# Distribution of Parental Genome Blocks in Recombinant Inbred Lines

**Olivier C. Martin*,†,1 and Frédéric Hospital‡**

*INRA, Université Paris-Sud, Centre National de la Recherche Scientifique, AgroParisTech, UMR0320/UMR8120 Génétique Végétale, F- 91190 Gif-sur-Yvette, France, †Université Paris-Sud, Centre National de la Recherche Scientifique, UMR 8626 Laboratoire de Physique Theorique et Modeles Statistiques, F-91405, Orsay, France, and ‡INRA, UMR1313 Génétique Animale et Biologie Intégrative, 78352 Jouy-en-Josas, France

**ABSTRACT** We consider recombinant inbred lines obtained by crossing two given homozygous parents and then applying multiple generations of self-crossings or full-sib matings. The chromosomal content of any such line forms a mosaic of blocks, each alternatively inherited *identically by descent* from one of the parents. Quantifying the statistical properties of such mosaic genomes has remained an open challenge for many years. Here, we solve this problem by taking a continuous chromosome picture and assuming crossovers to be noninterfering. Using a continuous-time random walk framework and Markov chain theory, we determine the statistical properties of these identical-by-descent blocks. We find that successive block lengths are only very slightly correlated. Furthermore, the blocks on the ends of chromosomes are larger on average than the others, a feature understandable from the nonexponential distribution of block lengths.

WITH the advent of dense genomic maps, in particular based on single-nucleotide polymorphism (SNP) data, the study of haplotypes has become central for modern analyses in population genetics (Buckler and Gore 2007; Carlton 2007; Frazer *et al.* 2007; Mott 2007; Jakobsson *et al.* 2008; Bryc *et al.* 2010). Here, the term haplotype refers to the series of alleles that an individual carries on a chromosome pair at a collection of (possibly many) loci and contrasts with single-locus genotypes that were the objects of many past studies. Haplotypic information can be used for association studies (Gold *et al.* 2008), for diversity studies (Lindblad-Toh *et al.* 2005), or for recognizing signals of positive selection using various measures of haplotype homozygosity (Sabeti *et al.* 2002; Zhang *et al.* 2006; Lencz *et al.* 2007; Tang *et al.* 2007; Curtis *et al.* 2008). Many approaches capitalize on the apparent "block" structure of haplotypes (Stumpf 2002; Cardon and Abecasis 2003; Wall and Pritchard 2003; Altshuler *et al.* 2005; Zheng and McPeek 2007). Various causes can be called upon to explain the apparent structuration of genomes in haplotype blocks (Tishkoff and Verrelli 2003; Zondervan and Cardon 2004; Pe'er *et al.* 2006), among which are recombination hotspots (Goldstein 2001; Jeffreys *et al.* 2001) and population structure (Pritchard *et al.* 2000; Grote 2007; Slate and Pemberton 2007). However, the situation is often complicated (Shifman *et al.* 2003; Yalcin *et al.* 2004; Cuppen 2005; Kauppi *et al.* 2005; Greenawalt *et al.* 2006; Moore *et al.* 2008). In particular, the theoretical properties of many of the objects mentioned above, *e.g.*, haplotype block lengths, remain largely unknown. Often, the distribution of blocks is declared "nonrandom" (Curtis *et al.* 2008) although the null hypothesis is not clearly specified.

The task of determining statistical properties of chromosomal block structures has arisen in many different contexts. These can be classified into two types according to the kind of populations considered and lead to different mathematical techniques. In the first class, one asks how the genome of one or more parents in a population gets broken up into blocks at successive generations and how different descendents may share *identical-by-descent* (IBD) blocks. The framework most generally taken allows for random mating between individuals, a stochastic number of offspring for each individual, and possibly population growth; because


doi: 10.1534/genetics.111.129700
Manuscript received December 24, 2010; accepted for publication July 12, 2011
Supporting information is available online at http://www.genetics.org/content/suppl/2011/08/12/genetics.111.129700.DC1.
[1]Corresponding author: INRA, UMR 0320 Génétique Végétale, F-91190 Gif-sur-Yvette, France. E-mail: olivier.martin@u-psud.fr

of this stochasticity, the mathematical theory of branching processes plays a key role. The second class on the contrary assumes complete knowledge of all genealogies and so is relevant only for controlled crosses. But because the corresponding framework is thus constrained, Markov chains can be used to follow the statistics of IBD blocks and even how the genomes of *all* founding parents get shared among descendants.

The mathematical treatments in the first class typically build on Fisher's "theory of junctions" (Fisher 1949). Fisher defined a *junction* in a chromosome as a boundary point between segments descended by different routes from the founders. Once formed by crossovers, junctions can be inherited, just like point mutations. Fisher (1949, 1954, 1959) and Bennett (1953) investigated the expected number of chromosomal regions separated by junctions for different systems of inbreeding (repeated selfing, repeated sib mating, repeated parent–offspring mating, etc.). Stam (1980) extended Fisher's theory of junctions to a random mating population of constant size and any number of generations. Furthermore, he was able to derive the "probability distribution of the heterogenic part of the genome" (and not just the expected number of fragments) by assuming that the fragments were exponentially distributed—a critical hypothesis that was justified by numerical simulations. Chapman and Thompson (2003) extended Stam's work to the case of a subdivided population. They were also able to relax the hypothesis of an exponential distribution and showed that the IBD tracts of chromosomes followed a distribution not quite exponential, having in fact a fat tail. They also determined how these properties were affected by the population size. Analogous work by Baird *et al.* (2003) focused on the case of a formally infinite population; this simplifies the problem because related individuals never mate with one another. Furthermore, they worked in the approximation of allowing only 0 or 1 crossover at each meiosis, a case sometimes referred to as complete interference; then each individual can carry at most just one block from the reference founding parent. Within this framework, they were able to treat the problem exactly, deriving in particular the distribution of the number of descendants containing blocks and the first two moments of these block sizes.

The mathematical treatment of the second class was initiated by Donelly (1983). Numerous studies since have derived exact mathematical results on different kinds of pedigree systems (Slatkin 1972; Franklin 1977; Donelly 1983; Guo 1994; Bickeboller and Thompson 1996a,b; Stefanov 2000; Browning and Browning 2002; Cannings 2003; Dimitropoulou and Cannings 2003; Ball and Stefanov 2005; Walters and Cannings 2005; Rodolphe *et al.* 2008). Such studies map the IBD problem to that of a random walk on a pedigree-dependent graph. It is that Markovian framework which we use here in the context of recombinant inbred lines, a particular kind of pedigree that has the additional complication of allowing for an infinite number of generations. We provide a description that is mathematically rigorous but also of practical use.

Recombinant inbred lines (RILs) can be derived by either self-fertilization (plants) or brother–sister matings (animals). RILs have become a tool of choice for animal and plant studies [genetic maps, QTL detection, and association studies (Churchill *et al.* 2004; Churchill 2007; Crow 2007; Keurentjes *et al.* 2007; Yu *et al.* 2008)]. Moreover, such lines are fixed and provide ever-lasting replicable reference homozygous genomes; these are very useful to dissect complex traits and estimate epistatic effects or genotype × environment interactions (Bergland *et al.* 2008; Maccaferri *et al.* 2008; Alcazar *et al.* 2009). To produce a RIL, one typically starts with $F_1$ hybrids derived from the cross of two homozygous parents, say $P_A$ and $P_a$. Offspring are generated from these $F_1$ and the process is repeated for many generations; this can be done by selfing [*single-seed descent* (SSD)] or by full-sib mating (hereafter referred to as "SIB"). At each generation mean heterozygosity decreases and in fact the process tends toward homozygosity at all loci. Due to the formation of crossovers during meiosis, the genomes at each generation are mixtures of the two parental ones, in which closer loci have a higher probability of descending from the same parent $P_A$ or $P_a$. The fixed genomes then form successions of blocks, each block being IBD to one of the two parents. In effect, we have a mosaic genome for the RIL, patching together pieces from each parent $P_A$ or $P_a$. What is the mean length of blocks? We shall see that it is 0.5 M in SSD and 0.25 M in SIB if the chromosome genetic length is large. But one may also ask what is the block length distribution, what is the mean number of blocks on a finite chromosome, or even what are the analogous statistics before all loci are fixed. As genome coverage becomes dense or as one approaches a nucleotide-level description of genomes [be it for association studies or genomic selection (Meuwissen *et al.* 2001)], one is inevitably driven toward a continuous chromosome picture, requiring block-like descriptions. Here, we address the need to work at this level where blocks are the elements of interest. In particular, we show how to calculate block statistics in RILs, using a mix of combinatorial analysis and probability theory.

## Model and Methods

### *Junctions*

For all our work, we deal with diploid organisms and are concerned with the construction of RILs in both SSD and SIB. Each chromosome pair is subject to independent dynamics, so without loss of generality we can focus on the case of a single chromosome pair. Furthermore, the objects of study are the IBD blocks, hereafter simply referred to as blocks. The $F_1$ is considered to be the zeroth generation ($g = 0$). To go from one generation to the next, we produce one offspring in the case of SSD and a brother–sister pair in the case of SIB. An offspring individual is the union of two gametes, each of which is produced by a parent through meiosis, during which there can be crossovers. (In the case

of SSD, there is just one parent.) Once a locus is fixed, it stays so forever. In SSD this simply requires the two homologous chromosomes to have the same allele at that locus; in SIB, it requires all four chromosomes to have the same allele.

We measure continuous positions on the chromosome in morgans, with the leftmost end of the chromosome corresponding to the origin of our axis; *i.e.*, $x = 0$. Following the work of Fisher (1949, 1954, 1959), Bennett (1953), and Donelly (1983), a crossover is referred to as a "junction" and is identified with an arbitrarily precise point on the chromosome. We assume that crossovers arise without interference; then the production or not of a junction in the interval $[x, x + dx]$ is independent of occurrences of junctions in any other interval. Here $x$ denotes the genetic position, $dx$ is infinitesimal, and junctions arise with density 1 along the chromosome. Figure 1 illustrates the use of these junctions when following successive generations under SSD. At one generation, consider the pair $(H, H')$ of homologous chromosomes. A meiosis takes place and results in an offspring chromosome (gamete) that is a mosaic of chromosome segments coming from either $H$ or $H'$. A junction separates two adjacent segments. We use a binary label (0 or 1) to specify the origin ($H$ or $H'$) of each segment. For any position $x$, having the list of its labelings for all generations $g$ allows us to determine the IBD content at $x$ as shown on the right-hand side of Figure 1. Note that the numbering of the junctions is done from left to right, *not* as a function of its occurrence in generations. The successive steps of the procedure are shown in Supporting Information, Figure S1: first one lays out the junctions and their numbering, then one introduces the binary labels across each junction, and finally one reconstructs the haplotypes (see File S1). As illustrated in Figure S2, the case of SIB mating is analogous, and again at each generation a chromosome consists of a mosaic of the founding parents' chromosomes (see File S1).

So far, the introduction of junctions can be formulated for any pedigree system. Many previous studies have done this and mapped the IBD problem to that of a continuous-time random walk on a pedigree-dependent graph (Slatkin 1972; Franklin 1977; Donelly 1983; Guo 1994; Bickeboller and Thompson 1996a,b; Stefanov 2000; Browning and Browning 2002; Cannings 2003; Dimitropoulou and Cannings 2003; Ball and Stefanov 2005; Walters and Cannings 2005; Rodolphe *et al.* 2008). Here we consider RILs and then the Markovian framework's pedigree-dependent graph is a hypercube. The sequence of binary labels at any given locus $x$ specifies a unique vertex of the hypercube, and how this vertex changes as one moves along the chromosome determines the block structure. Note that some of this mathematical framework is close in spirit to that used for studying the coalescent in the presence of recombination; there the central object is the so-called ancestral recombination graph, and the problem (Wiuf and Hein 1999; McVean and Cardin 2005) is to describe how this graph changes with position along a continuous chromosome. This is a very difficult problem and so the authors of those studies derived relatively few exact results.

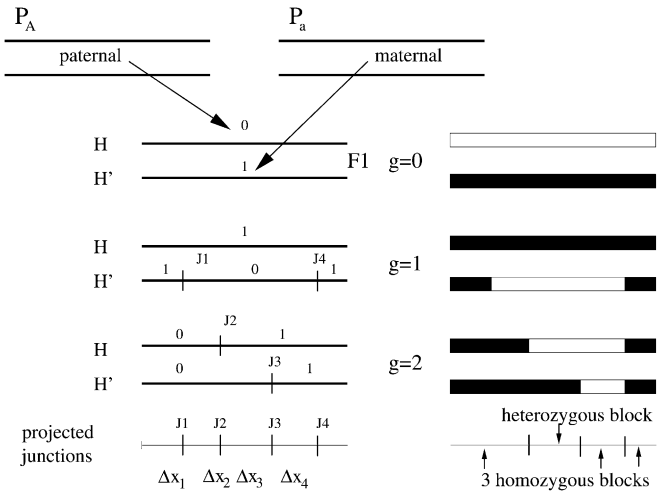



**Figure 1** Labeling in a SSD RIL. At each generation the homologs are called $H$ and $H'$. To keep track of the IBD property, for each point on the continuous chromosome under consideration we specify the origin ($H$ or $H'$ in the parent) using a 0–1 label, covering zones separated by junctions. The genotype at any generation $g$ can be reconstructed from these binary numbers as shown on the right. Note that a junction need not separate two blocks.

The application of this Markovian framework for SSD and SIB RILs requires considering all possible continuous-time walks on a high-dimensional hypercube. We do this in two steps. First, we enumerate by computer all possible discrete time random walks on that hypercube. Then we tackle the continuous waiting times of the original walks by analytical techniques. Finally, the numerical treatment of these analytical expressions is performed using Mathematica (Wolfram 1991). The C and Mathematica codes for these different tasks are provided in File S2.

### *The continuous-time Markov process*

When creating the generations 1 to $g$, $2g$ gametes are produced in the SSD mating scheme, and $4g$ gametes are produced in the SIB mating scheme. Denote this number by $N_c$ as it is also the number of (new) chromosomes produced in the RIL construction (remember we follow only one chromosome pair). Rather than follow each gamete from one generation to the next, it is (more) useful to consider all $N_c$ gametes *simultaneously*. This can be visualized by stacking the pairs of chromosomes for all generations on top of each other and then scanning the chromosome stack from left to right to see where the junctions appear in order of increasing $x$.

Of great importance is the fact that the junctions on these $N_c$ gametes are *independent* in all respects: having a junction on one gamete in $[x, x + dx]$ does not affect the probability of having another junction elsewhere, be it on the same gamete or on any other gamete. Because of this independence, one can think of the production of junctions among

the whole set of gametes as being a "continuous-time" Markov process (Feller 1950), where $x$ plays the role of time. For the interval $[x, x + dx]$, a junction arises with probability $N_c dx$, and then if such an *event* is realized the junction is assigned randomly to one of the $N_c$ gametes (each with probability $1/N_c$). The operation is then repeated for the interval $[x + dx, x + 2dx]$, and so forth. We thus have a Markov process where interevent intervals are independent and distributed as

$$\rho(\Delta x) = N_c \exp(-N_c \Delta x) \tag{1}$$

while junction assignments to chromosomes are done equiprobably.

### Discrete and continuous-time random walks on the hypercube

We initialize the binary labels at $x = 0$ randomly and uniformly because segregation is unbiased. The continuous-time Markov process extends these labels from $x = 0$ toward increasing $x$. At any given $x$, and using a {0, 1} notation for each binary label, we call $\mathcal{M}$ the map from the $N_c$ dimensional hypercube $\mathcal{H} = \{0,1\}^{N_c}>$ to the genotypes at generation $g$; this map can be thought of as a *coloring* of the vertices of the $N_c$-dimensional hypercube. There are as many colors as there are one-locus genotypes at generation $g$: 4 for SSD and 16 for SIB. Then the block pattern at generation $g$ can be "read off" by examining the succession of vertex colors visited by the Markov walk on the hypercube. Note that having a junction appear at $x$ corresponds to hopping to a random neighboring vertex on $\mathcal{H}$ at that time, while the residence time on each vertex is exponentially distributed (*cf*. Equation 1).

For the block *statistics*, we want to find the probability that the walk on $\mathcal{H}$ leads to a given pattern of successive colors. For this, we sum over all possible walks compatible with the desired pattern. The crucial point is that the continuous variables of the interjunction values affect the lengths of the blocks, but not the pattern of successive blocks. This allows us to decompose the problem of block statistics into two parts. The first comes from the discrete set of possibilities for the sequence of vertices visited on $\mathcal{H}$ (the "topology" of the junctions); we use the master equation of the discrete-time random walk on the hypercube to track these sequences. The second is associated with the continuous nature of the junction–junction intervals, which involves summing over known probability distributions derived from Equation 1.

### Extracting block length distributions

Consider the simplest observable: the length of the first block along the chromosome. If the block is heterozygous, then its length distribution in SSD is that of the distance of the first junction and is given by Equation 1. Indeed, the locus $x = 0$ must be heterozygous at generation $g$, but in SSD, at the very first hop of our random walk on the hypercube, the heterozygous block will end, starting a fixed block. The situation is more instructive if the first block is homozygous (fixed). To calculate the length distribution of that first block, we first consider all possibilities for the different walks from the starting vertex (defined from the situation at $x = 0$). A walk will maintain the homozygous structure at generation $g$ for perhaps a few hops and then one hop will change that. If $k$ is the first hop that ends the block, we can collect together all the discrete time walks that have the same $k$. We thus define $P^{(1)}(k)$ as the probability to perform $k - 1$ hops while staying at generation $g$ in the same fixed state as that of $x = 0$ and to then terminate the block at the $k$th hop. $P^{(1)}(k)$ is the sum of the probabilities of all discrete time walks on $\mathcal{H}$ that are compatible with staying in the first fixed state during exactly $k - 1$ hops. Because the number of such walks grows exponentially with $k$, it is best to determine this quantity by recursions rather than by enumerations. This is precisely what is done when using the associated master equation. Each iteration of that equation updates a vector on the hypercube and generates the successive $P^{(1)}(j)$. To obtain $P^{(1)}(k)$ one has to perform $k$ iterations of the master equation. File S1 specifies this master equation, the initialization of the vector iterated, and the relation between the iterated vector and $P^{(1)}(j)$; the C programs for implementing these iterations are also provided (see File S2).

Given the $P^{(1)}(k)$ probabilities, we can reintroduce the continuous times spent on each vertex of the hypercube to get the distribution of the length of the first bloc. Indeed, for SSD as well as for SIB, for all walks that contribute to this situation, $x$ will go from 0 to $x_k$ with $x_k$ distributed as a rescaled Gamma distribution,

$$\rho_k(x_k) = \frac{N_c^k x_k^{k-1} \exp(-N_c x_k)}{(k-1)!} \tag{2}$$

as this the distribution of the sum of $k$ independent exponentially distributed variables. The distribution of $\ell_1$, the length of the first block (assuming it is fixed) is then given by

$$\mu^{(1)}(\ell_1) = \sum_{k=1}^{\infty} P^{(1)}(k) \rho_k(\ell_1). \tag{3}$$

This result holds for an infinite chromosome. For a finite chromosome of length $L$, we note that if $\ell_1 > L$, we have "stepped off" the finite chromosome. Thus, if the value of $\ell_1$ (distributed as in Equation 3) is greater than $L$, we see that on the finite chromosome the block is actually only $L$ long. Thus to adapt Equation 3 to a finite chromosome, we simply keep the distribution as is when $\ell_1 < L$ while for all those values $\ell_1 \geq L$ we set $\ell_1 = L$. Mathematically, this generates a delta function at that point of weight given by the probability that $\ell_1 \geq L$. This derivation corresponds to a simple truncation of a distribution, and it can be extended to other observables. These include the length of the $n$th block,

which requires calculating the probabilities $P^{(n)}(k)$ of stepping off the $n$th block after $k$ steps, and the joint distribution of lengths for different blocks. Details on such derivations are given in File S1. The Mathematica codes for performing sums like those in Equation 3 are also provided in File S2.

## Results

### *Infinite and semi-infinite chromosomes*

***Mean block lengths:*** As the number $g$ of generations increases, alleles become fixed and since nothing changes thereafter the statistics of the blocks must have a limit at large $g$. In that situation, there is only alternation of IBD blocks homozygous of type $P_A$ or $P_a$. We focus on the statistics of such blocks, either at arbitrary $g$ or in the $g \to \infty$ limit, approximated by taking $g$ large enough. However, our computational framework applies to arbitrary blocks, homozygous or heterozygous. From the practical point of view, we are limited computationally by the part of the algorithm that follows occupation probabilities on the hypercube $\mathcal{H}$: executing this master equation on the computer uses $O(N_J N_c 2^{N_c})$ operations where $N_J$ is the maximum number of hops of the walks; this restricts our study to 14 generations in SSD and 7 generations in SIB.

A simple statistic of blocks is their mean length $\langle \ell \rangle$. This quantity is related to the density $\eta$ of block extremities: on a very large chromosome of size $L$, the number of blocks $n$ will satisfy $n/L \approx \eta$ while $\langle \ell \rangle \approx L/n$. For SSD RILs the density $\eta$ is known to be 2 while it is 4 in SIB. Such a result follows by considering a small interval $[x_1, x_2]$ and asking that the interval be recombinant. In SSD this occurs with probability $R = 2r/(1 + 2r)$, where $r$ is the recombination rate per meiosis between $x_1$ and $x_2$ (Haldane and Waddington 1931). Taking $x_2 - x_1$ to be infinitesimal, we get $r \approx (x_2 - x_1)$ (Haldane 1919) and so $R \approx 2(x_2 - x_1)$; noting that recombination then implies the presence of a block extremity in this interval, we see that the density of block extremities is 2. Setting $\eta = 2$, one obtains directly $\langle \ell \rangle = \frac{1}{2}$, valid at large $g$ and for large chromosomes. An identical reasoning gives $\langle \ell \rangle = \frac{1}{4}$ for SIB since in this case $R = 4r/(1 + 6r)$.

***Distribution of block lengths:*** Getting the *distribution* of a block length cannot be achieved by such shortcuts. Instead, we generalize Equation 3, again at very large $g$ and for a very long (semi-infinite) chromosome. Starting from $x = 0$, the successive block lengths are $\ell_1, \ell_2, \ldots$ ; as the block number $n$ increases, the distributions of $\ell_n$ tend toward a limiting distribution $\mu^*(\ell)$ that has no memory of the state at the chromosome's origin. The computation of the distribution at any given $n$, in direct analogy with what was done for $\ell_1$, requires calculating the probabilities $P^{(n)}(k)$ of staying on the $n$th block during $k - 1$ hops and stepping off at the $k$th one. Again, we use the master equation to compute these quantities iteratively; *cf.* File S1. Then the distribution of $\ell_n$ is obtained by replacing the $P^{(1)}(k)$ in Equation 3 by $P^{(n)}(k)$. These successive distributions are displayed for SSD in Figure 2. Note that the convergence in $n$ is very rapid; only the first block is visibly different from the others. In File S1, we provide a parameter-free approximation to $\mu^*(\ell)$ that works quite well as shown in Figure S3. In the case of SIB RILs, the convergence with $n$ is much slower as shown in Figure S4.

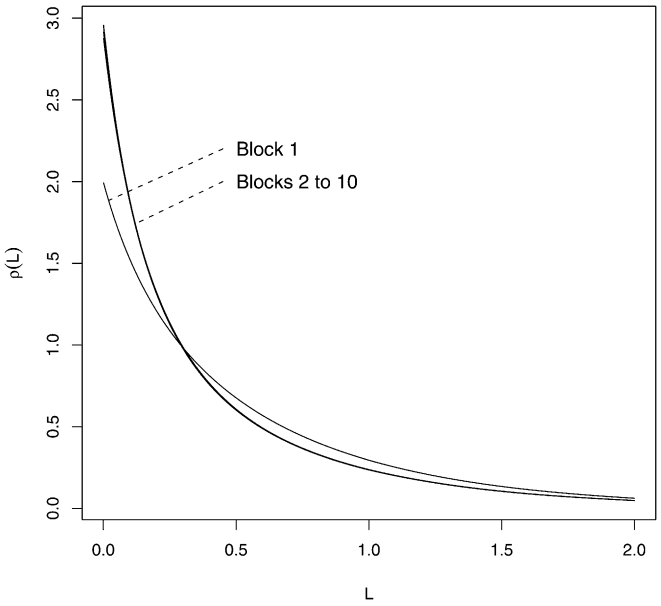



**Figure 2** Block length distribution. Displayed is the probability density of homozygous block length for block number 1, 2, . . . , 10 in SSD. The chromosome is semi-infinite, and the number of generations is large to be in fixation; except for the first block, the curves seem to superpose but in fact are distinct.

A log–log plot of these distributions shows that they are not exponential, though in the tail they all are well approximated by an exponential; such a form in the tail is a consequence of the spectral decomposition of the Markov process (see File S1). Note that these distributions for $\ell_1, \ell_2, \ldots$ must all decay at the same asymptotic rate. Another important point is that the distribution of $\ell_2$ is slightly different from that of $\ell_3$, proving that the successive lengths are *not independent*: the block lengths are *not* generated by a stationary renewal process.

Although we were not able to derive the analytic form of $\mu^*$, we nevertheless have

$$\mu^*(\ell \to 0) = 3 \text{ in SDD} \quad \text{and} \quad \mu^*(\ell \to 0) = 7 \text{ in SIB}. \qquad (4)$$

This can be proved by relating $\mu^*(\ell \to 0)$ to double-recombinant frequencies as follows. We take two successive intervals $I_{1,2}$ and $I_{2,3}$, each of length $dx$ (infinitesimal), and ask what the probability is that the intervals are both recombinant. There is a probability $2dx$ in SSD and $4dx$ in SIB that the first interval is recombinant (recall that the densities of block extremities are respectively 2 and 4). Then given this first extremity, the probability that there is another extremity in the second interval is $\mu^*(\ell \to 0)dx$, leading to a total probability of $2dx\mu^*(\ell \to 0)dx$ in SSD and $4dx\mu^*(\ell \to 0)dx$ in SIB. However, this double-recombinant

probability can also be computed (Martin and Hospital 2006) in terms of the three recombination frequencies $R_{1,2}$, $R_{2,3}$, and $R_{1,3}$ associated with the locus pairs (1, 2), (2, 3), and (1, 3). In the limit of small $dx$, it is $(2dx)^2 3/2$ for SSD and $(4dx)^2 7/4$ for SIB. Identifying the different expressions, we get the claimed results.

Analogous studies can be performed at given values of $g$, including or not heterozygous blocks. Recalling that fixation arises rather rapidly in SSD, it is no surprise that the statistics of homozygous blocks converge quickly to a large $g$ limit. For example, if one computes in SSD the length distribution of the first block when it is homozygous, one finds that it does not vary much for $g \geq 3$ as illustrated in Figure S5. For completeness, we show the analogous result in SIB in Figure S6.

***The first block is longer than the following ones:*** The system does not follow a stationary renewal process. Nevertheless, it turns out that the large difference we see between the first and the remaining blocks is not so much due to the memory from one block to the next but to the nonexponential distribution of block lengths. Even in the presence of memory from block to block, one has the following general relation on a semi-infinite chromosome at large $g$,

$$\left\langle \ell_1 \right\rangle = \frac{\left\langle \ell_\infty^2 \right\rangle}{2\langle \ell_\infty \rangle}, \tag{5}$$

where $\ell_1$ denotes the length of the first block and $\ell_\infty$ denotes that of faraway blocks. The proof boils down to considering the blocks on the infinite line and taking the origin of the (semi-infinite) chromosome at random. It falls inside a block of length $\ell_\infty$ with probability density proportional to $\ell_\infty$ itself. Denoting as before by $\mu^{(1)}(\ell_1)$ the probability density of the length of the first block, we have

$$\mu^{(1)}(\ell_1) = \frac{\int_{\ell_1}^{\infty} \mu^*(\ell_\infty) d\ell_\infty}{\langle \ell_\infty \rangle}. \tag{6}$$

(The reader can check that this is a normalized probability density.) Using this density, the computation of the first moment of $\ell_1$ leads directly to Equation 5. To interpret this result, note that Equation 5 implies that the relative difference $(\langle \ell_1 \rangle - \langle \ell_\infty \rangle)/\langle \ell_\infty \rangle$ is equal to the [relative variance of $\mu^*(\ell_\infty) - 1$]/2. When $\mu^*(\ell_\infty)$ is a pure exponential, this quantity vanishes and then $\langle \ell_1 \rangle = \langle \ell_\infty \rangle$. Thus we have $\langle \ell_1 \rangle > \langle \ell_\infty \rangle$ if and only if the relative variance of $\mu^*(\ell_\infty) > 1$, which is what we find to happen in this system. For instance in the SSD case, we have $\langle \ell_1 \rangle = 0.595$, to be compared with $\langle \ell_\infty \rangle = \frac{1}{2}$. The first block is thus on average substantially larger than the others.

From Figure 2 one can see that $\mu^{(1)}$ is more spread out than $\mu^*$; $\mu^*$ gives $\mu^{(1)}$ from Equation 6 so that $\mu^{(1)}(0) = 2$ in SSD and 4 in SIB by direct computation using $\langle \ell_\infty \rangle$. Note that $\mu^{(1)}(0)dx$ can also be interpreted as the probability of having the first block end between $x = 0$ and $x = dx$; since that is the same as the density of junctions times $dx$, *i.e.*, 2 or $4 \times dx$, we indeed recover the result $\mu^{(1)}(0) = 2$ for SSD and $\mu^{(1)}(0) = 4$ for SIB.

***The lengths of successive blocks are slightly correlated:*** Even though junctions are independent, each junction affects the IBD property in its neighborhood. Two positions will have nearly independent IBD only when they are distant along the chromosome because only in that case will there be many crossover events separating them. It thus seems natural to expect that the successive block lengths will not be independent in contrast to the underlying $\Delta x$ separating junctions. In fact this must be the case given that we found earlier that on a semi-infinite chromosome the distribution of lengths is different for the second and the third block.

Our framework allows one to compute joint distributions and thus the linear correlation coefficients

$$C(\ell_n, \ell_{n+1}) = \frac{\langle \ell_n \ell_{n+1} \rangle - \langle \ell_n \rangle \langle \ell_{n+1} \rangle}{\sigma_{\ell_n} \sigma_{\ell_{n+1}}}, \tag{7}$$

where $\sigma_{\ell_n}$ (resp. $\sigma_{\ell_{n+1}}$) is the standard deviation of $\ell_n$ (resp. $\ell_{n+1}$). The joint distribution of $\ell_n$ and $\ell_{n+1}$ is given by

$$\mu^{(n,n+1)}(\ell_n, \ell_{n+1}) = \sum_{k_n=1}^{\infty} \sum_{k_{n+1}=1}^{\infty} P^{(n,n+1)}(k_n, k_{n+1}) \rho_{k_n}(\ell_n) \rho_{k_{n+1}}(\ell_{n+1}),$$

where $P^{(n,n+1)}(k_n, k_{n+1})$ is the probability that the $n$th block ends after $k_n$ hops and the $n + 1$th after $k_{n+1}$. From this distribution, the mean of the product $\ell_n \ell_{n+1}$ is the sum of the probabilities $P^{(n,n+1)}(k_n, k_{n+1})$ times the average of $\ell_n$ times the average of $\ell_{n+1}$ (factorization), each of which is obtained from Equation 2. The linear correlation coefficient then reduces to

$$C(\ell_n, \ell_{n+1}) = \frac{\langle k_n k_{n+1} \rangle - \langle k_n \rangle \langle k_{n+1} \rangle}{\left[ \left( \sigma_{k_n}^2 + \langle k_n \rangle \right) \left( \sigma_{k_{n+1}}^2 + \langle k_{n+1} \rangle \right) \right]^{1/2}}, \tag{8}$$

where $\sigma_{k_n}$ (resp. $\sigma_{k_{n+1}}$) is the standard deviation of $k_n$ (resp. $k_{n+1}$). These quantities are directly obtainable from the probability $P^{(n,n+1)}(k_n, k_{n+1})$, as long as the number of generations is not too large. We have computed these quantities in SSD for a semi-infinite chromosome, for different $g$'s and choices of block numbers. For instance, if we consider the first and second blocks, assumed to be fixed, the value of $C(\ell_1, \ell_2)$ is –0.0197 at $g = 2$, –0.0125 at $g = 3$, –0.00841 at $g = 4, \ldots$, with a trend that is compatible with a vanishing limit at large $g$. We find the same trend for the following blocks too. Furthermore, at given $g$, $C(\ell_n, \ell_{n+1})$ rapidly converges to a limiting value as $n$ increases. This is illustrated in Table S1. The computer programs for obtaining these correlation coefficients are also provided (see File S2).

### *Case of finite chromosomes*

***Length distributions:*** Clearly on a finite chromosome of length $L$ the distributions of block lengths are modified. The

computations are more complicated, but remain feasible. As an illustration, consider the case where the chromosome has just two blocks. We use the $P^{(1,2)}(k_1, k_2)$ probabilities, and for each $(k_1, k_2)$ we impose the filter that $\ell_1 < L$ while $\ell_1 + \ell_2 > L$. Then we have for the probability densities $\mu^{(1)}(\ell_1) = \mu^{(1)}(\ell_2)$ and thus the distribution must be symmetric about $L/2$, and we also have $\langle \ell_1 \rangle = \langle \ell_2 \rangle = L/2$. The explicit expression for this density is

$$\mu^{(1)}(\ell_1) = \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} P^{(1,2)}(k_1, k_2) \rho_{k_1}(\ell_1) \int_{L-\ell_1}^{\infty} \rho_{k_2}(\ell_2) \quad . \quad (9)$$

From this it turns out that $\mu^{(1)}(\ell_1)$ has a minimum at $\ell_1 = L/2$: it is more likely to have one rather short and one rather long block than to have two blocks of approximately the same size.

***A comparison with experimental RIL data:*** It is appropriate to compare our theoretical computations with block statistics measured in experimental RILs. Since the block sizes are random variables, it is best to work with RIL data sets where (i) the block structure has been determined precisely, requiring high-density genotyping, and (ii) there are many lines, an easier task for SSD than for SIB crosses. Such a data set has been produced within the species *Arabidopsis thaliana* by Singer *et al.* (2006). These authors genotyped several hundred thousand loci via hybridization arrays, from which they determined block extremities in 100 SSD recombinant inbred lines derived from the crossing of Columbia and Landsberg homozygotes. Singer *et al.* provide the genetic map of their cross and the physical positions of the block extremities. From these data we determined the block lengths for each RIL and each of the five chromosomes. We display in Figure S7 the distributions of the first block length, for all five chromosomes. The solid line is the theoretical curve, corresponding to the infinite chromosome case but truncated to the genetic length of each chromosome. As was explained previously, if the (infinite chromosome) random variable $\ell_1$ is larger than the length $L$ of the chromosome, one sees in practice a block of length $\ell_1 = L$; this happens with a finite probability that is represented in Figure S7, using a solid dot. The histograms are for the experimental data, and we have included the 95% confidence intervals for each bin. We see that the theoretical predictions agree well with the experimental values except for the last bin of chromosome 1 and chromosome 4. Interestingly, for both of these the segregation data exhibit significant distortion; such distortion, typically caused by loci under selection pressures, can affect recombination rates. It is thus satisfying that the agreement between theory and experiment is as good as it is in spite of this distortion.

***Number of blocks:*** Clearly the typical number of blocks will grow with the total length $L$ of the chromosome. Furthermore, for large $g$, the density of block extremities is 2 in SSD (4 in SIB); thus at large $L$ the mean number of blocks should grow as $2L$ in SSD (as $4L$ in SIB). It is also of interest to determine the *distribution* of the number of blocks.

Consider first the probability that the whole chromosome at generation $g$ is in one single homozygous block. Starting at the left end of the chromosome, we must be in a fixed state: we choose it to be, for instance, the $P_A$ genotype. This constraint is used to set the occupation probabilities of the random walks on $\mathcal{H}$ before the first hop. Explicitly, at $x = 0$ we introduce $V_i^{(0)} = 0$ on vertex $i$ if its color is incompatible with the $P_A$ genotype; otherwise $V_i^{(0)}$ is a site-independent constant such that the total probabilities sum to 1. At each junction (hop), the master equation is used to update the vector of probabilities on the hypercube, and so for $K$ hops we have the vector $\{V_i^{(K)}\}_{i=1,\ldots,2^{N_c}}$. We iterate for up to a given total of $N_J$ junctions. In practice we cannot take $N_J = \infty$ because of the numerical nature of the algorithm; so instead we take $N_J$ sufficiently large so that only negligible probabilities are dropped in the truncation. As a ballpark estimate, $N_J$ must be large enough to have $N_J/N_c \gg L$, and then each gamete can have many junctions per morgan. During the application of the master equation to the vector, hops terminating the block are stored and we keep in a file the probabilities $\Pi^{(1)}(K)$ that the first block ends after $K$ hops, $1 \leq K \leq N_J$. Then the probability $p_1$ that there will be just one block in $0 \leq x \leq L$ is given by

$$p_1 = \sum_{K=1}^{\infty} \Pi^{(1)}(K) \int_{L}^{\infty} \rho_K(x) dx \quad (10)$$

times the probability of fixing at $x = 0$ (in SSD this is $1 - 1/2^g$; in SIB it is given by a recurrence relation). Note that these integrals correspond to incomplete Gamma functions, allowing for a relatively efficient computation (see File S2). As mentioned before, in practice the sum over $K$ is truncated to $K \leq N_J$ and one must check that the error induced by this truncation is small enough. We find that the probability to have a single block decreases exponentially when $L$ grows. For instance in SSD, at large $g$, the probability is 0.417 for $L = 0.5$ M, 0.189 for $L = 1.0$ M, and 0.088 for $L = 1.5$ M.

The previous approach can be extended to compute the probability that the chromosome consists of $n$ blocks as follows. For simplicity, consider directly the large $g$ limit so we can ignore hops on $\mathcal{H}$ leading to heterozygous genotypes at $g$. We use the master equation framework to keep track of the joint probabilities of being on a vertex of $\mathcal{H}$ and of being in the $m$th block, for the number of hops $K$ going from 0 to $N_J$. At each hop, there are transitions from vertex to vertex and potentially from block to block. If one hops to a vertex incompatible with the desired block pattern, the probability is set to 0. We keep track of the probabilities $\Pi^{(n)}(K)$ that the $n$th block terminates at the $K$th hop. The probability of having *at most* $n$ blocks when $0 \leq x \leq L$ is then given by the same formula as for one block (Equation 10) but replacing $\Pi^{(1)}(K)$ by $\Pi^{(n)}(K)$. (Note that $K$ is the sum of the number of hops for blocks $1, 2, \ldots, n$.) Repeating the computation for $n - 1$, we obtain the probability of having at most $n - 1$

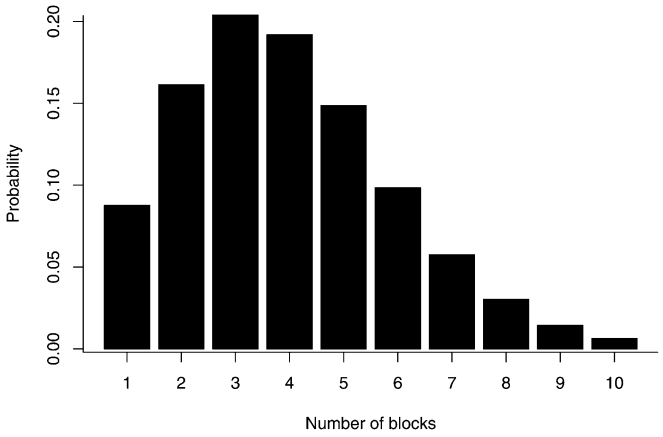



**Figure 3** Distribution of block number in SSD RIL. The histogram shows the frequencies of having 1, 2, . . . blocks in SSD at large number of generations; here $L$ = 1.5 M.

blocks in the region $0 \leq x \leq L$, and taking the difference of the two results we obtain the probability of having exactly $n$ blocks (see File S2). As an illustrative example, Figure 3 gives the probability distribution of $n$ at large $g$ in the case of SSD for a chromosome of length $L = 1.5$ M.

## Discussion

In the dense marker or continuous chromosome picture, genotypes appearing in a RIL form IBD blocks. The block structure is nontrivial in part because the consequence of a crossover at generation $g$ depends on crossovers arising at previous generations. To study the statistics of such blocks, we used a labeling procedure that allows for a mapping onto a Markov process. Such a process reduces our problem of blocks to linear operations (associated with a master equation that we implemented in a C code), followed by relatively standard analysis involving sums and integrals (that we treated via Mathematica). The sources for these computations are provided with this work (see File S2).

We illustrated a number of properties of block statistics, highlighting in particular several closed-form results. Although the joint statistics of blocks are quite complex, we found that block-to-block correlations were very weak and thus genotype frequencies are well approximated by a stationary renewal process. In addition, the distributions of block lengths are not too far from exponential. Thus approximating the RIL case using exponential distributions will lead to qualitatively correct results with an accuracy of ~20%. This level of accuracy should hold for other systems of crosses such as randomly mating populations, justifying the use of the exponential approximation in several previous studies (Stam 1980; Chapman and Thompson 2003).

We mainly stressed cases with complete fixation because the construction of RILs aims to have homozygous genotypes, but our formalism is applicable to both homozygous and heterozygous blocks. Note that within SSD RILs, as shown in *Results*, heterozygous blocks have an exponential distribution for their lengths; furthermore, heterozygous blocks *interrupt* the memory of the process in SSD; that is, two blocks separated by a heterozygous segment are independent. The reason is that there is only one way to be heterozygous in SSD (up to irrelevant exchanges of gametes at the same generation).

Clearly it is possible to extend our formalism to cases involving more than two parents; this may be of use when dealing with multiparental RILs that are being developed currently to have greater power in association studies (Churchill *et al.* 2004). We hope our results will stimulate work in this direction.

## Acknowledgments

We thank F. Rodolphe for discussing his results with us. This work has been supported by grants from the Agence Nationale de la Recherche: ANR-09-GENM-022-003 Single-Meiosis (to O.C.M.) and ANR-O6-BLANC-0128 (to F.H.).

## Literature Cited

Alcazar, R., A. V. Garcia, J. E. Parker, and M. Reymond, 2009 Incremental steps toward incompatibility revealed by Arabidopsis epistatic interactions modulating salicylic acid pathway activation. Proc. Natl. Acad. Sci. USA 106: 334–339.

Altshuler, D., L. D. Brooks, A. Chakravarti, F. S. Collins, M. J. Daly *et al.*, 2005 A haplotype map of the human genome. Nature 437: 1299–1320.

Baird, S., N. Barton, and A. Etheridge, 2003 The distribution of surviving blocks of an ancestral genome. Theor. Popul. Biol. 64: 451–471.

Ball, F., and V. Stefanov, 2005 Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. Math. Biosci. 196: 215–225.

Bennett, J., 1953 Junctions in inbreeding. Genetica 26: 392–406.

Bergland, A. O., A. Genissel, S. V. Nuzhdin, and M. Tatar, 2008 Quantitative trait loci affecting phenotypic plasticity and the allometric relationship of ovariole number and thorax length in *Drosophila melanogaster*. Genetics 180: 567–582.

Bickeboller, H., and E. Thompson, 1996a Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theor. Popul. Biol. 50: 66–90.

Bickeboller, H., and E. Thompson, 1996b The probability distribution of the amount of an individual's genome surviving to the following generation. Genetics 143: 1043–1049.

Browning, S., and B. Browning, 2002 On reducing the statespace of hidden Markov models for the identity by descent process. Theor. Popul. Biol. 62: 1–8.

Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA 107: 786–791.

Buckler, E., and M. Gore, 2007 An Arabidopsis haplotype map takes root. Nat. Genet. 39: 1056–1057.

Cannings, C., 2003 The identity by descent process along the chromosome. Hum. Hered. 56: 126–130.

Cardon, L. R., and G. R. Abecasis, 2003 Using haplotype blocks to map human complex trait loci. Trends Genet. 19: 135–140.

Carlton, J. M., 2007 Toward a malaria haplotype map. Nat. Genet. 39: 5–6.

Chapman, N. H., and E. A. Thompson, 2003 A model for the length of tracts of identity by descent in finite random mating populations. Theor. Popul. Biol. 64: 141–150.
Churchill, G. A., 2007 Recombinant inbred strain panels: a tool for systems genetics. Physiol. Genomics 31: 174–175.
Churchill, G., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The collaborative cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 36: 1133–1137.
Crow, J. F., 2007 Haldane, Bailey, Taylor and recombinant-inbred lines. Genetics 176: 729–732.
Cuppen, E., 2005 Haplotype-based genetics in mice and rats. Trends Genet. 21: 318–322.
Curtis, D., A. E. Vine, and J. Knight, 2008 Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann. Hum. Genet. 72: 261–278.
Dimitropoulou, P., and C. Cannings, 2003 Recsim and indstats: probabilities of identity in general genealogies. Bioinformatics 19: 790–791.
Donelly, K. P., 1983 The probability that related individuals share some section of genome identical by descent. Theor. Popul. Biol. 23: 34–63.
Feller, W., 1950 *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, New York.
Fisher, R., 1949 *The Theory of Inbreeding*. Oliver & Boyd, Edinburgh.
Fisher, R., 1954 A fuller theory of "junctions" in inbreeding. Heredity 8: 187–197.
Fisher, R., 1959 An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. Heredity 13: 523–542.
Franklin, I., 1977 The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theor. Popul. Biol. 11: 60–80.
Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
Gold, B., T. Kirchhoff, S. Stefanov, J. Lautenberger, A. Viale *et al.*, 2008 Genome-wide association study provides evidence for a breast cancer risk locus at 6q22–33. Proc. Natl. Acad. Sci. USA 105: 4340–4345.
Goldstein, D. B., 2001 Islands of linkage disequilibrium. Nat. Genet. 29: 109–111.
Greenawalt, D. M., X. F. Cui, Y. J. Wu, Y. Lin, H. Y. Wang *et al.*, 2006 Strong correlation between meiotic crossovers and haplotype structure in a 2.5-mb region on the long arm of chromosome 21. Genome Res. 16: 208–214.
Grote, M. N., 2007 A covariance structure model for the admixture of binary genetic variation. Genetics 176: 2405–2420.
Guo, S., 1994 Computation of identity by descent proportions shared by two siblings. Am. J. Hum. Genet. 54: 1104–1109.
Haldane, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genet. 8: 299–309.
Haldane, J. B. S., and C. H. Waddington, 1931 Inbreeding and linkage. Genetics 16: 357–374.
Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.
Jeffreys, A. J., L. Kauppi, and R. Neumann, 2001 Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. Nat. Genet. 29: 217–222.
Kauppi, L., M. P. H. Stumpf, and A. J. Jeffreys, 2005 Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region. Genomics 86: 13–24.
Keurentjes, J. J. B., J. Y. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken *et al.*, 2007 Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. USA 104: 1708–1713.
Lencz, T., C. Lambert, P. DeRosse, K. E. Burdick, T. V. Morgan *et al.*, 2007 Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc. Natl. Acad. Sci. USA 104: 19942–19947.
Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.
Maccaferri, M., M. C. Sanguineti, S. Corneti, J. L. A. Ortega, M. Ben Salem *et al.*, 2008 Quantitative trait loci for grain yield and adaptation of durum wheat (*Triticum durum* desf.) across a wide range of water availability. Genetics 178: 489–511.
Martin, O. C., and F. Hospital, 2006 Two- and three-locus tests for linkage analysis using recombinant inbred lines. Genetics 173: 451–459.
McVean, G. A. T., and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. B Biol. Sci. 360: 1387–1393.
Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
Moore, R. C., M. Henry, and H. Stevens, 2008 Local patterns of nucleotide polymorphism are highly variable in the selfing species Arabidopsis thaliana. J. Mol. Evol. 66: 116–129.
Mott, R., 2007 A haplotype map for the laboratory mouse. Nat. Genet. 39: 1054–1056.
Pe'er, I., Y. R. Chretien, P. I. W. de Bakker, J. C. Barrett, M. J. Daly *et al.*, 2006 Biases and reconciliation in estimates of linkage disequilibrium in the human genome. Am. J. Hum. Genet. 78: 588–603.
Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
Rodolphe, F., J. Martin, and E. Della-Chiesa, 2008 Theoretical description of chromosome architecture after multiple backcrossing. Theor. Popul. Biol. 73: 289–299.
Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
Shifman, S., J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi, 2003 Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. 12: 771–776.
Singer, T., Y. P. Fan, H. S. Chang, T. Zhu, S. P. Hazen *et al.*, 2006 A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet. 2: e144.
Slate, J., and J. M. Pemberton, 2007 Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. J. Evol. Biol. 20: 1415–1427.
Slatkin, M., 1972 On treating the chromosome as the unit of selection. Genetics 72: 157–168.
Stam, P., 1980 The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35: 131–155.
Stefanov, V. T., 2000 Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. Genetics 156: 1403–1410.
Stumpf, M. P. H., 2002 Haplotype diversity and the block structure of linkage disequilibrium. Trends Genet. 18: 226–228.
Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 5: e071.
Tishkoff, S. A., and B. C. Verrelli, 2003 Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. Curr. Opin. Genet. Dev. 13: 569–575.

Wall, J. D., and J. K. Pritchard, 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. 4: 587–597.

Walters, K., and C. Cannings, 2005 The probability density of the total IBD length over a single autosome in unilineal relationships. Theor. Popul. Biol. 68: 55–63.

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. 55: 248–259.

Wolfram, S., 1991 *Mathematica: A System for Doing Mathematics by Computer*, Ed. 2. Addison Wesley Longman Publishing, Redwood City, CA.

Yalcin, B., J. Fullerton, S. Miller, D. A. Keays, S. Brady *et al.*, 2004 Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. Proc. Natl. Acad. Sci. USA 101: 9734–9739.

Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539–551.

Zhang, C., D. K. Bailey, T. Awad, G. Y. Liu, G. L. Xing *et al.*, 2006 A whole genome long-range haplotype (wglrh) test for detecting imprints of positive selection in human populations. Bioinformatics 22: 2122–2128.

Zheng, M. X., and M. S. McPeek, 2007 Multipoint linkage-disequilibrium mapping with haplotype-block structure. Am. J. Hum. Genet. 80: 112–125.

Zondervan, K. T., and L. R. Cardon, 2004 The complex interplay among factors that influence allelic association. Nat. Rev. Genet. 5: 89–100.

*Communicating editor: I. Hoeschele*

GENETICS

**Supporting Information**
http://www.genetics.org/content/suppl/2011/08/12/genetics.111.129700.DC1

# Distribution of Parental Genome Blocks in Recombinant Inbred Lines

**Olivier C. Martin and Frédéric Hospital**


DOI: 10.1534/genetics.111.129700

File S1

Supporting Text

# Methods

## General aspects

Let $P_A$ and $P_a$ be the parents of the original $F_1$ hybrid (cf. Fig. 1, main text). Let $N_c$ be the total number of gametes produced. The probability of staying heterozygous at a given locus can be computed without much difficulty. In SSD, one has a probability $1/2$ of maintaining heterozygousy from one generation to the next, so the probability of remaining heterozygous after $g$ crossings is $2^{-g}$. In SIB, fixation is reached more slowly as should be clear from the fact that there are four chromosomes rather than only two. To determine the dependence with $g$, one can use a recursion for the probabilities to have 0, 1, 2, 3, or 4 copies of the parental $P_A$ allele. At large $g$, this recursion shows that the probability of being heterogyzgous at a locus decays as $0.8357^g$; in practice this means that about $\ln(1/2)/\ln(0.8357) = 3.86$ times as many generations are required in SIB than in SSD to reach the same level of homozygosity. A similar study could be performed using two loci, but that would not tell us anything about the blocks we are interested in; instead, it is necessary to tackle the continuous chromosome framework straight-on.

The objects of study are the IBD (identical by descent) blocks, which we refer to as ``blocks''; one can think of the two parents $P_A$ and $P_a$ as having different alleles at all loci; that way IBD blocks are also identical in state and can be determined directly from the allelic state of the loci, assuming no mutations occurred in the RIL construction process. However, this is just for convenience, and our work does not require that all loci be polymorphic in the parents. The blocks that we describe refer in all cases to chromosomal stretches that are identical by descent to the parents of the $F_1$ hybrid.

Our goal is to understand how genomic blocks are organized in SSD and SIB recombinant inbred lines. The statistics of such blocks will depend on the number $g$ of generations used to construct the RIL. However, since all loci eventually get fixed after sufficient generations, there will be a limiting probability law as $g$ increases. In the case of an infinite chromosome, the mean size of blocks and the probability density of very short blocks can be derived as shown in the main text. For other statistical properties of blocks, in particular for finite chromosomes, the problem is more complex. Fortunately, it can be mapped to a random walk process on a hypercube, in direct analogy with the case of other studies on various pedigrees [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The enumeration of all walks can be treated by computer; the C source code we have developed for that is provided in the Supplemental Information (see File S2). Furthermore, the continuous waiting time variables of the random walks can be treated by standard mathematical analysis; we have performed these calculations using Mathematica [13]) and the associated Mathematica notebooks are provided in the Supplemental Information (see File S2). Our mathematical and computational approach builds on the work of Donelly [1], Stefanov [6, 7] as well as Rodolphe et al. [12]. We now give some details about the mapping, the algorithmic approach for enumeration, and the application to block statistics.

## Junctions and labelings

Each (haploid) gamete or homologous chromosome is produced from a diploid parent; we refer to a gamete/chromosome as being ``paternal'' or ``maternal'' depending on whether it has been produced via a male or female meiosis. This situation is shown for SSD in Fig. 1 (main text) and also in Fig. S1 using a slightly different viewpoint, with the use of the $H$ and $H'$ labels for the paternal and maternal origins. In animals, the sex of the parent determines the nature of the gametes, while in most plants it is the organ. Since for our purposes a diploid individual is the union of two such gametes, we shall say for the chromosome pair of interest that there is one chromosome that is paternal (coming from the male meiosis) and one that is maternal (coming from the female meiosis).

Given a gamete, produced by a parent $P$, we need to determine its composition in $P_A$ and $P_a$ genomes. The conceptually simplest way to do this would be through the allelic content at each position along the gamete; unfortunately such an approach leads to complex dependencies which we have found to be untractable. Instead, we follow [6, 12] and specify which genomic

regions come from the paternal chromosome of $P$ and which come from the maternal chromosome of $P$, regardless of the regions' genotypes; this choice will allow a mapping to a Markov process and ultimately to a manageable framework for determining all quantities of interest. To each point at position $x$ on the gamete, we assign ``0'' when genetic material comes from the paternal chromosome of $P$, and ``1'' when it comes from the maternal chromosome. A chromosome then consists of regions labeled by 0's and 1's, separated by crossing-over (junction) events.

Our labeling scheme is illustrated for the SSD case in Fig. 1 (main text) and in Fig. S1; note that the parental line $P_A$ (respectively $P_a$) contributes entirely to the paternal (respectively maternal) chromosome of the hybrid $F_1$ at generation $g = 0$. From these labelings, we can determine the identity by descent for any position. For instance, in the figure, the individual at generation $g = 1$ has a ``maternal'' chromosome which has contributions from both $P_A$ and $P_a$ (the crossover points are represented by small vertical segments) and clearly all the loci outside of the region between junctions $J1$ and $J4$ are fixed already at this stage (in fact they are IBD to the parent $P_a$). The SSD offspring at $g = 2$ further fixes some more loci, and we see that the regions adjacent to $J4$ are now fixed and that this $J4$ junction delimits two blocks. On the other hand, $J1$ has not been passed on to $g = 2$ and so will not delimit two blocks, in fact the region around that point is fixed with alleles from $P_a$.

In SIB matings, one produces a male and a female individual at each generation so there are now four gametes to follow from one generation to the next. Our labeling scheme is the same as in SSD: each chromosome (gamete) is labeled with a binary ``0-1'' coding at position $x$ depending on whether the associated genetic material comes from the paternal or maternal chromosome of its parent. Note that now one gamete comes from one parent, and the other gamete from the other parent, so these paternal and maternal ``origins'' correspond in SIB to the sex (male or female) of the grand-parent from which the genetic material came. An illustrative example of the resulting mosaic structures is given in Fig. S2.

Note: were one to use allelic content rather than the ``0'' and ``1'' labelings, the introduction of a new junction at generation $g$ would change the alleles of all generations beyond that and the dynamics would become untractable. Our ``0'' and ``1'' labelings only refer to the material at the previous generation, and not directly to the original parents $P_A$ and $P_a$. If one wants to know the final allelic state at a locus in the last generation, one has to follow the series of relative heredity events ``0'' and ``1'' at that locus, back to the original parent. However, it is because the ``0'' and ``1'' labelling is only relative to the previous generation, and hence has no memory, that the Markov machinery can be applied.

Given all the labelings, we want to examine the blocks arising at generation $g$. It should be clear that a junction does not necessarily signal the extremity of a block: for example, the region around a junction can very well not be passed on to offspring, as was illustrated in Fig. S1 for junction $J1$. But if we know the ``0'' and ``1'' labels of all $N_c$ gametes at a given point, then we can reconstruct the IBD origin ($P_A$ or $P_a$) of the alleles at that point; in fact this map $\mathcal{M}$ from $\{0,1\}^{N_c}$ to genotypes is $x$ independent and corresponds to a coloring of the vertices of the $N_c$-dimensional hypercube. There are as many colors as there are one-locus genotypes at a given generation: 4 for SSD and 16 for SIB. (This number distinguishes $Aa$ from $aA$, so that there is no phase ambiguity.)

## The master equation to treat the random walks

Each random walk will lead to a sequence of colors on $\mathcal{H}$ and thus will be compatible or not with a desired block pattern. For instance, one might ask for a first block fixed of the $P_a$ type, followed by a heterozygous block, followed by another block of the $P_a$ type, etc. To keep the explanations as simple as possible, we shall often restrict the cases to fixed (homozygous) blocks but the Markov approach is applicable in complete generality to homozygous as well as heterozygous blocks. We want to go over all walks compatible with the desired block pattern, taking into account their weights, and then sum all their weights to get the probability of having the predefined pattern.

Before dealing with the full continuous time Markov process beginning at $x = 0$ and ending at $x = L$, let us ignore the residence times and focus only on the sequence of vertices visited by the walk on the hypercube. We use the Master equation approach [14]: within a computer algorithm, we follow the probability $V_i$ of being on vertex $i$ of $\mathcal{H}$ before each hop. At a given hop (junction), we consider each vertex, divide the probability at that vertex into $N_c$ equal parts, and send one part to each of the neighboring vertices. If the receiving vertex is not compatible with the block pattern requested, the part sent is instead thrown away, i.e., set to 0. These operations define the updating of the probabilities on each vertex after the hop; the implementation

corresponds mathematically to applying a (sparse) matrix to the vector of probabilities $V^{(k)}$ after hop $k$ to obtain the probabilities after hop $(k+1)$:

$$V_i^{(k+1)} = \sum_{\langle ij \rangle} M_{i,j} V_j^{(k)} \tag{S1}$$

The indices $i$ and $j$ label hypercube vertices, and the sum is over the $N_c$ neighbors of vertex $i$. This Master equation, encoded in the matrix $M$, in effect considers all possible random walks on $\mathcal{H}$ in a systematic way. (Without loss of generality, we shall assume that the initial vector of probabilities $V^{(0)}$ is set so as to be compatible with the desired block pattern.) From this deterministic evolution of probabilities, we can extract numerous relevant quantities, the simplest one being the probability $P^{(n)}(k)$ that the $n$'th block lasts exactly $k$ hops. Later, these probabilities will be used to construct the block size distributions. Other quantities of use are the joint probabilities $P^{(n,n+1)}(k_n, k_{n+1})$ that two successive blocks $n$ and $n+1$ terminate after their $k_n$th and $k_{n+1}$th hop; from these we can get joint block size distributions, etc.

## The role of chromosome length

Implicitly the total length $L$ of the chromosome has been ignored in the previous discussion. However, $L$ influences both the lengths and the number of blocks, so the computation of the block sizes given in the previous section has to be modified to take into account $L$. Clearly one must force the last block to contain the value $x = L$; such a constraint can be imposed via a ``filter'' on the events. Suppose for instance one wants to find the probability of having a pattern of 3 blocks when $0 \leq x \leq L$. One can first compute (on the infinite chromosome) the quantities $P^{(1,2,3)}(k_1, k_2, k_3)$ via the Master equation. For given $k_1$, $k_2$ and $k_3$, block $i$ will have a length $\ell_i$ distributed according to $\rho_{k_i}$; the filter will then simply be the constraint $\ell_1 + \ell_2 < L < \ell_1 + \ell_2 + \ell_3$. One must then integrate over all values of the $\ell_i$ compatible with these constraints, and finally sum over all cases of $k_1, k_2, k_3$ along with their probabilities. Here and in most other cases, there are short-cuts to make the computation more efficient because the filter applies only to the last block. In this example, a short-cut consists in applying the Master equation to give the quantities $P^{([1-2],3)}(k_{[1-2]}, k_3)$ associated with the cases where the union of the first and second blocks (both constrained to be non empty) uses $k_{[1-2]}$ hops and the third uses $k_3$ as before. Now instead of having triple integrals, one has double integrals, and the overall probability $P_3$ of having the desired pattern with three blocks becomes

$$\begin{aligned} P_3 = \sum_{k_{[1-2]}=2}^{\infty} \sum_{k_3=1}^{\infty} & P^{([1-2],3)}(k_{[1-2]}, k_3) \\ & \int_{\ell_{[1-2]}=0}^{L} d\ell_{[1-2]} \rho_{k_{[1-2]}}(\ell_{[1-2]}) \\ & \int_{\ell_3 = L - \ell_{[1-2]}}^{\infty} d\ell_3 \rho_{k_3}(\ell_3)\,. \end{aligned} \tag{S2}$$

Such sums and integrals can be treated by mathematical software and we have used Mathematica [13] (see File S2). It turns out that further simplifications are often possible. For instance to determine the probability of having $n$ blocks when $0 \leq x \leq L$, one can first compute the probabilities of having at most $m$ blocks; determining these requires only summing over the one index $k_{[1-\ldots-m]}$ and integrating over the one variable $\ell_{[1-\ldots-m]}$ with the constraint that $\ell_{[1-\ldots-m]} > L$ (the filter). Given these probabilities, quantities such as $P_3$ are obtained by simple differences as is illustrated in the main text.

# Results

## A spectral decomposition for the length distribution

The continuous time stochastic process on the hypercube $\mathcal{H}$ admits a spectral decomposition [15]. This means that the evolution of probabilities on the hypercube can be represented in terms of a diagonal operator (just like the matrix in Eq. S1 can be diagonalized), the spectrum being the corresponding eigenvalues $\{\lambda_j\}_{j=1,2^{N_c}}$. The probability of being in a block of given color can then be

written as

$$\sum_i{}' V_i(x) = \sum_{j=1}^{2^{N_c}} \alpha_j e^{-\lambda_j x} \tag{S3}$$

where the first sum is restricted to vertices $i$ of the hypercube that have the color under consideration, $x$ is the continuous position along the chromosome, and the coefficients $\alpha_j$ depend on the eigenvectors. We then see that the probability of staying within a given color for some length (such as the block size) decomposes into a sum of exponentials. As a consequence, the probability density of block sizes is dominated at large sizes by a single exponential of rate $\lambda^*$, the smallest of the $\lambda_j$; in fact this rate is the same for all blocks. Note that for $g = \infty$, formally one has an infinite sum of exponentials which a priori could change this result, but in reality that is not the case because these terms, coming from larger and larger $g$, have contributions that decay fast with $g$.

A similar analysis can be applied to the values of $P^{(n)}(k)$ (at large $n$) from which the steady-state block size distribution are computed. Indeed, exploiting again the Markov process at any given $n$, these probabilities can be represented by a sum of exponentials, and we find that a single exponential dominates at large $k$.

## A good SRP approximation

Even though junctions are independent, each junction affects the IBD property in its neighborhood; only at large distances (many crossover events) will we get decorrelation as is often the case with Markov processes. As explained in the main text, the successive block lengths are not independent, in contrast to the underlying $\Delta x$ separating junctions. Nevertheless, since we found the correlations between block lengths to be weak, it is of interest to consider the approximation whereby these correlations are simply neglected, leading to a stationary renewal process (SRP) [14] when considering the infinite chromosome. The block lengths are then taken to be i.i.d. random variables of distribution $\mu^*(\ell)$. This distribution was obtained in the main part of the paper. In the absence of a closed form expression for $\mu^*(\ell)$, we have empirically fitted it. For instance in the case of SSD, we find that the function $\mu^*(\ell) \approx 3e^{-A\ell}/(1 + B\ell)$ works quite well. In this parameter-free function, the $3$ comes from $\mu^*(0) = 3$ while the values $A = 0.9789$ and $B = 4.1295$ are set so that the distribution is normalized and $\langle \ell \rangle = 0.5$. Fig. S3 shows $\mu^*(\ell)$ and its approximation which clearly is very satisfactory.

Although such an SRP framework is formulated for an infinite chromosome, it can be applied to finite chromosomes by considering that the interval $[0, L]$ acts as a filter, just as it did when deriving block number distributions.

# References


[1] Donelly KP (1983) The probability that related individuals share some section of genome identical by descent. Theor Pop Biol 23:34--63.

[2] Franklin I (1977) The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theor Pop Biol 11:60--80.

[3] Guo S (1994) Computation of identity by descent proportions shared by two siblings. Am J Hum Genet 54:11041109.

[4] Bickeboller H, Thompson E (1996) Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theor Pop Biol 50:66--90.

[5] Bickeboller H, Thompson E (1996) The probability distribution of the amount of an individual's genome surviving to the following generation. Genetics 143:1043--1049.

[6] Stefanov VT (2000) Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. Genetics 156:1403--1410

[7] Ball F, Stefanov VT (2005) Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. Math Biosci 196:215--225.

[8] Browning S, Browning B (2002) On reducing the statespace of hidden markov models for the identity by descent process. Theor Pop Biol 62:1--8.

[9] Cannings C (2003) The identity by descent process along the chromosome. Hum Hered 56:126--130.

[10] Dimitropoulou P, Cannings C (2003) Recsim and indstats: probabilities of identity in general genealogies. Bioinformatics 19:790--791.

[11] Walters K, Cannings C (2005) The probability density of the total IBD length over a single autosome in unilineal relationships. Theor Pop Biol 68:55--63.

[12] Rodolphe F, Martin J, Della-Chiesa E (2008) Theoretical description of chromosome architecture after multiple back-crossing. Theor Pop Biol 73:289--299.

[13] Wolfram S (1991) Mathematica: a system for doing mathematics by computer (2nd ed.) (Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA).

[14] Feller W (1950) An introduction to probability theory and its applications (John Wiley and sons, New York, NY).

[15] Sharpe M (1988) General theory of Markov processes (Academic Press, San Diego).

[16] Singer T, Fan YP, Chang HS, Zhu T, Hazen SP, Briggs SP (2006) A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. Plos Genetics 2(9):e144.

Table S1 Linear correlation coefficients $C_{n,n+1}$ between successive block lengths (for blocks $n$ and $n+1$) at increasing values of $g$. Note that all coefficients are small, and approach $0$ as $g$ increases.

| $g$ | $C_{1,2}$ | $C_{2,3}$ | $C_{3,4}$ | $C_{4,5}$ |
|---|---|---|---|---|
| 2 | -0.0197 | -0.0334 | -0.0221 | -0.0285 |
| 3 | -0.0125 | -0.0237 | -0.0195 | -0.0205 |
| 4 | -0.00841 | -0.0165 | -0.0141 | -0.0144 |
| 5 | -0.00597 | -0.0115 | -0.00967 | -0.00970 |
| 6 | -0.00446 | -0.00803 | -0.00609 | -0.00618 |

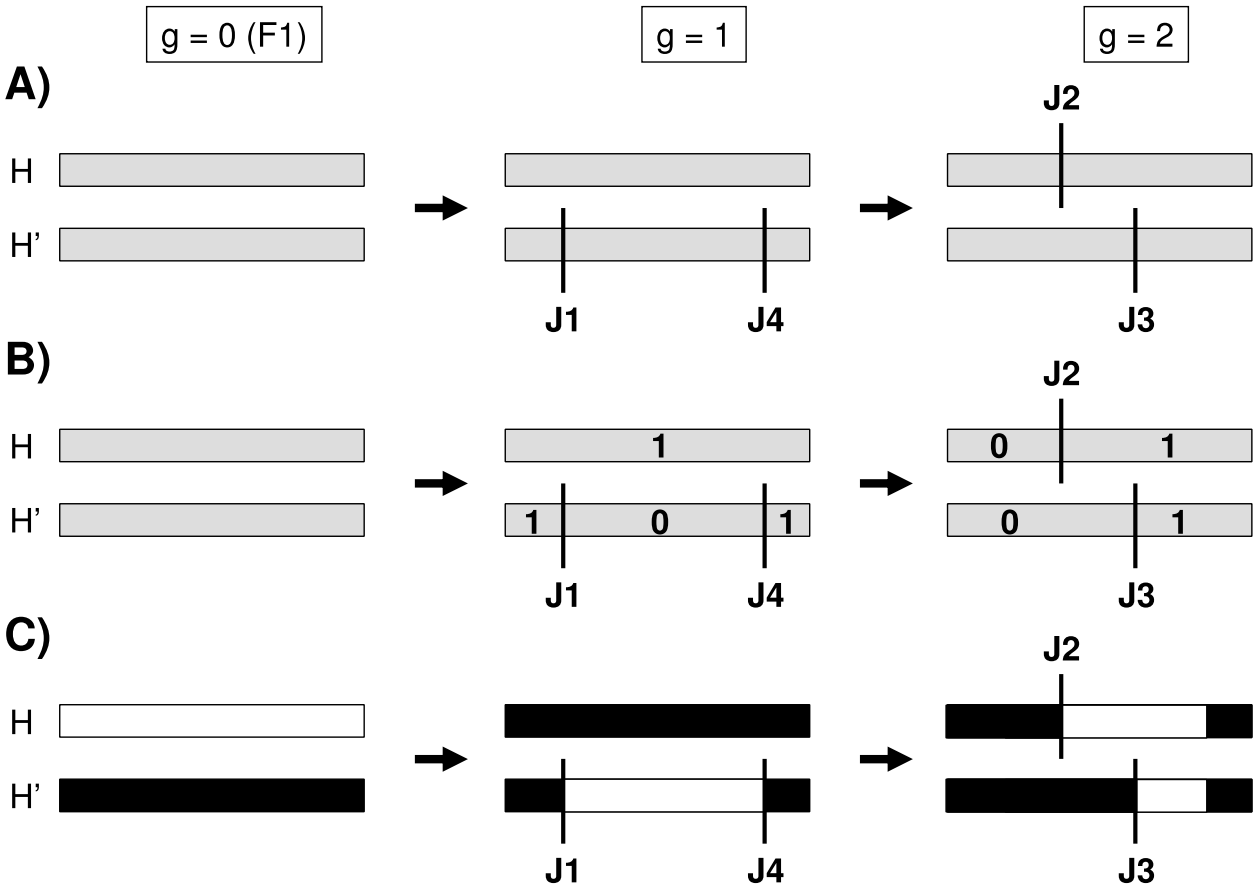



Figure S1: Principle of genotype monitoring using junctions, illustrated in the case of SSD. Only one pair of homologous chromosomes (H, H') is depicted, for three generations ($g = 0$ to 2, from left to right). The process is decomposed into three steps for the sake of clarity: A) at all generations, junctions (vertical bars) are placed randomly on chromosomes. The example shows four junctions, J1 to J4. The junctions are indexed from left to right when all chromosomes/generations are projected onto the same map (see text for details). B) Relative inheritance (IBD) from one generation to the next is described using 0 and 1 labels, with 0 (resp. 1) indicating that the chromosome segment is a copy of the H (resp. H') chromosome of the previous generation. The labels are random at the left side of the chromosomes (corresponding to $x = 0$, then necessarily switch ($0 \to 1$ or $1 \to 0$) at each junction. It is important to note that during steps A) and B) the genotypes (allelic content) of the chromosomes are not considered. This is emphasized by a uniform gray colouring of all chromosomes. C) Genotypes are deduced. Knowing the true genotype of the F1 (assumed here to be fully heterozygous white/black), one can then follow the relative inheritance patterns of step B) from one generation to the next and deduce the genotypes of the chromosomes in the last generation. It is important to keep in mind that there is no direct relationship between the 0/1 labels and the white/black genotypes (except for $g = 1$).

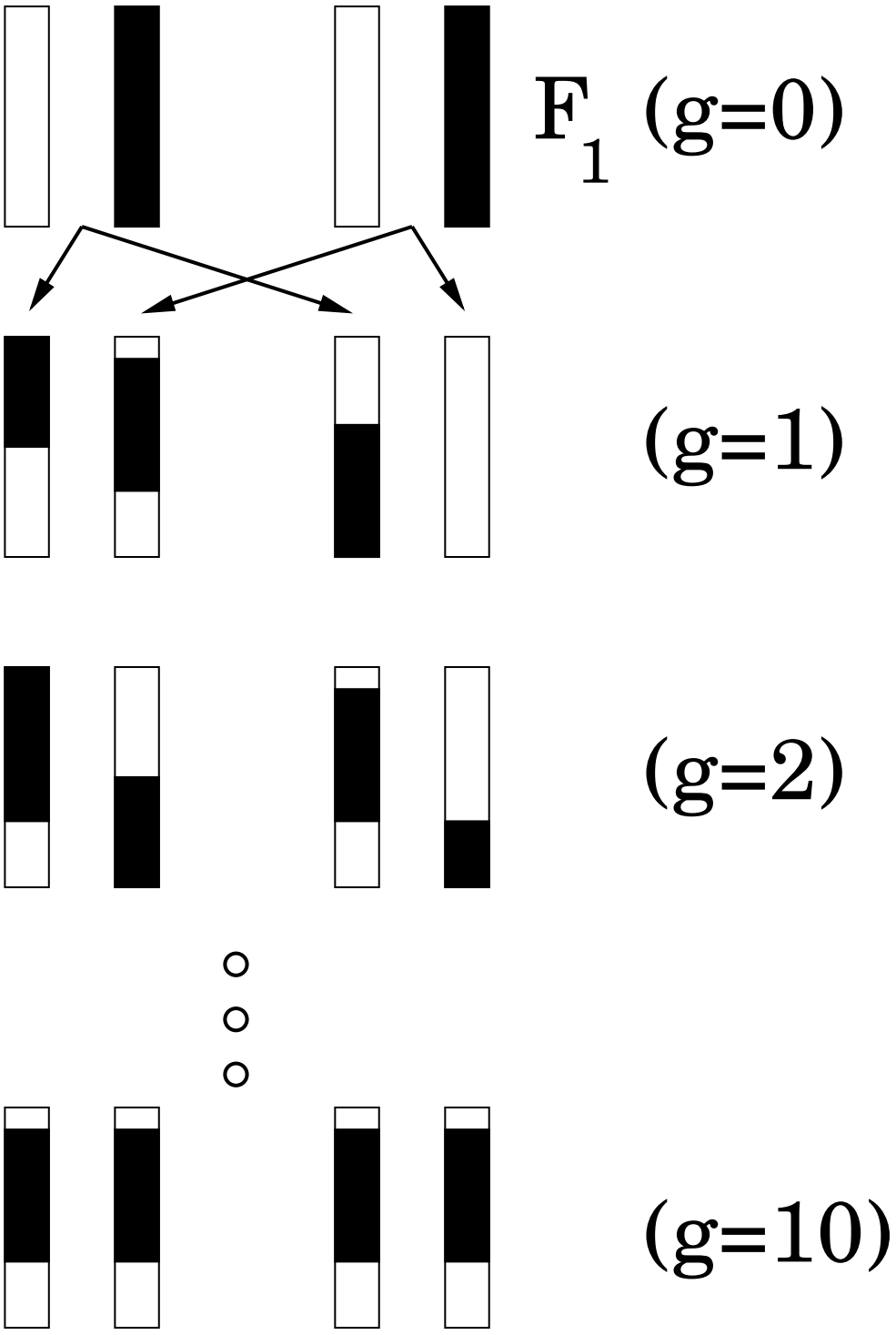



Figure S2: Two parental lines are crossed in SIB to produce a recombinant inbred line. A given chromosome from this line then forms a mosaic of blocks, each IBD with one of the parents. The final block pattern here consists of $3$ alternating fixed blocks.

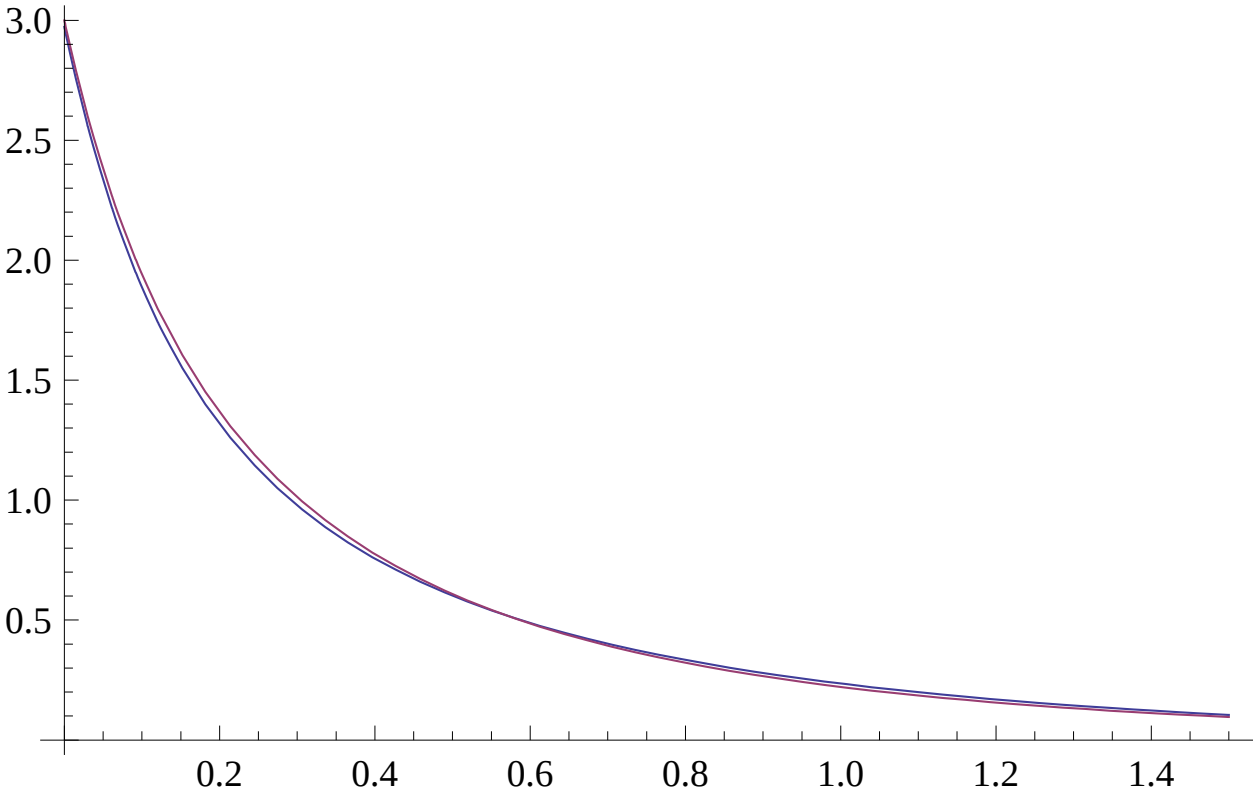



Figure S3: Probability density $\mu^*(\ell)$ of a block far away from the origin and the parameter-free approximation $3e^{-A\ell}/(1+B\ell)$ with $A = 0.9789$ and $B = 4.1295$ so that the distribution is normalized and $\langle \ell \rangle = 0.5$. The number of generations is large, and the RILs use SSD.

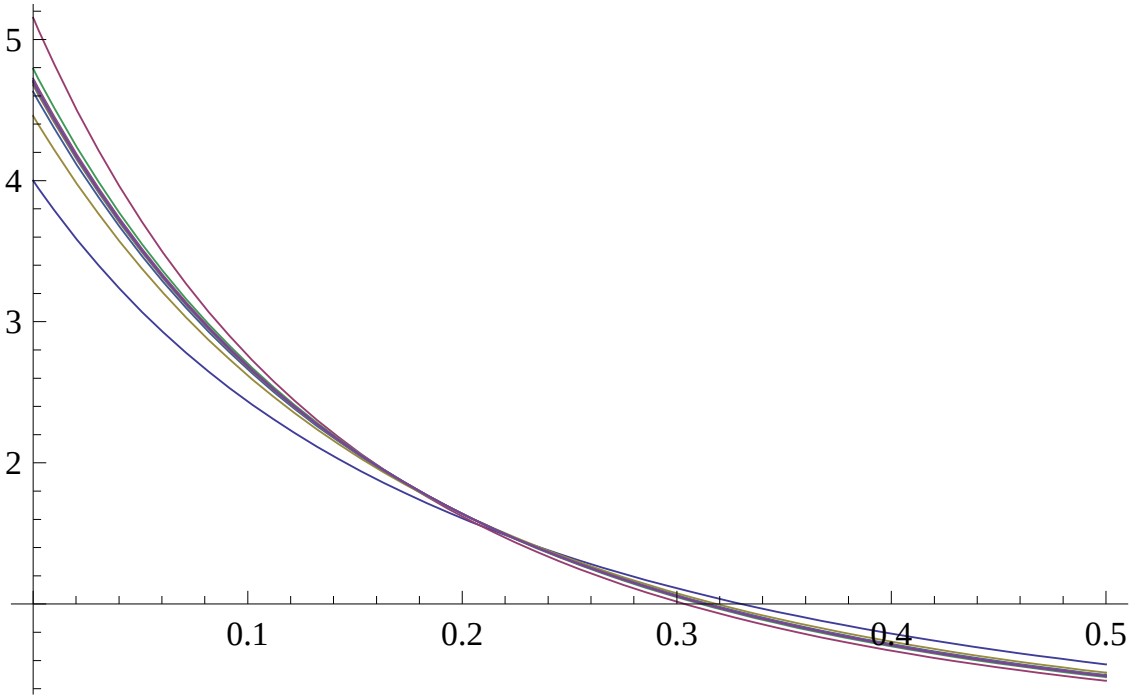



Figure S4: Probability density of homozygous block size for block number 1, 2, ... 10 in SIB. The chromosome is semi-infinite, and the number of generations is 7. Convergence in block number is oscillatory. One sees that the memory effects are much larger than in the SSD RILs.

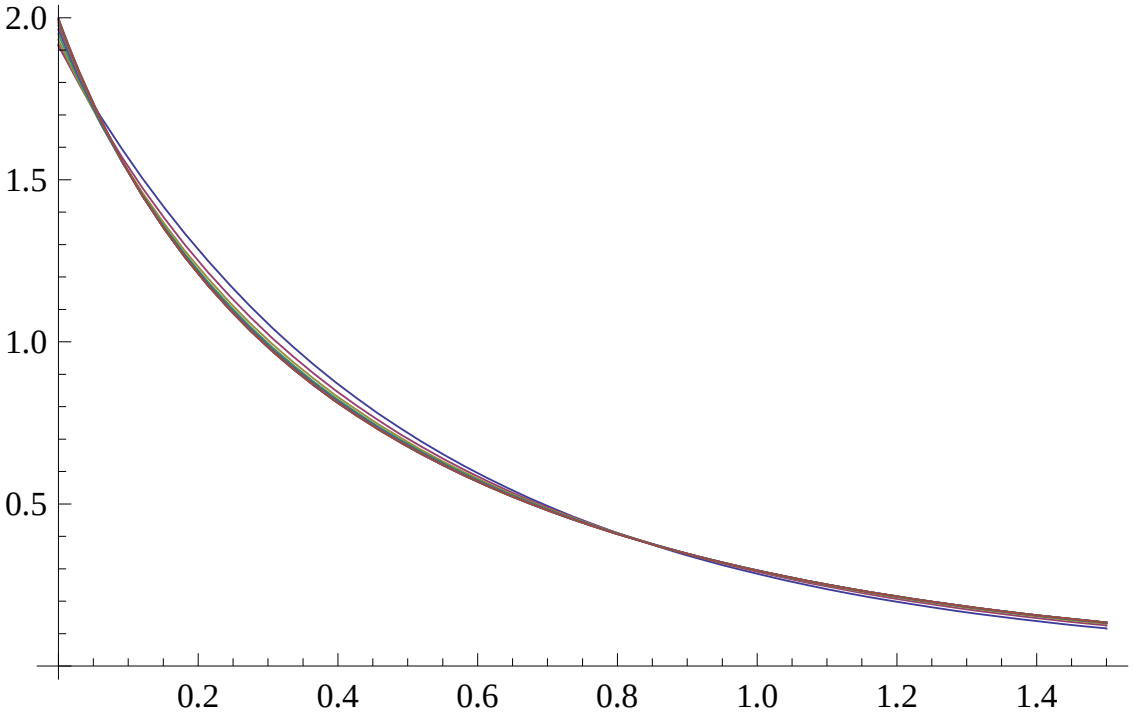



Figure S5: Probability density of the length of the first block on a long chromosome, assuming it is homozygous. Results are for $g = 2$ to $g = 12$ in SSD, exhibiting the rapid convergence of block statistics with $g$.

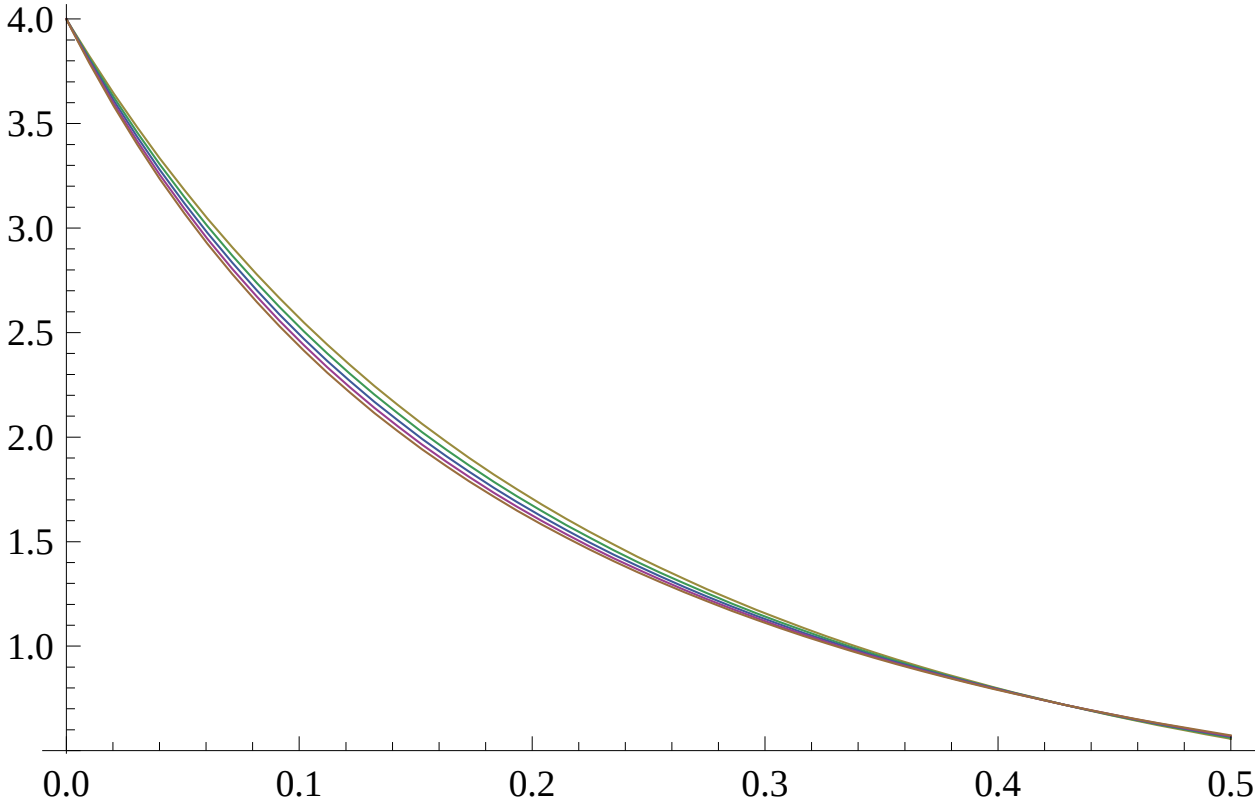
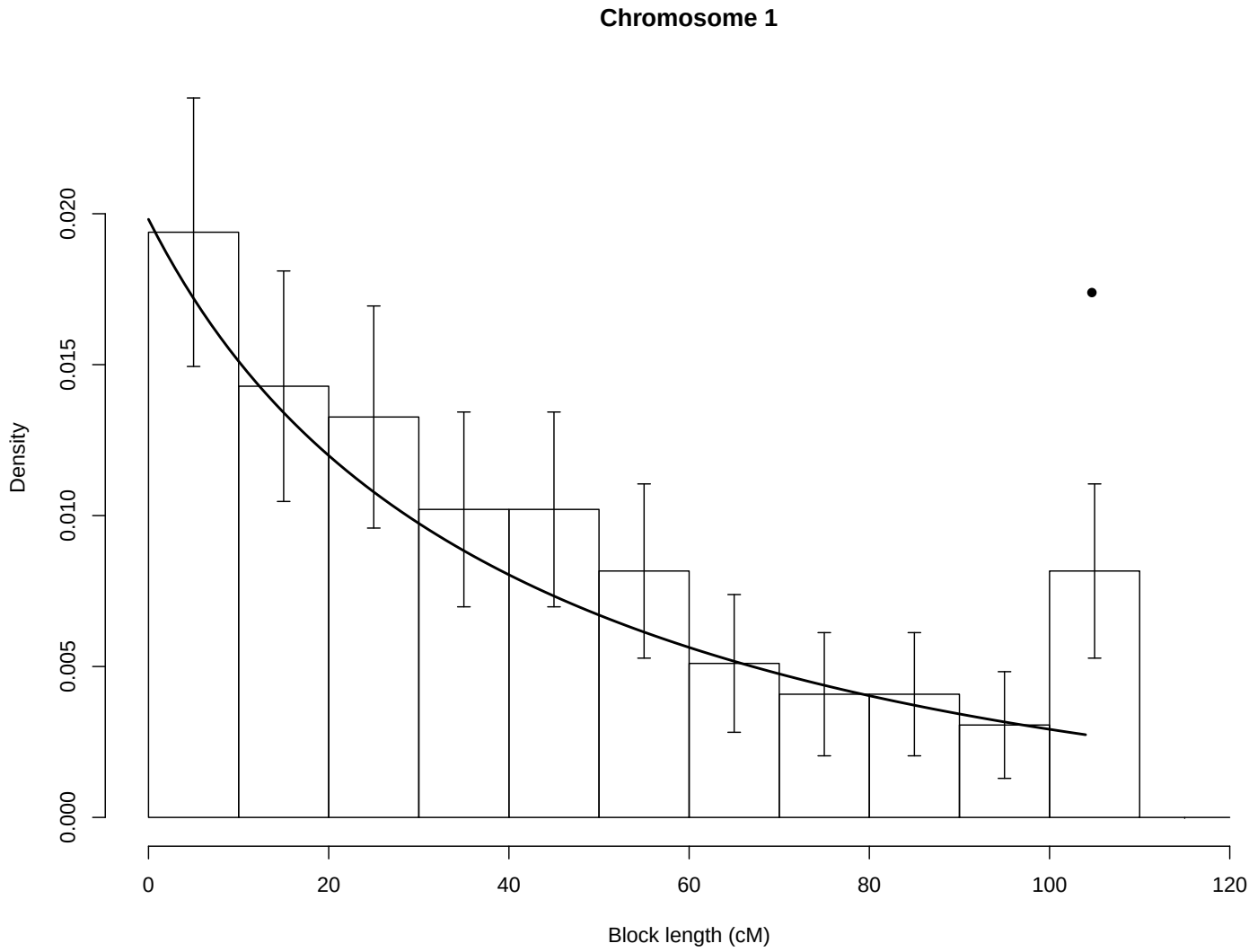
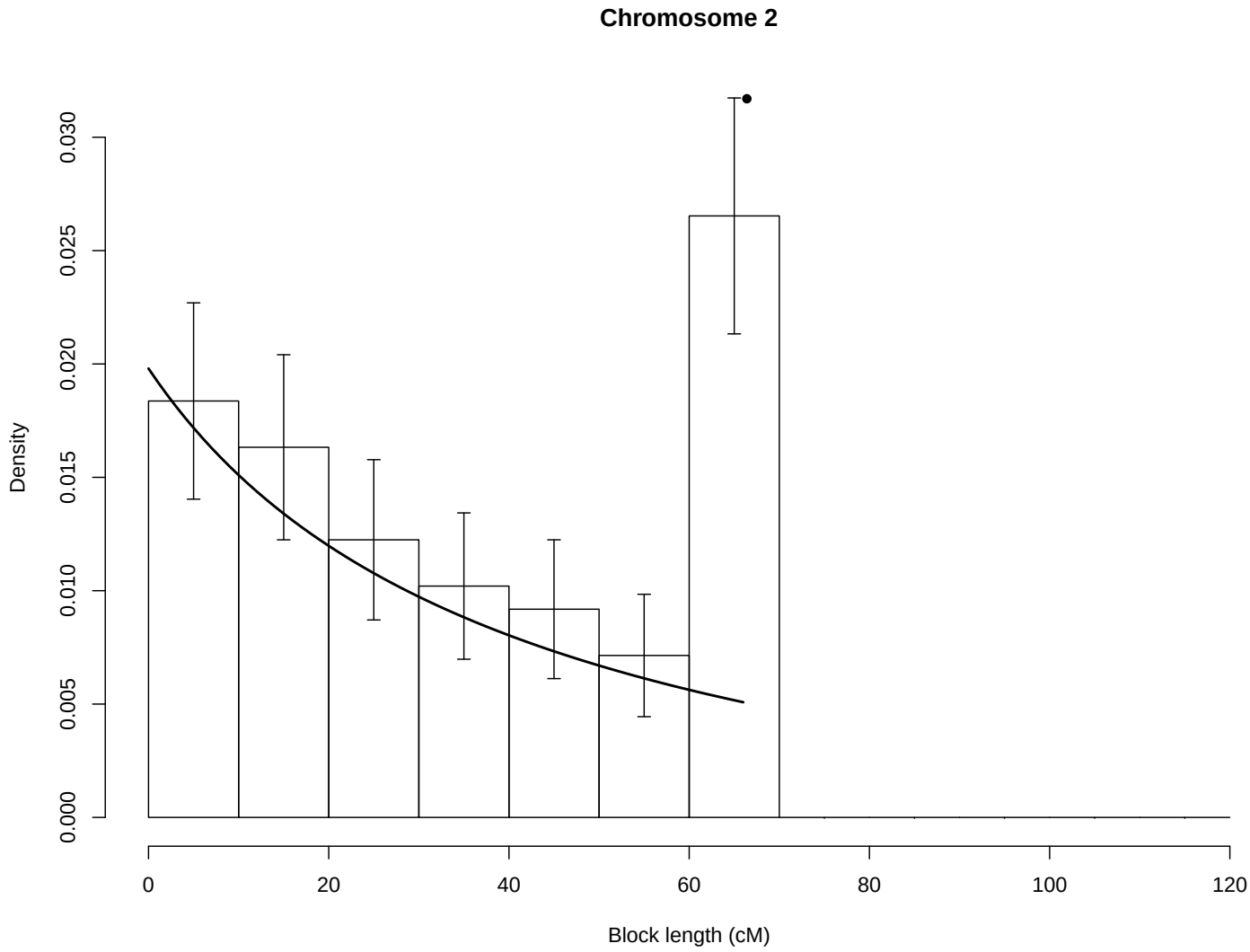


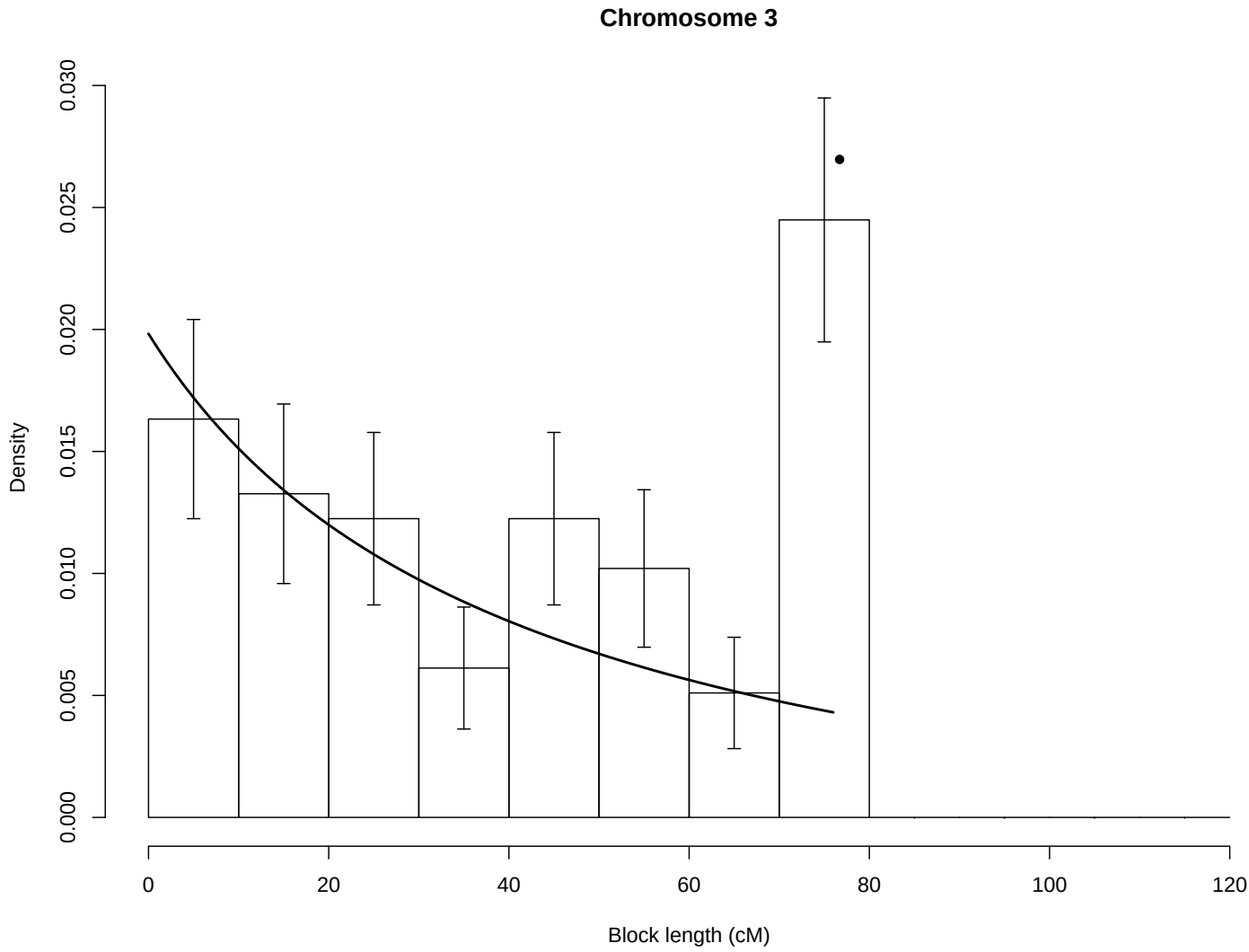

Figure S6: Probability density of the length of the first block on a long chromosome, assuming it is homozygous. Results are for $g = 1$ to $g = 7$ in SIB, exhibiting the convergence of block statistics with $g$.


Chromosome 1
Density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
100
120
Block length (cM)

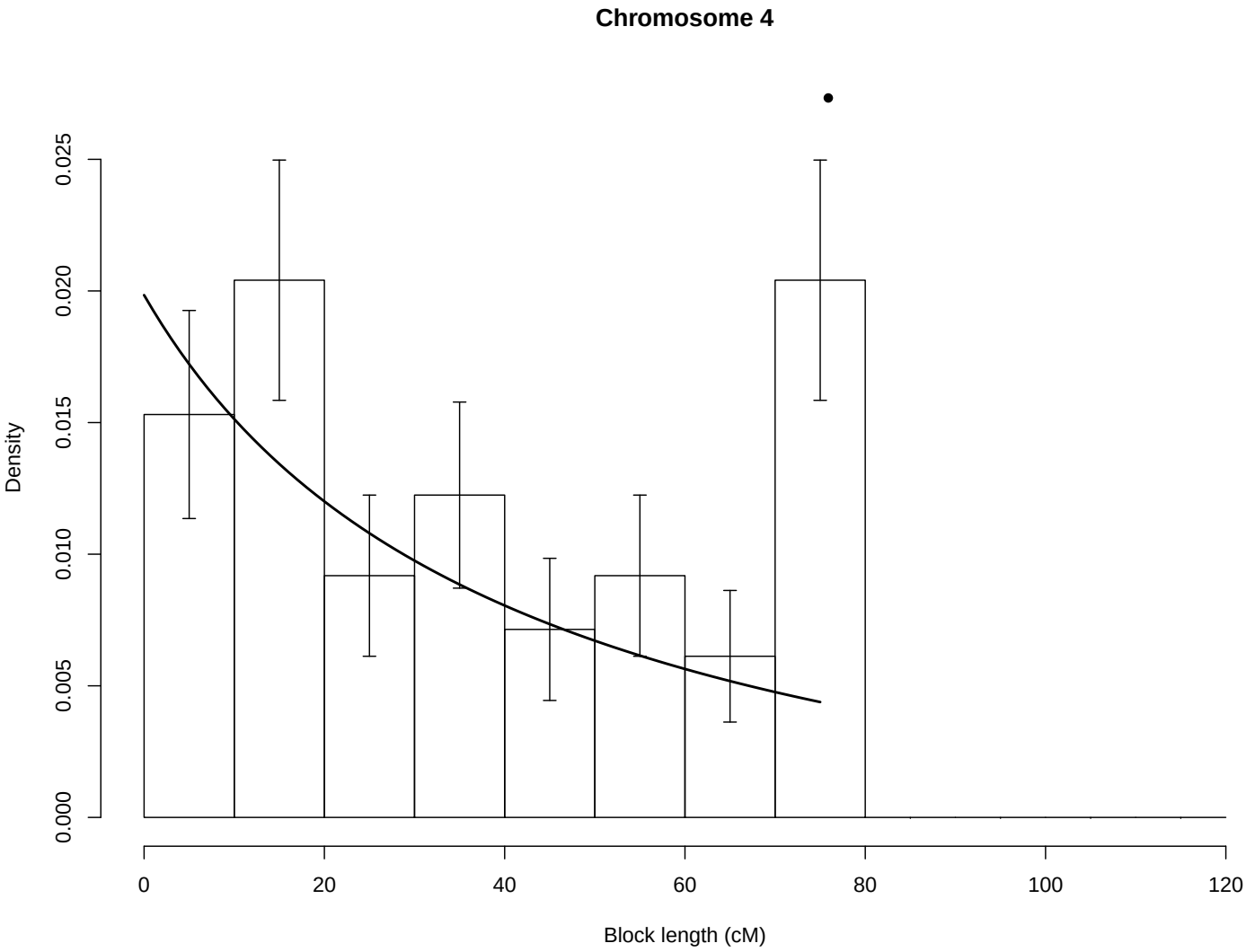
(a)


Chromosome 2
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
100
120
Block length (cM)

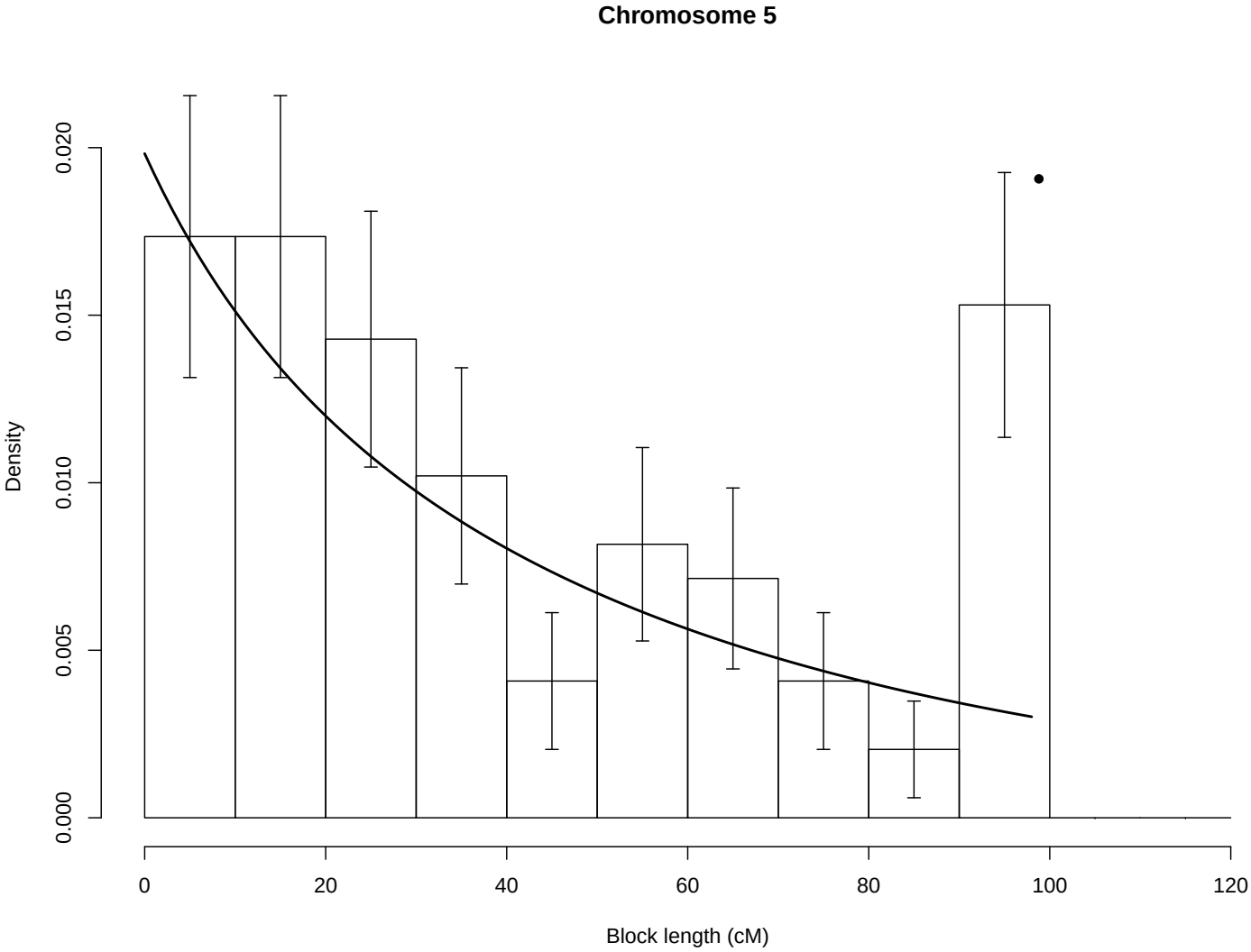
(b)


Chromosome 3
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
100
120
Block length (cM)

(c)


Chromosome 4
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
100
120
Block length (cM)

(d)




(e)

Figure S7: Comparison of theoretical predictions to experimental data for the length of the first block in SSD RILs. The experimental data is from Singer et al. [16] for 100 RILs and high density genotyping. The continuous curve is the theoretical distribution with the filled dot corresponding to the cases with no second block. The histograms are for the experimental data, and $95\%$ confidence intervals are also represented. Results for chromosomes 1 through 5 are shown on the successive subfigures.