assumed the recombination rate to be uniformly distributed along the chromosome. Such an as-

788 assumption especially applies to chromosomes mapped in Morgan,
 789 but does not apply to chromosomes where we track haplotype
 790 blocks in physical distance (basepairs). Typically, the recombina-
 791 tion rate across the chromosome changes considerably, where it
 792 is not uncommon to recover an increase in the recombination rate
 793 towards the peripheral ends of the chromosomes (Lukaszewski
 794 and Curtis 1993; Roesti *et al.* 2012). To study the effect of such
 795 changes in recombination rate, we have simulated haplotype
 796 block dynamics assuming a 10 fold increase in the probability of
 797 a crossover event close to the peripheral ends compared to close
 798 to the center of the chromosome. A 10 fold increase in recombina-
 799 tion rate is in line with empirical studies (Lukaszewski and
 800 Curtis 1993; Roesti *et al.* 2012), but should mainly be interpreted
 801 as an illustration of the effect of elevated recombination rates to-
 802 wards the peripheral ends, rather than an attempt to accurately
 803 mimic empirical patterns. Because crossover events are more
 804 likely towards the peripheral ends, the probability of a crossover
 805 event taking place at a site that has experienced crossover be-
 806 fore increases. As a result we notice that the accumulation of
 807 blocks is slower than under the uniform recombination rate pre-
 808 dictions (Figure 7). The maximum number of blocks remains
 809 unaffected, and different initial ratios between the two ancestral
 810 species p does not change the general pattern of a slowdown of
 811 accumulation of haplotype blocks.

812 B. Hotspots

813 Non-uniform recombination rates might also manifest them-
 814 selves as a result of recombination hotspots. Recombination
 815 hotspots delineate areas in the genome that recombine more of-
 816 ten than other areas, and are well documented in many different
 817 species (Gerton *et al.* 2000; Myers *et al.* 2005; Mackiewicz *et al.*
 818 2013; Singhal *et al.* 2015; Arbeithuber *et al.* 2015; Smagulova *et al.*
 819 2011). The density of recombination hotspots is generally large
 820 with hotspots occurring every 200kb. Depending on the size of
 821 the chromosome, this results in a total number of hotspots per
 822 chromosome between 300 and 1200. To demonstrate the effect of
 823 hotspots on haplotype block dynamics, we model a chromosome
 824 consisting of 1000 genomic elements ($L = 1000$). We assume
 825 100 recombination hotspots scattered randomly across the chro-
 826 mosome, where the recombination rate is 9 times the normal
 827 recombination rate, such that 50% of all recombination events
 828 take place in recombination hotspots (McVean *et al.* 2004). The
 829 locations of the recombination hotspots were determined *a priori*
 830 and kept constant across replicates and parameter settings. We
 831 find that recombination hotspots slow down the formation of
 832 haplotype blocks (Figure 8), as recombination is more likely to
 833 take place at a site that has previously recombined, compared
 834 to a uniform recombination rate. The maximum number of
 835 haplotype blocks remains unaffected.

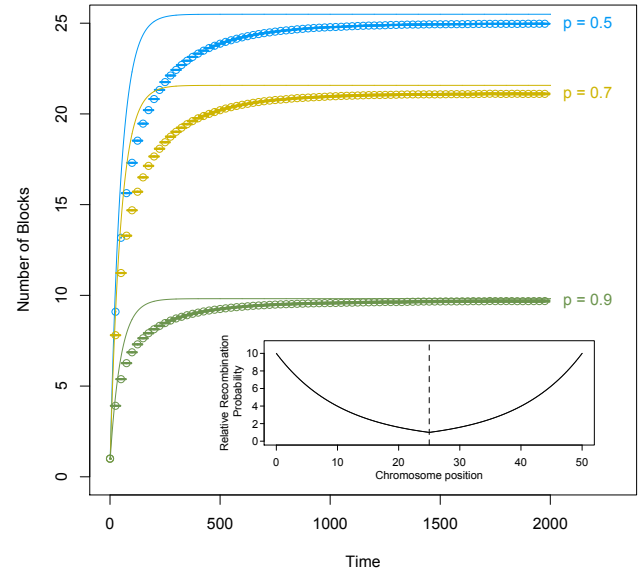


Figure 7 Individual based simulation results for the mean number of blocks over time, assuming an exponential recombination rate distribution with a 10 times higher recombination rate towards the peripheral ends of the chromosome. The recombination rate of site i along a chromosome of length L , assuming the centromere is located at position $L/2$ is then given by: $P(R, i) = \exp\left(\log(10) \frac{|2i-L|}{L}\right)$. The population size N is 100,000 individuals and the number of chromosome elements L is 50. The solid line shows the mean number of blocks assuming a uniform recombination rate and infinite population size. Error bars show the standard error of the mean across 100 replicates. The inset graph shows the recombination rate across the chromosome relative to the recombination rate at the centromere, with the dotted line indicating the position of the centromere (at $L/2$).

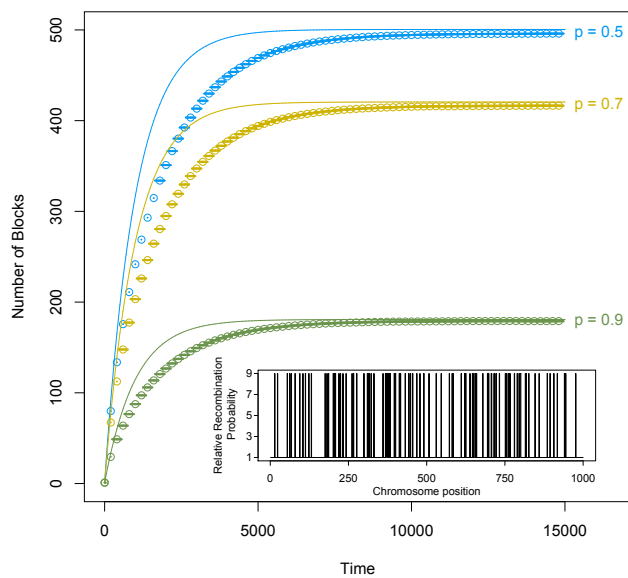


Figure 8 Individual based simulation results for the mean number of blocks over time, assuming 100 randomly placed recombination hotspots on a chromosome consisting of 1000 genomic elements ($L = 1000$), the population size is 100,000. The recombination rate in a hotspot is 9 times the recombination rate outside the hotspots, as a result of which, 50% of all recombination events take place in the recombination hotspots. The dotted line shows the mean number of blocks assuming a uniform recombination rate and infinite population size without hotspots. Error bars show the standard error of the mean across 100 replicates. The inset graph shows the locations of the hotspots across the simulated chromosome.