Population size and the length of the chromosome blocks identical by descent over generations

Mathieu Tiret and Frédéric Hospital

Génétique Animale, INRA, Jouy-en-Josas, 78352, France

ABSTRACT In all populations, as the time runs, crossovers break apart ancestor haplotypes, forming smaller blocks at each generation. Some blocks, and eventually all of them, become identical by descent because of the genetic drift. We have in this paper developed and benchmarked a theoretical prediction of the mean length of such blocks and used it to study a simple population model assuming panmixia, no selfing and drift as the only evolutionary pressure. Besides, we have on the one hand derived, for any user defined error threshold, the range of the parameters this prediction is reliable for, and on the other hand shown that the mean length remains constant over time in ideally large populations.

KEYWORDS Identity-by-descent; Population genetics; Junctions; Wright-Fisher model; Poisson process

Identity by descent (IBD) was formally defined for two alleles only, and the definition was subsequently extended to a pair of chromosome segments. Two segments are said to be identical by descent if they are copies of a common ancestor segment without having undergone any mutation. Eventually, Stam (1980) proposed to widen the analyses of identity by descent into an entire population by studying every individual as a pair of chromosome segments. Henceforth, a portion of the chromosome will be referenced as a *segment* and a portion of a segment will be referenced as a *block*.

In a population undergoing genetic drift, as the time runs, crossovers break apart the ancestor haplotypes, forming smaller blocks at each generation. Some blocks might be identical by descent, in which case they are called "IBD blocks". In this paper, we will focus on 1) the distribution in the

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 20th November, 2015%
[1]Corresponding author: Tel: +331-3465-2198; E-mail address: mtiret@jouy.inra.fr

population of the length of an IBD block, and 2) the distribution of the number and the total length of IBD blocks per individual.

The theoretical framework dealing with these distributions uses a particular population model in order to take account of recombinations: the individuals of this population are diploid, and, following the idea of Stam (1980), they are modeled as a pair of homologous segments with a fixed length L (in Morgan). The population size, denoted N, is considered constant over generations, there is no evolutionary pressure but the genetic drift, and the generations do not overlap. Moreover, the segments of the founder – or initial – population are all different, that is, none of the $2N$ segments are identical by descent. Finally, the model assumes panmixia without selfing as in Stam (1980), so that two homologous segments of an individual are necessarily derived from two different individuals.

In this model, a segment can be considered as either a discrete or a continuous object. These two ways of modeling are rarely equivalent (Bickeböller and Thompson 1996a,b), and it is not always obvious to know which model should be used. To make the mathematical analyses easier, the segment is here considered as a continuous object: the recombination process can thus be modeled with a classic Poisson process with a rate equal to 1, neglecting crossover interference (Fisher 1949, 1954; Stam 1980; Donnelly 1983; Bickeböller and Thompson 1996a,b; Ball and Stefanov 2005; Chapman and Thompson 2002, 2003; Cannings 2003; Martin and Hospital 2011).

Theory of junctions

In order to describe the transmission over time of IBD blocks in such population, Fisher (1949) developed the so-called theory of junctions. A *junction* is here a crossover point delimiting two blocks coming from different founders. When considering two segments, it is possible to distinguish two types of junctions: *external junctions*, which are the edges of IBD blocks, and *internal junctions*, which are the other junctions (see Figure 1).

An edge of an IBD block is either an external junction or one of the segment edges. Then, knowing the segment length and the number of external junctions, it is possible to infer the distribution in the population of the length of an IBD block. Let $J_{N,t}$ be the expected number of IBD block edges per individual at generation t. In order to compute $J_{N,t}$, Fisher (1949) introduced $H_{N,t}$ and $Z_{N,t}$, which are respectively the expected non IBD proportion of a segment and the expected number of external

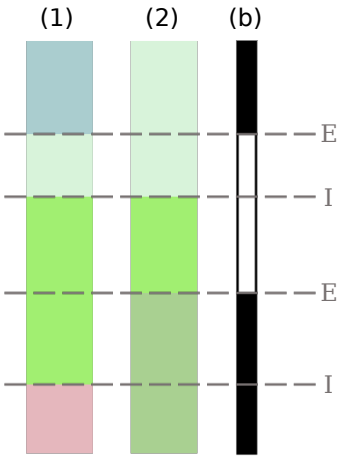

Figure 1 Two segments (1) and (2) of an individual, some time after founding. The different colors represent different founder segments. The white and black bar on the right indicates IBD (white) and non IBD (black) blocks, and each dotted line indicates a junction denoted either E if it is an external junction, or I if it is an internal junction.

junctions per Morgan *within the segment* (that is, excluding the segment edges). Besides, let us call IBD segment edge (ISE) a segment edge that is also an edge of an IBD block, and $I_{N,t}$ the expected number of ISE. Fisher (1949) estimated $I_{N,t}$ as $2 \cdot (1 - H_{N,t})$ and deduced the following equation:

$$J_{N,t} = L \cdot Z_{N,t} \; + \; I_{N,t} = L \cdot Z_{N,t} \; + \; 2 \cdot (1 - H_{N,t})$$

Therefore, the expected number $M_{N,t}$ of IBD blocks per individual is (Fisher 1949):

$$M_{N,t} \; = \; J_{N,t}/2 \; = \; 0.5 \cdot L \cdot Z_{N,t} \; + \; 1 - H_{N,t} \tag{1}$$

Fisher (1949) figured out a theoretical expression of both $H_{N,t}$ and $Z_{N,t}$ only for some very particular cases of relatedness shared by the individuals because of the complexity of the problem. Using identity relations between three genes and their recurrence relations, Stam (1980) derived an expression of $H_{N,t}$ and $Z_{N,t}$ for any case of relatedness (see Appendix A for the their formulation). We recall that all these expressions stand for one individual.

Thereby, Fisher (1949) and Stam (1980) described the identity by descent in a population at a *specific time*. Our objective here is to study the evolution of the identity by descent *over time* with a simple population model. To this end, we have developed and benchmarked a prediction of the expected length of an IBD block.

Model and methods

Simulations

We have implemented a program generating pseudo-data to compare the predictions with. This program simulates over generations the aforementioned population model. Here, a segment is modeled not as a set of nucleotides but as a continuous object, and hence the program records on the one hand the starting and ending edges of each block and on the other hand the origin of this block, which is the label of the founder segment this block belonged to.

With this program, we have simulated 10000 replicates over 3000 generations, with 10 different population sizes (from 10 to 100 individuals) and with a segment length ranging from 1 to 5 Morgan. The program has been implemented in C++ (version C++11, compiled with g++ 4.9.2), and graphical outputs have been obtained with R (version 3.2.2).

Length of identical by descent blocks

Let $X_{N,t}$ be the expected length of one IBD block. Stam (1980) and Chapman and Thompson (2003), assuming both that the dispatching of junctions over a genome followed a stationary process, obtained the following formulation of $X_{N,t}$:

$$X_{N,t} = \frac{L\,(1 - H_{N,t})}{0.5 \cdot L \cdot Z_{N,t}} \tag{2}$$

There is however a problem with this assumption: $X_{N,t}$ as formulated in equation (2) tends to infinity as t tends to infinity, whereas the length of an IBD block necessarily ranges from 0 to L. The problem comes from $Z_{N,t}$ that tends to zero as t tends to infinity (see Appendix A). The formulation of $Z_{N,t}$ is correct though, and its asymptotic behavior is indeed expected: genetic drift makes every segment identical after a while, making thus the number of external junctions $Z_{N,t}$ tend towards zero. The use, however, of the multiplicative inverse of $Z_{N,t}$ is incorrect, implying that the stationary process assumption is seemingly incompatible with a study of $X_{N,t}$ over time.

Using equation (1), it is possible to derive another expression of $X_{N,t}$:

$$
\begin{aligned}
X_{N,t} &= \frac{\text{Length per individual of the segment that is IBD}}{\text{Number per individual of IBD blocks}} \\
&= \frac{L\,(1-H_{N,t})}{M_{N,t}} \\
&= \frac{L\,(1-H_{N,t})}{(0.5\cdot L\cdot Z_{N,t})+(1-H_{N,t})}
\end{aligned}
\tag{3}
$$

For the problem mentioned above, only equation (3) will hereafter be considered.

Predictions and observations

The theoretical predictions used in this paper are summarized in Table 1. We recall that, according to equation (3), $X_{N,t}$ is a combination of $H_{N,t}$, $Z_{N,t}$ and $I_{N,t}$.

Description	**prediction**	**observation**
Non IBD proportion	$\widehat{H}_{N,t}$	$h_{N,t}$
Number of external junctions	$\widehat{Z}_{N,t}$	$z_{N,t}$
Number of ISE	$\widehat{I}_{N,t}$	$i_{N,t}$
Length of one IBD block	$\widehat{X}_{N,t}$	$x_{N,t}$

Table 1 Predictions and observations.

Of all observations, only the average length $x_{N,t}$ of an IBD block could be defined in different ways. We have chosen here to define $x_{N,t}$ as:

$$
x_{N,t} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} l_{ij}}{\sum_{i=1}^{N} n_i} \tag{4}
$$

with n_i the number of IBD blocks in the i$^{\text{th}}$ individual and l_{ij} the length of the j$^{\text{th}}$ block in the i$^{\text{th}}$ individual. Thus, $x_{N,t}$ is defined not as an individual wise but as a population wise observation.

The focused range of time and the prediction error

We can see on Figure 2 and 3 that $X_{N,t}$ begins with a high peak, due to the fact that all the founder segments are different. Since we are not interested in this artifactual peak, we will hereafter focus

N	10	20	30	40	50	60	70	80	90	100
$\widehat{H}_{N,t}$	5.96	2.90	1.92	1.26	1.01	0.93	0.75	0.64	0.55	0.48
$\widehat{Z}_{N,t}$	6.73	3.20	2.41	1.40	1.13	1.17	0.94	0.78	0.70	0.59
$\widehat{I}_{N,t}$	2.81	1.09	0.82	0.60	0.55	0.46	0.47	0.53	0.41	0.48
$\widehat{X}_{N,t}$	48.11	41.12	35.88	34.54	32.92	30.96	29.58	28.82	28.42	27.67

Table 2 The thousandfold value of the average prediction error $\tilde{pe}$ of the different predictions $H_{N,t}$, $Z_{N,t}$, $I_{N,t}$ and $X_{N,t}$.

on a range of time starting at $T_{N,min}$, the time at which $x_{N,t}$ reaches its minimum after the peak, and ending at generation 3000. The latter is a fixed value, and the former mainly depends on the population size (see Appendix B).

The fit between predictions and observations was ascertained using the prediction error pe. With $\widehat{a}_t$ the value at generation t of a prediction and a_t the value of the corresponding observation, we have $pe\{\widehat{a}_t\} = |\widehat{a}_t - a_t| / \max_t a_t$.

We have also provided the average value over time of the prediction error, denoted $\tilde{pe}$. Our main interest being $X_{N,t}$, we have searched which component of $X_{N,t}$ had the greatest prediction error: $H_{N,t}$, $Z_{N,t}$, $I_{N,t}$ or their combination.

Results

The prediction error

Figure 2 shows the prediction and the observation of $X_{N,t}$ over time, for various population sizes and segment lengths. The average prediction error over $T_{N,min}$ to generation 3000 are summarized in Table 2.

Table 2 shows that the average prediction error $\tilde{pe}\{\widehat{X}_{N,t}\}$ is at least ten times greater than the other average prediction errors, whatever the population size: it seems that a large part of $\tilde{pe}\{\widehat{X}_{N,t}\}$ comes from the combination of $H_{N,t}$, $Z_{N,t}$ and $I_{N,t}$ rather than from each prediction.

We have plotted the prediction error $pe\{\widehat{X}_{N,t}\}$ on Figure 3. This figure shows that the error is not constant over time, and has rather a pattern divisible in three phases. Let α be a fixed error threshold, ranging from 0 to 1. We define the *first phase* as the phase during which $pe\{\widehat{X}_{N,t}\}$ is less than α;

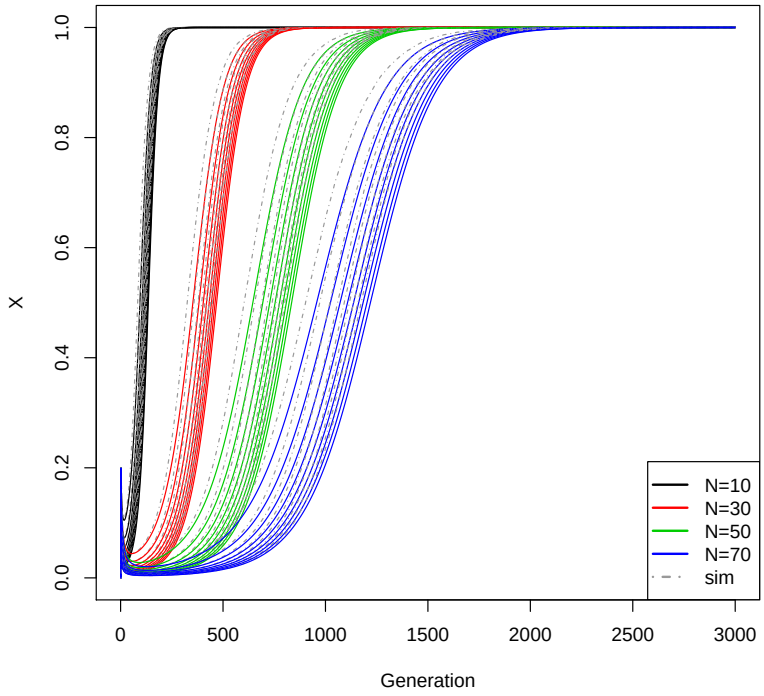

Figure 2 Comparing theoretical prediction and simulation values of $X_{N,t}$, with a population size of 10 (black line), 30 (red line), 50 (green line) or 70 individuals (blue line). The solid lines are the estimations, and the dotted lines are the observations (10000 replicates, denoted `sim` in the legend). The different lines of the same color corresponds to the different segment length, ranging from 1 to 5 Morgan.

N	10	20	30	40	50	60	70	80	90	100
μ	124.67	277.28	442.39	614.83	795.89	979.67	1167.78	1357.06	1554.44	1749.39
σ	29.67	58.83	86.28	112.39	142.44	168.11	193.78	220.50	251.00	277.61
$\mathbf{R^2}$	98.88	98.97	98.86	98.54	98.76	98.45	98.17	98.21	98.40	98.13

Table 3 The parameters (the mean μ and the standard deviation σ) of the fitted normal distribution and the R-squared (in percent) of this fitting.

Time	**Description**	**Formulation**	$\mathbf{R^2}$
$T_{1,\alpha}$	Beginning of the transition phase	$\mu - \sqrt{-2\sigma^2 \log \alpha}$	99.70
$T_{2,\alpha}$	Ending of the transition phase	$\mu + \sqrt{-2\sigma^2 \log \alpha}$	99.15

Table 4 The beginning and the ending times of the transition phase, their formulation and the R-squared of the formulation. α is a real number ranging from 0 to 1, μ and σ are respectively the mean and the standard deviation of the fitted normal distribution.

then the *transition phase* during which $pe\{\widehat{X}_{N,t}\}$ is greater than α; finally, the *final phase* during which $pe\{\widehat{X}_{N,t}\}$ is less than α again. Figure 3 shows that the transition phase, during which $pe\{\widehat{X}_{N,t}\}$ is the greatest, corresponds to the increasing phase of $X_{N,t}$.

Figure 3 also shows that $pe\{\widehat{X}_{N,t}\}$ is almost a Gauss-like function and so a normal density function, up to a multiplicative constant. After having fitted $pe\{\widehat{X}_{N,t}\}$ for each population size to a normal gaussian function (see Appendix C), we have deduced the parameters of this function and summarized them in Table 3.

According to this fitting and for any value of α, we have finally derived the beginning time $T_{1,\alpha}$ and the ending time $T_{2,\alpha}$ of the transition phase (summarized in Table 4). Between $T_{N,min}$ and $T_{1,\alpha}$ and from $T_{2,\alpha}$ until generation 3000, the average prediction error is less than α. Furthermore, using the linear regressions of μ and σ, we have derived that the $T_{1,\alpha} \simeq 18.17N - 92.95 + (2.74N + 3.74)\sqrt{-2\log\alpha})$ (for further details, see Appendix C):

The minimum length of IBD blocks

Figure 3 shows that our prediction of the expected length of an IBD block is reliable during the first phase, between $T_{N,min}$ and $T_{1,\alpha}$. During this phase, $X_{N,t}$ increases only slightly, hence its narrow value range. Besides, the first phase lasts longer as the population size increases, since $T_{1,\alpha}$ increases

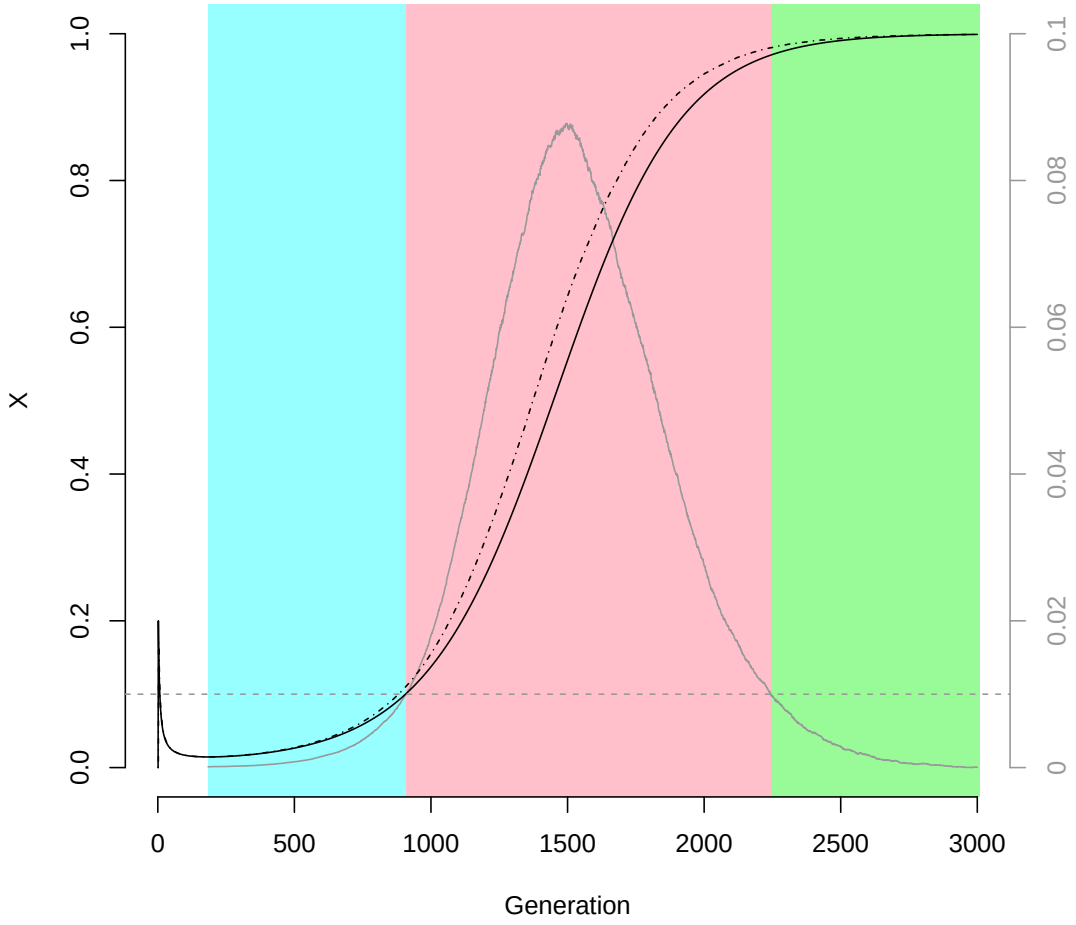

Figure 3 A superposition of the prediction error $pe\{\widehat{X}_{N,t}\}$, $x_{N,t}$ and $\widehat{X}_{N,t}$ for a population of 100 individuals, α of 0.01 (the dashed horizontal line) and a segment length of 1 Morgan. The black solid line is $\widehat{X}_{N,t}$, the black dotted line is $x_{N,t}$ and the solid gray line is the prediction error $pe\{\widehat{X}_{N,t}\}$, scaled for visual purposes. Its axis is on the right. The blue region corresponds to the first phase, the pink region to the transition phase and the green region to the final phase.

ten times faster than $T_{N,min}$ does (see Appendix B). Consequently, for a fixed threshold α, the larger a population is, the flatter $X_{N,t}$ seems to be. The minimum value of $X_{N,t}$ during the first phase can be estimated as $3/2N$ according to the method of Newton (see Appendix B).

Discussion

Knowing that the minimum value of $X_{N,t}$ during the first phase can be estimated as $3/2N$, $X_{N,t}$ could therefore be, in an ideally large population, considered constant and equal to $3/2N$ during the first phase. As Table 3 and Table 4 show, the duration of the first phase rapidly increases with the population size, flattening $X_{N,t}$ and making the ideally large population assumption stand even for a population with 100 individuals, whose first phase lasts about 1,600 generations. It is hence consistent to assume that most populations deriving from highly diverse founders are likely to be in the first phase nowadays, only if it is consistent to assume the model of Haldane, a constant population size and that genetic drift is the only evolutionary pressure.

Mutations were neglected here because the genome was modeled as a continuous object. Indeed, mutations are points, and punctual items do not exist in such continuous models. Compared with the discrete approach, the continuous approach has the advantage of easing the mathematical analyses, but in counterpart it has the shortcoming to assume that there is at the first sight no mutation (it is possible to extend the continuous approach with the infinite allele or the infinite site assumption though), whereas the occurrence of recombination is on average of the same order of magnitude as the occurrence of mutations (the recombination rate is around 10^{-8} per nucleotide per generation for humans). Neglecting mutations is therefore an important limitation to overcome in the future.

Finally, an extension of this framework will consist on the one hand in theoretically determining the variance of the distribution of $X_{N,t}$ and on the other hand in focusing on the impact of the founder population and its structure: assuming that every founder segment is different, as we did here (according to Stam 1980), is more than unlikely in a real population. It will be important in further studies to assess whether this structure changes the whole dynamic or only, as for a Markovian process, the beginning of the process.

Acknowledgements

This work has been supported by grants from the metaprogram SelGen, Institut Nationale de Recherche Agronomique, INRA (to F.H.). We are also grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing help and computing resources.

Literature Cited

Baird, S. J. E., N. H. Barton, and A. M. Etheridge, 2003 The distribution of surviving blocks of an ancestral genome. Theoretical Population Biology **64**: 451–471.

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

Bickeböller, H. and E. A. Thompson, 1996a Distribution of genome shared IBD by half-sibs: approximation by the poisson clumping heuristic. Theoretical population biology **50**: 66–90.

Bickeböller, H. and E. A. 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. L. Browning, 2002 On reducing the statespace of hidden markov models for the identity by descent process. Theoretical Population Biology **62**: 1–8.

Browning, S. R., 2008 Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics **178**: 2123–2132.

Cannings, C., 2003 The identity by descent process along the chromosome. Human Heredity **56**: 126–130.

Chapman, N. H. and E. A. Thompson, 2002 The effect of population history on the lengths of ancestral chromosome segments. Genetics **162**: 449–458.

Chapman, N. H. and E. A. Thompson, 2003 A model for the length of tracts of identity by descent in finite random mating populations. Theoretical Population Biology **64**: 141–150.

Dimitropoulou, P. and C. Cannings, 2003 RECSIM and INDSTATS: probabilities of identity in general genealogies. Bioinformatics **19**: 790–791.

Donnelly, K., 1983 The probability that related individuals share some section of the genome identical by descent. Popul. Biol. **23**: 34.

Fisher, R. A., 1949 *The Theory of Inbreeding*. Oliver & Boyd, Edinburgh.

Fisher, R. A., 1954 A fuller theory of "junctions" in inbreeding. Heredity **8**: 187–197.

Fisher, R. A., 1959 An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. Heredity **13**: 523–542.

Goldstein, D. B., 2001 Islands of linkage disequilibrium. Nature genetics **29**.

Haldane, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. Genetics **8**: 299–309.

Hayes, B. J. and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Karlin, S. and H. M. Taylor, 1975 *A First Course in Stochastic Processes*. Academic Press, San Diego.

Martin, O. C. and F. Hospital, 2011 Distribution of parental genome blocks in recombinant inbred lines. Genetics **189**: 645–654.

Rodolphe, F., J. Martin, and E. Della-Chiesa, 2008 Theoretical description of chromosome architecture after multiple back-crossing. Theoretical Population Biology **73**: 289–299.

Sabeti, P. C., V. K. Iyengar, H. K. Reeve, and T. Eisner, 2002 Paternal inheritance of a female moth's mating preference. Nature **419**: 830–832.

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

Stumpf, M. P., 2002 Haplotype diversity and the block structure of linkage disequilibrium. TRENDS in Genetics **18**: 226–228.

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

Appendices

A. Mathematical formulation of *H* and *Z*

The exact expression of $H_{N,t}$ and $Z_{N,t}$ was described by Stam (1980) and reads:

$$
\begin{aligned}
H_{N,t} &= \frac{1-\lambda_3}{\lambda_1-\lambda_3}\,\lambda_1^t \;-\; \frac{1-\lambda_1}{\lambda_1-\lambda_3}\,\lambda_3^t \\
Z_{N,t} &= (c_1 + c_2\,t)\,\lambda_1^t \;+\; (c_3 + c_4\,t)\,\lambda_3^t \;+\; c_5\,\lambda_5^t \;+\; c_6\,\lambda_6^t
\end{aligned}
$$

with c_i's and λ_i's values depending on N, as follows:

$$
\begin{aligned}
\lambda_1 &= (N-1+\sqrt{N^2+1})/2N \\
\lambda_3 &= (N+1+\sqrt{N^2+1})/2N \\
\lambda_5 &= (a+D)/2N^2 \\
\lambda_6 &= (a-D)/2N^2 \\
c_1 &= -\frac{2N}{\sqrt{N^2+1}}\left(\frac{N}{N^2+1} + \frac{(N^2-2)(N+\sqrt{N^2+1})+2}{N^2+(N-2)\sqrt{N^2+1}}\right) \\
c_2 &= \frac{N^2+N+2+(N+2)\sqrt{N^2+1}}{N^2+1} \\
c_3 &= \frac{2N}{\sqrt{N^2+1}}\left(\frac{N}{N^2+1} + \frac{a_n+b_n\sqrt{N^2+1}+c_n\cdot D+d_n\cdot D\sqrt{N^2+1}}{a_d+b_d\sqrt{N^2+1}+c_d\cdot D+d_d\cdot D\sqrt{N^2+1}}\right) \\
c_4 &= \frac{N^2+N+2-(N+2)\sqrt{N^2+1}}{N^2+1} \\
c_5 &= \frac{N}{D}\frac{4N^5-6N^4+2N^3-16N^2+32N-16+D(4N^3-12N+8)}{4N^3-5N^2+4N-4} \\
c_6 &= \frac{N}{D}\frac{-4N^5+6N^4-2N^3+16N^2-32N+16+D(4N^3-12N+8)}{4N^3-5N^2+4N-4}
\end{aligned}
$$

with

$$a = N + (N-1)(N-2)$$

$$D = \sqrt{a^2 + 2N(N-1)(N-2)}$$

$$a_n = N^5 - N^4 - N^3 + 2N^2 - 2N$$

$$b_n = -N^4 + N^3 + 2N^2 - 2N$$

$$c_n = -N^2 + 3N - 2$$

$$d_n = -2 + N$$

$$a_d = -N^5 + 3N^4 - 4N^3 + 6N^2 - 8N + 4$$

$$b_d = N^4 - 2N^3 + 2N^2 - 4N + 4$$

$$c_d = N^3 - 2N + 2$$

$$d_d = -N^2 + N$$

To ease mathematical analyses, we have also derived the series expansion as N tends towards infinity of those coefficients:

$$\begin{aligned}
\lambda_1 &= 1 - 1/2N + o(N^{-2}) \\
\lambda_3 &= -1/2N + o(N^{-2}) \\
\lambda_5 &= 1 - 3/2N + o(N^{-2}) \\
\lambda_6 &= -1/2N + o(N^{-2}) \\
c_1 &= -2N - 2 + 1/N + o(N^{-2}) \\
c_2 &= 2 + 3/N + o(N^{-2}) \\
c_3 &= 7/2N + o(N^{-2}) \\
c_4 &= -1/N + o(N^{-2}) \\
c_5 &= 2N + 2 - 3/N + o(N^{-2}) \\
c_6 &= 0.5 - 15/8N + o(N^{-2})
\end{aligned}$$

Hence, we can derive that:

$$\begin{aligned}
H_{N,t} &\simeq (1 + 1/2N)\exp\{-t/2N\} \\
Z_{N,t} &\simeq (-2N - 2 + 2t)\exp\{-t/2N\} + 2N\exp\{-3t/2N\}
\end{aligned}$$

It is now trivial to deduce that $Z_{N,t}$ tends towards zero as t tends towards infinity.

B. The minimum value of *X*

We have derived with a linear regression on simulations that $T_{N,min} \simeq 1.87N - 1.27$ with an adjusted R-squared of 0.9992 and that $\min_t \{X_{N,t}\} \simeq 1/(0.66N + 2.40)$ with an adjusted R-squared of 1. Besides, using the Newton's method on theoretical expressions of the predictions, we have derived that $\min_t \{X_{N,t}\} \simeq 1/(0.66N + 2.80)$. Both approximations of $\min_t \{X_{N,t}\}$ are equivalent and could roughly be approximated as $3/2N$.

C. Normal adjustment

We estimated here the parameters μ and σ of the fitted normal distribution of pe as follows:

$$\begin{aligned} \mu &= T_{N,min} + 0.5(pe^{-1}_{]T_{N,min},t_{max}]}\{e^{-0.5}\} + pe^{-1}_{]t_{max},3000]}\{e^{-0.5}\}) \\ \sigma &= 0.5(pe^{-1}_{]T_{N,min},t_{max}]}\{e^{-0.5}\} - pe^{-1}_{]t_{max},3000]}\{e^{-0.5}\}) \end{aligned}$$

with t_{max} the time at which pe reaches its maximum and pe^{-1}_A the inverse function of pe on A.

The coefficient of determination R^2 of the normal fitting was computed as follows:

$$R^2 = 1 - \frac{\sum_{t=T_{N,min}}^{3000}(pe\{\widehat{X}_{N,t}\} - S \cdot f_{\mu,\sigma}\{t\})^2}{\sum_{t=T_{N,min}}^{3000}(pe\{\widehat{X}_{N,t}\} - \tilde{pe}\{\widehat{X}_{N,t}\})^2}$$

with S a corrective coefficient (equal to $\sum_{t=T_{N,min}}^{3000} pe\{\widehat{X}_{N,t}\}$) to make pe have its integration over the real numbers equal to 1, $f_{\mu,\sigma}$ the density function of the normal distribution with a mean μ and a standard deviation σ, and $\tilde{pe}\{\widehat{X}_{N,t}\}$ the average value of pe over $T_{N,min}$ to the generation 3000.

With a linear regression, we have derived that $\mu \simeq 18.17N - 92.95$ with an adjusted R-squared of 0.9773 and $\sigma \simeq 2.74N + 3.47$ with an adjusted R-squared of 0.9959.