THEORETICAL DESCRIPTION OF CHROMOSOME ARCHITECTURE AFTER MULTIPLE BACK-CROSSING

by

François Rodolphe, Juliette Martin and Emmanuelle Della-Chiesa

Research Report No. 2
January 2007

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

THEORETICAL DESCRIPTION OF CHROMOSOME ARCHITECTURE AFTER MULTIPLE BACK-CROSSING

François Rodolphe[†], Juliette Martin[†,*], Emmanuelle Della-Chiesa[‡]

[†] INRA, Unité Mathématique, Informatique et Génome, UR1077, F - 78350 Jouy-en-Josas
[*] INSERM, Equipe de Bioinformatique Génomique et Moléculaire U726, Université Paris 7, F75005 Paris.
[‡] Laboratoire Statistique et Génome, UMR CNRS 8071 - INRA 1152 - Université d'Évry, F - 91000 Evry

Abstract

This paper provides a theoretical description of the chromosome architecture resulting from a given number of generations in a back-cross. It is worth considering chromosome architecture as depending on a marked point process, whose properties depend on the crossing-over model used. Resulting architecture is presented here for two different models : no interference, and complete interference. Exact distributions, with easy-to-compute formulae, are derived for quantities of interest, as the length of donor or receiver fragments, for any chromosome length and for both crossing-over models. Examples are presented to illustrate the use of these distributions in introgression programs or in population genetics.

Introduction

Introgression is a technique often used for a long period, for instance, by plant breeders who want to introduce a monogenic character available in a wild genotype (the donor) into a cultivated variety (the receiver), without altering other characteristics.

Introgression starts with the hybrid (generation 0), which is crossed with the receiver. Product genotypes containing one copy of the desired gene are selected, and crossed again with the receiver, and so on for several generations. At each stage, genotypes bear a complete set of receiver chromosomes, and the other half of their genome, owing to crossing-over, is a mosaic, which, due to selection, bears the desired gene. It is well known that the expected length of the donor chromosome fragment bearing the gene of interest, reduces at each generation, and that the donor genome is progressively washed out from the rest of the genome (Fisher R. 1949, Hanson W. D. 1959, Stam P. and Zeven C., 1981). Naveira H. and Barbadilla A. (1992) give an extensive review of the question, and provide exact expressions for the mean and standard deviation of the length, when there is no interference. Hill (1993) provides expressions for the two first moments of donor contribution (the proportion of the complete genome copied from the donor).

The aim of this note is to derive a general and complete description of the mosaic chromosome structure. It can be adapted to different crossing-over models, and makes possible the computation of the probability of any event of interest like the number of remaining segments of the donor chromosome. In an introgression program, these computations provide a tool for experimental design. This description is obtained by constructing a marked point process which contains all needed information.

Materials and methods

Genetic models and point process construction

We assume throughout this paper, chromosomes to be independent, hence only one will be considered, and crossing-over to occur at each generation independently from the past. Genetic distances will be considered with two different crossing-over models. In the first one (W), no interference between cross-overs is assumed : they occur according to a homogeneous Poisson process. In the second one (C), complete interference between cross-overs is assumed : one and only one cross-over occurs at each generation, on each chromosome arm.

At each generation i, a point process, X^i, describes the crossing-over. X^i_j is the coordinate of the jth cross-over having occured at generation i. Consider now the marked point process (X^+, Y) which results from the superposition of all point processes having occured at each generation between 1 and n ; points are renumbered and each point, X^+_j, bears the mark, Y_j, of the generation at which the corresponding cross-over occured (figure 1).

Our claim is that chromosome architecture is determined by the marked point process (X^+, Y), whose properties depend on the crossing-over model. This provides a way to a complete description of the chromosome architecture

Structure of the mosaic chromosome

Consider first the case of introgression illustrated in figure 1. Due to selection in the experimental design, the locus of interest, L, is copied from the donor ; the segment, bearing this locus, will be interrupted by the next cross-over, in both directions. Consider now another fragment copied from the donor chromosome, if any, which does not bear the selected gene. It starts at a given point of the process X^+ and stops at the next.

Let us now examine the relationship between the marked point process (X^+, Y) and the reappearence of a donor genome fragment along the chromosome. We start on L, the locus to be introgressed. The first point, X_1^+, encountered stops the fragment and starts a new interval, copied from the receiver chromosome. Its mark, Y_1, indicates at which generation this interruption occured. All the segment limited by X_1^+ and the next point of X^+ bearing the same mark, has been copied from the receiver at generation Y_1. From then on, in a back-cross, this entire segment will remain a copy of the receiver. This holds at any generation. Hence donor chromosome will reappear if and only if all marks appeared, each one, an even number of times (zero is even) since we left the last donor segment, see figure 1.

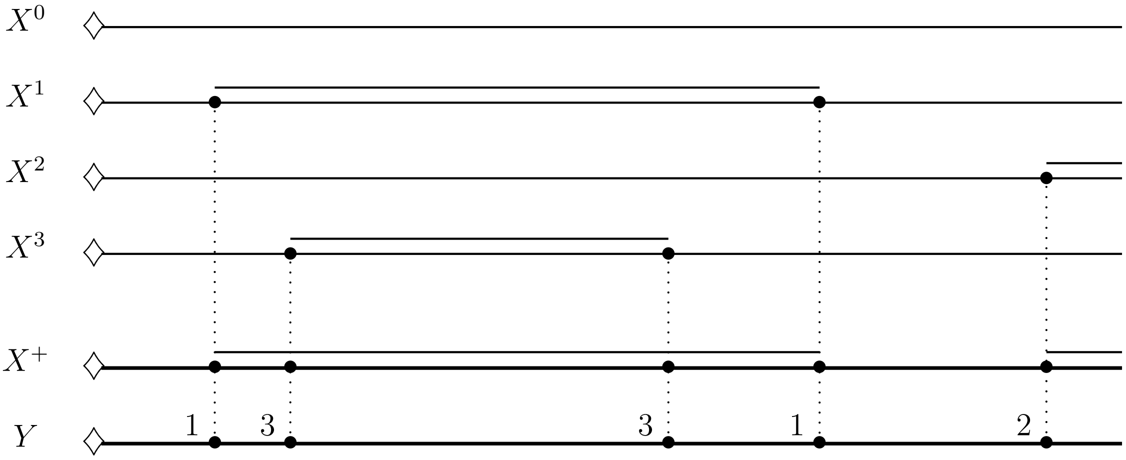

FIG. 1 – Construction of the marked point process (X^+, Y), based on the crossing-over processes X^i at each generation i. Bullets represent cross-overs, diamond the gene to be introgressed. Chromosome fragments copied from the receiver at each generation and at the end of the experiment, are indicated by a double line.

Crossing-over without interference on an infinite chromosome

When no interference is assumed, crossing-over occur at each generation, according to a homogeneous Poisson process, with intensity one, by definition. In this section, the problem will be considered on the line; in the next sections it will be considered on a finite chromosome.

The resulting point process, X^+, after n generations, is itself a homogeneous Poisson process of intensity n. Processes X^i being independent, Poisson (memoryless), all with the same intensity, marks are distributed independently and at random : Y is a series of independent variables with uniform distribution on the integers $\{1, 2, \cdots n\}$ (a homogeneous Bernoulli process). Moreover, processes X^+ and Y are independent.

Donor chromosome fragments

Waiting time for the next point in process X^+ has an exponential distribution of parameter n, with density of probability $g_n^W(z) = nexp(-nz)$. Total length of the donor chromosome fragment bearing the gene to be introgressed, corresponds to two independent waiting times : one on the right, one on the left. It is distributed as $(g_n^W)^{*2}(z) = n^2 z.exp(-nz)$, the convolution square of $g_n^W(z)$. On the contrary, unselected fragments inherited from the donor, have an exponential distribution of parameter n, with density of probability $g_n^W(z)$.

Receiver chromosome fragments

Consider two urns ("even" and "odd") with n balls numbered from 1 to n. Starting in a donor fragment, we put all balls in the "even" urn. Each time we encounter a point X_j^+ in the X^+ process, we move the ball numbered Y_j. The "even" urn contains the balls numbered as the marks which appeared an even number of times, since we left the last donor fragment. Donor chromosome will reappear when and only when "odd" urn gets empty. The number of balls moved, before returning to zero, equals one plus the number of intervals covered before meeting again the donor chromosome, in the process X^+.

Because of the nature of the process Y, the number of balls in the "odd" urn is an Ehrenfest promenade with parameter n. We can now derive the receiver chromosome fragment length probability distribution after n generations. Such a fragment can be considered as a concatenation of intervals of lengths independent and identically exponentially distributed, with parameter n. The number of fragments to be concatenated is distributed as the zero return time, minus one, of an Ehrenfest promenade with parameter n.

Let Z be the length of a receiver chromosome fragment, we have for Z, after n generations, the following density of probability :

$$f_n^W(z) = \sum_{i=1}^{\infty} \pi_{2i}^{(n)} (g_n^W)^{*(2i-1)}(z) = \sum_{i=1}^{\infty} \pi_{2i}^{(n)} n^{2i-1} z^{2(i-1)} exp(-nz)/(2(i-1))!$$

where $\pi_{2i}^{(n)}$ is the probability, zero return time equals $2i$ (zero return times are even). If $n = 1$ this formula reduces to $f_1^W(z) = g_1^W(z) = exp(-z)$ since, if there is only one ball in an Ehrenfest promenade, zero return time equals 2 with probability one. The kth convolution power of the exponential distribution of parameter n, $(g_n^W)^{*k}(z) = n^k z^{k-1} exp(-nz)/(k-1)!$, is the density of probability of the total length of k successive segments. It is known as an Erlang distribution, in fact, a gamma distribution with integer parameter k.

Let $G_n^{W,k}(z)$ be the cumulative distribution function corresponding to $(g_n^W)^{*k}(z)$, $\forall k$:

$$G_n^{W,k}(z) = 1 - exp(-nz) \sum_{i=0}^{k-1} (nz)^i / i!$$

The cumulative distribution function $F_n^W(\cdot)$ of Z, corresponding to f_n^W, has the following expression :

$$F_n^W(z) = \sum_{i=1}^{\infty} \pi_{2i}^{(n)} (G_n^W)^{(2i-1)}(z) = 1 - exp(-nz) \sum_{i=1}^{\infty} \pi_{2i}^{(n)} \sum_{j=0}^{2(i-1)} (nz)^j / j! \qquad (1)$$

This formula is not suitable for computations. A computable formula, 14, is provided in appendices and . According to this formula, F_n^W is a linear combination of n exponentials, depending on constants β_k and d_k. Therefore, all expressions which will be derived below, can be directly computed as finite linear combinations of exponentials too, with no need of numerical integration.
These distributions are shown on figure 2 for a generation number between 1 and 8. As it can be seen on these figures, the slope of these cumulative distribution functions is always 1 at the origin. Indeed :

$$\frac{dF_n^W}{dz}(0) = f_n^W(0) = \sum_{i=1}^{\infty} \pi_{2i}^{(n)} g_n^{*(2i-1)}(0) = \pi_2^{(n)} 1 = \frac{1}{n} n = 1, \forall n \qquad (2)$$

This means that receiver genome fragments, whose lengths generally increase very fast with n (mean value is $(2^n - 1)/n$, see appendix), keep a probability of taking tiny values, very small but quite unsensitive to n.

Whole chromosome architecture

As an Ehrenfest promenade is a Markov chain, each return to zero is a renewal, hence all fragments have independent lengths. After n generations of back-cross, resulting mosaic chromosome is a concatenation of fragments of independent lengths, alternatively copied from the donor and the receiver chromosomes. Fragment lengths are distributed according to densities g_n^W for the donor, and f_n^W for the receiver. The donor fragment bearing the unique gene to be introgressed has a length distribution $(g_n^W)^{*2}$.

Results

Crossing-over without interference on a finite chromosome

We now consider a chromosome of length K (in Morgans). A finite chromosome corresponds to a window of length K on the line. As we are no longer considering the process on the line, but on a finite chromosome, we must here take into account that any segment can be truncated by chromosome ends.

Introgression : length of the segment bearing the selected gene on a finite chromosome arm without interference

Let B_l and B_r be the left and right ends of the segment copied from the donor and bearing, at locus L, the selected gene. Both segment ends are independent stochastic variables. Their joint probability density is, for $0 \leq x \leq l$ and $l \leq y \leq K$:

$$\begin{aligned} p_n^W(x,y) = & \ (ne^{-n(l-x)} + e^{-nl}\delta_0(x))(ne^{-n(y-l)} + e^{-n(K-l)}\delta_K(y)) \\ & = e^{-n(y-x)}(n^2 + n(\delta_0(x) + \delta_K(y)) + \delta_0(x)\delta_K(y) \end{aligned}$$

where $\delta_a(x)$ represents Dirac measure on a. Probability density of length Z of the interval $[B_l, B_r]$ is $h_{n,l}^W(z) = \int_{0\vee(l-z)}^{l\wedge(K-z)} p_n(x, x+z)dx$ on $[0, K]$, where $a \wedge b$ and $a \vee b$ stay for the minimum and the maximum of a and b.

$$h_{n,l}^W(z) = e^{-nz}(n^2(l \wedge (K-z) - 0 \vee (l-z)) + n(1_{\{z \geq l\}} + 1_{\{z \geq K-l\}}) + \delta_K(z)) \quad (3)$$

Let $H_{n,l}^W(z) = P[Z \leq z]$ be the cumulative distribution function of the interval length. If $l < \frac{K}{2}$, $H_{n,l}^W(z)$ has the following expressions :

$$\begin{array}{rl} \textit{if } z \leq l & \int_0^z e^{-nu}n^2 u du = 1 - e^{-nz}(1+nz) \\ \textit{if } l \leq z \leq K-l & 1 - e^{-nl}(1+nl) + \int_l^z e^{-nu}(n^2 l + n)du = 1 - e^{-nz}(1+nl) \\ \textit{if } K-l \leq z & 1 - e^{-n(K-l)}(1+nl) + \int_{K-l}^z e^{-nu}(n^2(K-u) + 2n)du + e^{-nK}1_{\{z=K\}} \\ & = 1 - e^{-nz}(1 + n(K-z)) + e^{-nK}1_{\{z=K\}} \end{array}$$

Using a symmetry for the case $l > \frac{K}{2}$, we finally get :

$$\forall l \in [0, K] \ H_{n,l}^W(z) = 1 - e^{-nz}(1 + n(z \wedge (K-z) \wedge (K-l) \wedge l)) + e^{-nK}1_{\{z=K\}} \quad (4)$$

Mathematical expectation of Z is :

$$E_{n,l}^W[Z] = \int_0^K (1 - H_{n,l}^W(z))dz = \frac{1}{n}(2 - e^{-nl} - e^{-n(K-l)}) \quad (5)$$

On an infinite chromosome, length expectation is $\frac{2}{n}$, twice the expected length of an unselected donor genome segment. Here, $\frac{-1}{n}(e^{-nl} + e^{-n(K-l)})$ is the correction term due to chromosome finiteness; it consists of the truncation probability, e^{-nl} to the left and $e^{-n(K-l)}$ to the right, times $\frac{1}{n}$, the expectation of each truncation.

No selection : donor segment length on a finite chromosome arm without interference

Left chromosome end, as any point of the line, is copied from the donor with probability 2^{-n}. If not, it belongs to a receiver chromosome fragment, whose length z is distributed according to the density $\frac{n}{2^n-1} z f_n^W(z)$, the normalized product of the density $f_n^W(z)$ of any receiver chromosome fragment length by its length z. Conditionally on z, chromosome end is uniformly distributed on this interval.

Probability of donor disappearance : Let $q_{n,K}^W$ be the probability, the donor genome totally disappeared at generation n. Probability $1 - q_{n,K}^W$ that the chromosome of length K bears a donor genome fragment, equals the probability it starts on such a segment plus, otherwise, the probability such a segment starts before chromosome ends. Thus :

$$\begin{aligned}
1 - q^W_{n,K} &= 2^{-n} + (1-2^{-n}) \int_0^\infty \frac{n}{2^n-1} z f^W_n(z) \frac{K \wedge z}{z} dz \\
&= 2^{-n} + (1-2^{-n}) (\int_K^\infty \frac{nK}{2^n-1} f^W_n(z) dz + \int_0^K \frac{n}{2^n-1} z f^W_n(z) dz) \\
&= 2^{-n}(1 + nK(1 - F^W_n(K)) + nKF^W_n(K) - n \int_0^K F^W_n(z) dz)
\end{aligned}$$

$$1 - q^W_{n,K} = 2^{-n}(1 + n \int_0^K (1 - F^W_n(z)) dz) \qquad (6)$$

Note that if $K = 0$ then $q^W_{n,K} = 1 - 2^{-n}$, and if $K \uparrow \infty$ then $q^W_{n,K} \to 0$ as expected on an infinite chromosome.

Donor fragment length distribution : On an unlimited chromosome, the length of unselected donor fragments is distributed with probability density $g^W_n(z)$. On a finite chromosome, truncation must be taken into account. Moreover the number of donor fragments left on the chromosome is theoretically unlimited, but the length of such fragments is differently distributed according to their status. Calculation is done here for the first (or last) fragment. Four mutually exclusive situations have to be considered, whether, or not, this fragment is limited by the chromosome end on its left or on its right; donor disappearance is considered as the presence of a fragment of length zero. The length probability density $g^W_{n,K}(z)$, $z \in [0, K]$ of such an extremal donor fragment length is :

$$\begin{aligned}
g^W_{n,K}(z) &= q^W_{n,K}\delta_0(z) + 2^{-n}(\int_z^\infty n^2 x e^{-nx} \frac{1}{x} dx + \delta_K(z) \int_K^\infty n^2 x e^{-nx} \frac{(x-K)}{x} dx) \\
&+ (1-2^{-n})(\int_0^\infty \frac{n u f^W_n(u)}{2^n-1} \frac{(K-z) \wedge u}{u} n e^{-nz} du + \int_{K-z}^\infty \frac{n u f^W_n(u)}{2^n-1} \frac{1}{u} e^{-nz} du) \\
g^W_{n,K}(z) &= q^W_{n,K}\delta_0(z) + 2^{-n}(n e^{-nz} + \delta_K(z) e^{-nK} \\
&+ n^2 e^{-nz}((K-z)F^W_n(K-z) - \int_0^{K-z} F^W_n(u) du + (K-z)(1 - F^W_n(K-z)))) \\
&+ n e^{-nz}(1 - F^W_n(K-z)))
\end{aligned}$$

We finally obtain for $g^W_{n,K}(z)$, $(z \in [0, K])$

$$q^W_{n,K}\delta_0(z) + 2^{-n}e^{-nK}\delta_K(z) + 2^{-n} n e^{-nz}(2 - F^W_n(K-z) + n(\int_0^{K-z} (1 - F^W_n(u)) du)) \quad (7)$$

The cumulative distribution function $G^W_{n,K}(z) = \int_0^z g^W_{n,K}(u) du$, results from a straightforward calculus :

$$G^W_{n,K}(z) = 1 + 2^{-n}e^{-nK}1_{\{z=K\}} - 2^{-n}e^{-nz}(1 + n \int_0^{K-z} (1 - F^W_n(u)) du) \qquad (8)$$

Defining $\Phi_n(H) = n\int_0^H e^{-nx}F_n^W(H-x)dx$, mathematical expectation of these fragments, conditional on their existence, is :

$$D_{n,K}^W = \frac{\frac{1}{n}\Phi_n(K) + \int_0^K (1-F_n^W(u))du}{1+n\int_0^K (1-F_n^W(u))du} \tag{9}$$

Note that Φ_n is the length cumulative distribution function of the concatenation of a donor and a receiver fragment. It tends to 1 as K tends to infinity, and $D_{n,K}^W$ to $\frac{1}{n}$ as expected on an infinite chromosome.

Probability of donor multiple occurrences : Let $\sigma_{n,K}^W$ be the probability there are two or more donor genome fragments left on the chromosome.

$$\begin{aligned}
\sigma_{n,K}^W \quad &= 2^{-n}\int_0^K ne^{-nx}F_n^W(K-x)dx \\
&+(1-2^{-n})\int_0^\infty \int_0^{K\wedge z}\int_0^{K-y} \frac{nzf_n^W(z)}{2^n-1}\frac{1}{z}ne^{-nx}F_n^W(K-y-x)dxdydz \\
= \quad & n2^{-n}\int_0^K e^{-nx}F_n^W(K-x)dx + n^2 2^{-n}(\int_0^K f_n^W(z)\int_0^z\int_0^{K-y} e^{-nx}F_n^W(K-y-x)dxdydz \\
&+\int_K^\infty f_n^W(z)\int_0^K\int_0^{K-y} e^{-nx}F_n^W(K-y-x)dxdydz)
\end{aligned}$$

Integrating by parts, with $\Phi_n(H) = n\int_0^H e^{-nx}F_n^W(H-x)dx$, we have :

$$\int_0^K f_n^W(z)\int_0^z \Phi_n(K-y)dydz = F_n^W(K)\int_0^K \Phi_n(K-y)dy - \int_0^K F_n^W(z)\Phi_n(K-z)dz$$

$$\text{and also, } \int_K^\infty f_n^W(z)\int_0^K \Phi_n(K-y)dydz = (1-F_n^W(K))\int_0^K \Phi_n(K-y)dy$$

then finally,

$$\sigma_{n,K}^W = 2^{-n}(\Phi_n(K) + n\int_0^K (1-F_n^W(z))\Phi_n(K-z)dz) \tag{10}$$

Crossing-over with complete interference

In this section, complete interference is assumed : at each generation, a unique cross-over occurs on each chromosome arm. By definition of genetic distances, the length of any chromosome arm is 1 and the cross-over is uniformly distributed on it. Each process X^i reduces to one point. Let l be the position, on the chromosome arm, of the gene to be introgressed (on locus L). The probability for the cross-over to occur between the centromer and the locus L is l.

The point process, X^+, from the centromer to the end of the chromosome arm is constituted by exactly n independent uniformly distributed points; each mark appears once. Because of selection, the locus of interest, L, is always copied from the donor; if a cross-over occurs before (respectively, after) L, all the segment on its left (respectively, right) is copied from the receiver. The interval which bears L, is the only one copied from the donor. It is limited on its right by B_r, the first point of X^+ after L, or 1, if all cross-overs occured before L. Similarly, it is limited on its left by B_l, the last point of X^+ before L, or 0, if all cross-overs occured after L.

Introgression : length of the segment bearing the selected gene on a finite chromosome arm with complete interference

The number N of cross-overs having occured before L has a binomial distribution with parameters n and l. Given N, cross-overs are distributed uniformly either on $[0, l]$ or $[l, 1]$, and independent. We have :

$$\begin{aligned} P[N = k] = & \quad \tbinom{n}{k} l^k (1-l)^{n-k} \\ P[B_l = 0] = & \quad (1-l)^n \ (then\ N = 0) \\ P[0 < B_l < x | N = k] = & \quad (x/l)^k, \forall k > 0 \\ P[l < B_r < y | N = k] = & \quad 1 - ((1-y)/(1-l))^{n-k}, \forall k < n \\ P[B_r = 1] = & \quad l^n \ (then\ N = n) \end{aligned}$$

Let $p_n^C(x, y)$ be the joint probability density of B_l and B_r. $p_n^C(x, y) =$

$$\begin{aligned} &(1-l)^n \delta_0(x) \tfrac{n}{1-l} (\tfrac{1-y}{1-l})^{n-1} + \textstyle\sum_{k=1}^{n-1} \tbinom{n}{k} l^k (1-l)^{n-k} \tfrac{k(n-k)}{l(1-l)} (\tfrac{x}{l})^{k-1} (\tfrac{1-y}{1-l})^{n-k-1} + l^n \delta_1(y) \tfrac{n}{l} (\tfrac{x}{l})^{n-1} \\ &= n\delta_0(x)(1-y)^{n-1} + \textstyle\sum_{k=1}^{n-1} \tfrac{n!}{(k-1)!(n-k-1)!} x^{k-1} (1-y)^{n-k-1} + n\delta_1(y) x^{n-1} \\ &= n\delta_0(x)(1-y)^{n-1} + n(n-1)(1-y+x)^{n-2} + n\delta_1(y) x^{n-1} \end{aligned}$$

which leads to the joint cumulative distribution function of B_l and B_r (with $0 \leq x \leq l \leq y \leq 1$) :

$$P^C[B_l \leq x \ and\ B_r \leq y] = (1-l+x)^n - (1-y+x)^n + x^n 1_{\{y=1\}}$$

Let Z be the length of the interval $[B_l, B_r]$. Its density of probability $h_{n,l}^C(z)$, on $[0, 1]$, is $\int_{0 \vee (l-z)}^{l \wedge (1-z)} p_n(x, x+z) dx$, which leads to :

$$h_{n,l}^C(z) = n(1-z)^{n-1}(1_{\{z>l\}} + 1_{\{z>1-l\}}) + n(n-1)(1-z)^{n-2}(l \wedge (1-z) - 0 \vee (l-z)) \quad (11)$$

Let $H_{n,l}^C(z) = \int_0^z h_{n,l}^C(u) du = P[Z \leq z]$ be the cumulative distribution function of Z; if $l < \frac{1}{2}$, then

$$\begin{aligned} &if\ z \leq l && H_{n,l}^C(z) = 1 - (1-z)^{n-1}(1 + (n-1)z) \\ &if\ l \leq z \leq 1-l && H_{n,l}^C(z) = 1 - (1-z)^{n-1}(1 + nl - z) \\ &if\ 1-l \leq z && H_{n,l}^C(z) = 1 - (1-z)^n(n+1) \end{aligned}$$

Using a symmetry for the case $l > \frac{1}{2}$, we finally get :

$$\forall l \in [0, 1] \ \ H_{n,l}^C(z) = 1 - (1-z)^{n-1}(1 - z + n(z \wedge (1-z) \wedge (1-l) \wedge l)) \quad (12)$$

Mathematical expectation of Z is $E_{n,l}^C[Z] = \int_0^1 (1 - H_{n,l}^C(z)) dz$

$$E_{n,l}^C[Z] = \frac{1}{n+1}(2 - l^{n+1} - (1-l)^{n+1}) \quad (13)$$

No selection : Unselected donor segment length on a finite chromosome arm with complete interference

There is zero or one segment copied from the donor. For such a segment to exist, receiver chromosome must be copied at each generation in such a way that one of the $n+1$ intervals defined by the process X^+ on the chromosome remains of the donor type. Given an interval, the probability it belongs to the donor type is 2^{-n} ; probability, $1 - q_n^C$, that a segment of the donor chromosome remains after n generations is $(n+1)2^{-n}$. Expected length of such an interval (provided it exists) is $D_n^C = \frac{1}{n+1}$ (there are always $n+1$ intervals). Probability distribution of these segments, conditional on their existence, is the one of the intervals in the process X^+, with cumulative distribution function $1-(1-z)^n$, and density of probability $n(1-z)^{n-1}$. Their unconditioned cumulative distribution function is $G_n^C(z) = 1 - (1-q_n^C)(1-z)^n$, and their density is $g_n^C(z) = q_n^C \delta_0(z) + (1-q_n^C)n(1-z)^{n-1}$.

Comparison of both crossing-over models

In order to compare both crossing-over models, no interference or complete interference, we must consider similar situations. Model without interference will therefore be considered, in this section, on a chromosome arm of length $K = 1$ Morgan, with, eventually, a selected locus L at l.

The major qualitative difference is certainly that, if interference is complete, there is a unique donor chromosome segment in an introgression scheme (at most one, if there is no selection), whereas, without interference, donor genome has a positive probability to reappear, and the number of such fragments is theoretically unlimited. We already noted (equation 2) that, even after a large generation number, recurrence probability of a very short receiver fragment remains almost constant. But somewhat surprisingly, this seems to be of little relevance in practice, as shown by numerical results (fig. 3 and 4, fig. 5, fig. 6, fig. 7, fig. 9 and fig. 10).

Otherwise, qualitative and quantitative differences between both models are rather small. Selected genes are beared by fragments, which are, in the average and not taking into account shortening due to truncation, twice as long as fragments obtained without selection, provided they exist.

Examples of use of these computations

When designing a backcross, generally one has first to define the desired maximal size of the introgressed segment. Computations presented in this article allow to manage this step : length cumulative distribution function of the donor segment bearing the selected gene can be easily computed for any chromosome length, for any generation number and for different crossing-over models (equations 4, 12 and figures 3, 4) with its mathematical expectation (equations 5, 13 and figure 5). For example, for a gene to be introgressed standing on the middle of a 1 Morgan long chromosome and assuming no interference, after 5 generations of backcrossing, the donor segment carrying the selected gene has probability .80 to be smaller than

55 cM (one half to be less than 34 cM) (see table 1 and figure 3). The gain per generation can be measured by the length reduction for a given probability (here, about 10 cM between generations 5 and 6). Similar computations can also be made for a model with complete interference (see Table 2 and figure 4).

Conversely the number of generations needed to achieve a given length reduction can also be computed. For example, if a length less than 50 cM is wanted, 4 generations are needed to ensure this happens with a probability of at least .5 (see Tables 1 and 2).

Another characteristic important to be defined in a backcross design, is the proportion of donor fragments tolerated in unselected chromosomes. Whereas some publications only study the structure of the selected chromosome (Frish and Melchinger 2001, Hospital 2001), we here also provide formulae for the distribution of donor fragment length and number on unselected chromosomes. Hence, in addition to the control of the selected chromosome, receiver genome return on other chromosomes can also be controlled. For any generation number, we can compute (i) the probability of donor disappearance (equations 6, subsection and figure 7), (ii) the probability of donor multiple occurrences in a model without interference (multiple occurences cannot occur with complete interference) (equation 10 and figure 8), (iii) the cumulative distribution function of unselected donor segment length (equation 8, and figures 9, 10), (iv) the mathematical expectation of unselected donor segment length, conditional on their existence (equation 9, subsection and figure 6).

Discussion

We presented here a general and complete description of the mosaic chromosome structure, which requires only simple computations ; in particular, no numerical integration is needed. Equations provided here make possible the computation of the probability of any event of interest. Moreover, they can be adapted to different crossing-over models.

Computations presented in this article allow to design introgression programs as they can provide the probability of any feature on selected and unselected chromosomes after any generation number. Before beginning an introgression program, the user is able to find in a few seconds an optimal compromise between objectives and financial means.

Many publications presented similar results on marker assisted backcrossing, i.e. probabilities to have a given size of donor segments conditional on the genotype at markers (Hill 1993, Hospital 2001, Frish and Melchinger 2001, Servin et al 2002, Frish and Melchinger 2005). Here, computations are as easy as the published ones but are unconditionned. Hence, in a first step, computation with equations presented here can be made to evaluate the expected structure of the genome and to help to choice marker positions for a marker assisted backcrossing.

Such a description, in terms of marked point processes, could be extended to other crosses and provide a fruitfull point of view.

It is also of larger concern, for instance in population genetics. In any population, a chromosome can be considered as a mosaic of fragments inherited from the ancestors n generations backward. Each of these fragments is inherited through one of the 2^n different possible paths, and its length is distributed as in a back-cross after n generations. In a population, chromosomes bearing a point mutation remain identical to the original chromosome, where the mutation occured, all along a segment, which is one such fragment, or several of them, owing to the fact that contiguous fragments can be eventually copied from the same ancestor. Haplotype determination in a sample provides a dating tool : the shorter the conserved haplotype, the higher the number of generations. But, owing to possible recombinations in the sample coalescent, a correct description requires to take into account possible recombinations between copies of the ancestral chromosome. Theoretical distributions as provided here, can help building dating confidence intervals.

For conservation policies directed toward small endangered populations, it is crucial to maintain a sufficiently large genetic diversity. Similar calculations, provided the genealogy is known, make possible to estimate which proportion of a given ancestor genome is still present in the population.

Appendix : Ehrenfest promenade

Consider two urns containing n balls in total. Each time, a ball is chosen at random and moved to the other urn. The Ehrenfest promenade of parameter n is the number of balls in a given urn. It is a Markov chain with period 2 (odd,even), therefore return times are even. An Ehrenfest promenade is ergodic and all return times have finite expectations.

Zero return time distribution

The Ehrenfest promenade has the following transition matrix :

$$P_n = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \frac{1}{n} & 0 & \frac{n-1}{n} & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \frac{2}{n} & 0 & \frac{n-2}{n} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \frac{3}{n} & 0 & \frac{n-3}{n} & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \cdots & \frac{n-1}{n} & 0 & \frac{1}{n} \\ 0 & \cdots & \cdots & \cdots & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

Binomial distribution with parameter $(n, 1/2)$, is the stationnary distribution. Zero return time expectation is the inverse of the stationnary probability of 0, hence equals 2^n. Because of periodicity, zero return time can be calculated from the chain sampled at even instants, with a state space reduced to even values. The transition matrix to be considered is (in this example with an ending corresponding to n even) :

$$M_n = \begin{bmatrix} \frac{1}{n} & \frac{n-1}{n} & 0 & \cdot & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \frac{2}{n^2} & \frac{5n-8}{n^2} & \frac{n^2-5n+6}{n^2} & \cdot & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & \cdot & \frac{i(i-1)}{n^2} & \frac{n+2i(n-i)}{n^2} & \frac{(n-i)(n-i-1)}{n^2} & \cdot & 0 & 0 \\ \cdots & 0 & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdots & 0 & \cdot & 0 & \cdots & \cdots & 0 & \frac{n-1}{n} & \frac{1}{n} \end{bmatrix}$$

Index i is used here for the row corresponding to row i of P_n^2, whose rows and columns are numbered as the states of the process, from 0 to n. Row named i is hence the $(\frac{i}{2}+1)-th$ row of M_n. The number of rows or columns of M_n, $m+1$, is the number of even states for the Ehrenfest promenade ; $m = \frac{n-1}{2}$ if n is odd, $m = \frac{n}{2}$ if n is even. M_n is tridiagonal and can be written as :

$$M_n = \begin{bmatrix} \alpha & v \\ u & A \end{bmatrix}$$

with $\alpha = 1/n$, $u = (2/n^2, 0, \cdots 0)'$, $v = ((n-1)/n, 0, \cdots 0)$.
Zero return time probabilities are given by : $\pi_2^{(n)} = 1/n$, $\pi_{2i}^{(n)} = vA^{i-2}u$, $i \geq 2$.

Appendix : Deriving a computable formula for F_n^W

A general method to compute $\pi_{2i}^{(n)} = vA^{i-2}u$ starts with a diagonalization of matrix A. A is real, nonnegative and tridiagonal, hence has real eigenvalues ; A is strictly submarkovian, hence has eigenvalues of modulus strictly less than one (Perron-Frobenius theorem). A very effective algorithm exists for the numerical diagonalization of tridiagonal matrices (Horn R. A. 1990).

Let ω_k, $k = 1, \ldots m$ be a complete set of linearly independent left-eigenvectors of A, with corresponding eigenvalues d_k. Let $c_1' = (1, 0, \ldots 0)$, then, since $\omega_1, \ldots \omega_k$ are linearly independent, there is a unique set of values $\gamma_1, \ldots \gamma_k$ such that $c_1' = \sum_{k=1}^m \gamma_k \omega_k$. Then $vA^i u = \frac{n-1}{n} \sum_{k=1}^m \gamma_k \omega_k A^i c_1 \frac{2}{n^2}$. Let $\beta_k = \frac{2(n-1)}{n^3} \gamma_k < \omega_k, c_1 >$, then $vA^i u = \sum_{k=1}^m \beta_k d_k^i$, $i \geq 0$.

Using these considerations and equation 1, $F_n^W(z)$ can now be written as a finite linear combination of exponentials, with known coefficients and parameters.

$$
\begin{aligned}
F_n^W(z) &= 1 - e^{-nz} \sum_{i=1}^{\infty} \pi_{2i}^{(n)} \sum_{j=0}^{2(i-1)} \frac{(nz)^j}{j!} \\
&= 1 - e^{-nz} (\pi_2^{(n)} + \sum_{i=2}^{\infty} \pi_{2i}^{(n)} (1 + \sum_{j=1}^{2(i-1)} \frac{(nz)^j}{j!})) \\
&= 1 - e^{-nz} (1 + \sum_{i=2}^{\infty} \pi_{2i}^{(n)} \sum_{j=1}^{2(i-1)} \frac{(nz)^j}{j!}) \\
&= 1 - e^{-nz} (1 + \sum_{i=0}^{\infty} vA^i u \sum_{j=1}^{2(i+1)} \frac{(nz)^j}{j!}) \\
&= 1 - e^{-nz} (1 + \sum_{i=0}^{\infty} \sum_{k=1}^{m} \beta_k d_k^i \sum_{j=1}^{2(i+1)} \frac{(nz)^j}{j!}) \\
&= 1 - e^{-nz} (1 + \sum_{k=1}^{m} \beta_k \sum_{i=0}^{\infty} d_k^i \sum_{j=1}^{2(i+1)} \frac{(nz)^j}{j!}) \\
&= 1 - e^{-nz} (1 + \sum_{k=1}^{m} \beta_k \sum_{i=0}^{\infty} d_k^i \sum_{j=1}^{(i+1)} (\frac{(nz)^{2j}}{(2j)!} + \frac{(nz)^{2j-1}}{(2j-1)!})) \\
&= 1 - e^{-nz} (1 + \sum_{k=1}^{m} \beta_k \sum_{j=1}^{\infty} (\frac{(nz)^{2j}}{(2j)!} + \frac{(nz)^{2j-1}}{(2j-1)!}) \sum_{i=j-1}^{\infty} d_k^i) \\
&= 1 - e^{-nz} (1 + \sum_{k=1}^{m} \beta_k \sum_{j=1}^{\infty} d_k^{j-1} (\frac{(nz)^{2j}}{(2j)!} + \frac{(nz)^{2j-1}}{(2j-1)!}) \frac{1}{1-d_k}) \\
&= 1 - e^{-nz} (1 + \sum_{k=1}^{m} \frac{\beta_k}{d_k(1-d_k)} \sum_{j=1}^{\infty} (\frac{(\sqrt{d_k} nz)^{2j}}{(2j)!} + \sqrt{d_k} \frac{(\sqrt{d_k} nz)^{2j-1}}{(2j-1)!})) \\
&= 1 - e^{-nz} (1 + \sum_{k=1}^{m} \beta_k \frac{-2 + e^{\sqrt{d_k} nz} + e^{-\sqrt{d_k} nz} - \sqrt{d_k}(e^{-\sqrt{d_k} nz} - e^{\sqrt{d_k} nz})}{2 d_k (1-d_k)}) \\
&= 1 - e^{-nz} (1 - \sum_{k=1}^{m} \frac{\beta_k}{d_k(1-d_k)}) - e^{-nz} \sum_{k=1}^{m} \frac{\beta_k}{2} (\frac{1+\sqrt{d_k}}{d_k(1-d_k)} e^{\sqrt{d_k} nz} + \frac{1-\sqrt{d_k}}{d_k(1-d_k)} e^{-\sqrt{d_k} nz})
\end{aligned}
$$

We finally get : $F_n^W(z) =$

$$
1 - (1 - \sum_{k=1}^{m} \frac{\beta_k}{d_k(1-d_k)}) e^{-nz} - \sum_{k=1}^{m} \frac{\beta_k}{d_k(1-d_k)} (\frac{1+\sqrt{d_k}}{2} e^{-nz(1-\sqrt{d_k})} + \frac{1-\sqrt{d_k}}{2} e^{-nz(1+\sqrt{d_k})}) \tag{14}
$$

Moments of F_n^W can be derived from (14). For instance $E[Z] = \frac{1}{n}(1 + \sum_{k=1}^{m} \frac{2\beta_k}{(1-d_k)^2})$ (yet known to equal $(2^n - 1)/n$), and $E[Z^2] = \frac{2}{n^2}(1 + \sum_{k=1}^{m} \frac{\beta_k(5-d_k)}{(1-d_k)^3})$

Numerical examples

For $n = 1$, $\pi_2^{(1)} = 1$. Hence : $F_1^W(z) = 1 - e^{-z}$, with mean 1 and variance 1.

For $n = 2$,

$$M_2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

$\pi_{2j}^{(2)} = 2^{-j}$, $j > 0$ (remember $\pi_0^{(2)} = 0$). Straightforward calculation leads to :
$F_2^W(z) = 1 - \frac{2+\sqrt{2}}{4}e^{-(2-\sqrt{2})z} - \frac{2-\sqrt{2}}{4}e^{-(2+\sqrt{2})z}$, with mean 3/2 and variance 11/4.

For $n = 3$,

$$M_3 = \begin{bmatrix} 1/3 & 2/3 \\ 2/9 & 7/9 \end{bmatrix}$$

$\pi_2^{(3)} = 1/3$, $\pi_{2j}^{(3)} = (4/27) \times (7/9)^{j-2}$, $j > 0$.
$F_3^W(z) = 1 - \frac{1}{7}e^{-3z} - \frac{3+\sqrt{7}}{7}e^{-(3-\sqrt{7})z} - \frac{3-\sqrt{7}}{7}e^{-(3+\sqrt{7})z}$, with mean 7/3 and variance 67/9.

For $n = 4$,

$$M_4 = \begin{bmatrix} 1/4 & 3/4 & 0 \\ 1/8 & 3/4 & 1/8 \\ 0 & 3/4 & 1/4 \end{bmatrix}$$

eigenvalues d_1, d_2 and left-eigenvectors ω_1, ω_2 of A are :
$d_1 = \frac{4+\sqrt{10}}{8}$, $\omega_1 = (2 + \sqrt{10}, 1)$, and $d_2 = \frac{4-\sqrt{10}}{8}$, $\omega_2 = (2 - \sqrt{10}, 1)$,
$v = \frac{3}{4\times 2\sqrt{10}}(\omega_1 - \omega_2)$, $< \omega_1, u >= \frac{2+\sqrt{10}}{8}$, $< \omega_2, u >= \frac{2-\sqrt{10}}{8}$.
$\beta_1 = \frac{15+3\sqrt{10}}{320}$, $\beta_2 = \frac{15-3\sqrt{10}}{320}$, and $\sum_{k=1}^{2} \frac{\beta_k}{d_k(1-d_k)} = 1$,
$F_4^W(z) = 1 - \frac{5+\sqrt{10}}{20}(1 + \frac{\sqrt{4+\sqrt{10}}}{2\sqrt{2}})e^{-(4-\sqrt{8+2\sqrt{10}})z} - \frac{5+\sqrt{10}}{20}(1 - \frac{\sqrt{4+\sqrt{10}}}{2\sqrt{2}})e^{-(4+\sqrt{8+2\sqrt{10}})z}$
$- \frac{5-\sqrt{10}}{20}(1 + \frac{\sqrt{4-\sqrt{10}}}{2\sqrt{2}})e^{-(4-\sqrt{8-2\sqrt{10}})z} - \frac{5-\sqrt{10}}{20}(1 - \frac{\sqrt{4-\sqrt{10}}}{2\sqrt{2}})e^{-(4+\sqrt{8-2\sqrt{10}})z}$,
with mean 15/4 and variance 973/48.

References

- Fisher R. 1949 Theory of junctions in inbreeding, in *Theory of inbreeding* (120 p.)
- Hanson W. D. 1959, Early generation analysis of lengths of heterozygous with backcrossing or selfing. *Genetics* 44, 833-837
- Stam P. and Zeven C. 1981. The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* 30, 227-238
- Horn R. A. 1990 Matrix analysis *Cambridge University Press* (560 p.)

- Naveira H. and Barbadilla A. 1992. The theoretical distribution of lengths of intact chromosome segments around a locus held heterozygous with backcrossing in a diploid species, *Genetics* 130 :205-209.
- Hill W. 1993. Variation in the genetic composition in backcrossing programs. *Journal of Heredity* 84 :212-213.
- Frish M. and Melchinger A. E. 2001. The length of the intact donor chromosome segment around a target gene in marker-assisted backcrossing, *Genetics* 157 :1343-1356.
- Frish M. and Melchinger A. E. 2005. Selection theory for marker-assisted backcrossing, *Genetics* 170 :909-917.
- Hospital F. 2001. Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker- assisted backcross programs. *Genetics* 158 :1363-1379.
- Servin B., Dillmann C., Decoux G., and Hospital F. 2002. Mdm : a program to compute fully informative genotype frequencies in complex breeding schemes. *Journal of Heredity* 93 :227-228.

n	x such as $P(Z < x) = 80\%$	x such as $P(Z < x) = 50\%$	$P(Z < 0.5)$	$P(Z < 0.3)$
1	$\simeq 1$	0.84	0.09	0.04
2	0.89	0.63	0.26	0.12
3	0.73	0.53	0.44	0.23
4	0.63	0.42	0.59	0.34
5	0.55	0.34	0.71	0.44

TAB. 1 – Probability of the length Z of the donor segment carrying the introgressed gene on a chromosome of length $K = 1$ M, without interference, at different backcrossing generation numbers (n). Introgressed gene lies at $l = 0.5$ M.

n	x such as $P(Z < x) = 80\%$	x such as $P(Z < x) = 50\%$	$P(Z < 0.5)$	$P(Z < 0.3)$
1	0.90	0.75	0	0
2	0.74	0.59	0.25	0.09
3	0.63	0.49	0.50	0.22
4	0.55	0.39	0.69	0.35
5	0.49	0.31	0.81	0.47

TAB. 2 – Probability of the length Z of the donor segment carrying the introgressed gene on a chromosome of length $K = 1$ M, with complete interference, at different backcrossing generation numbers (n). Introgressed gene lies at $l = 0.5$ M.

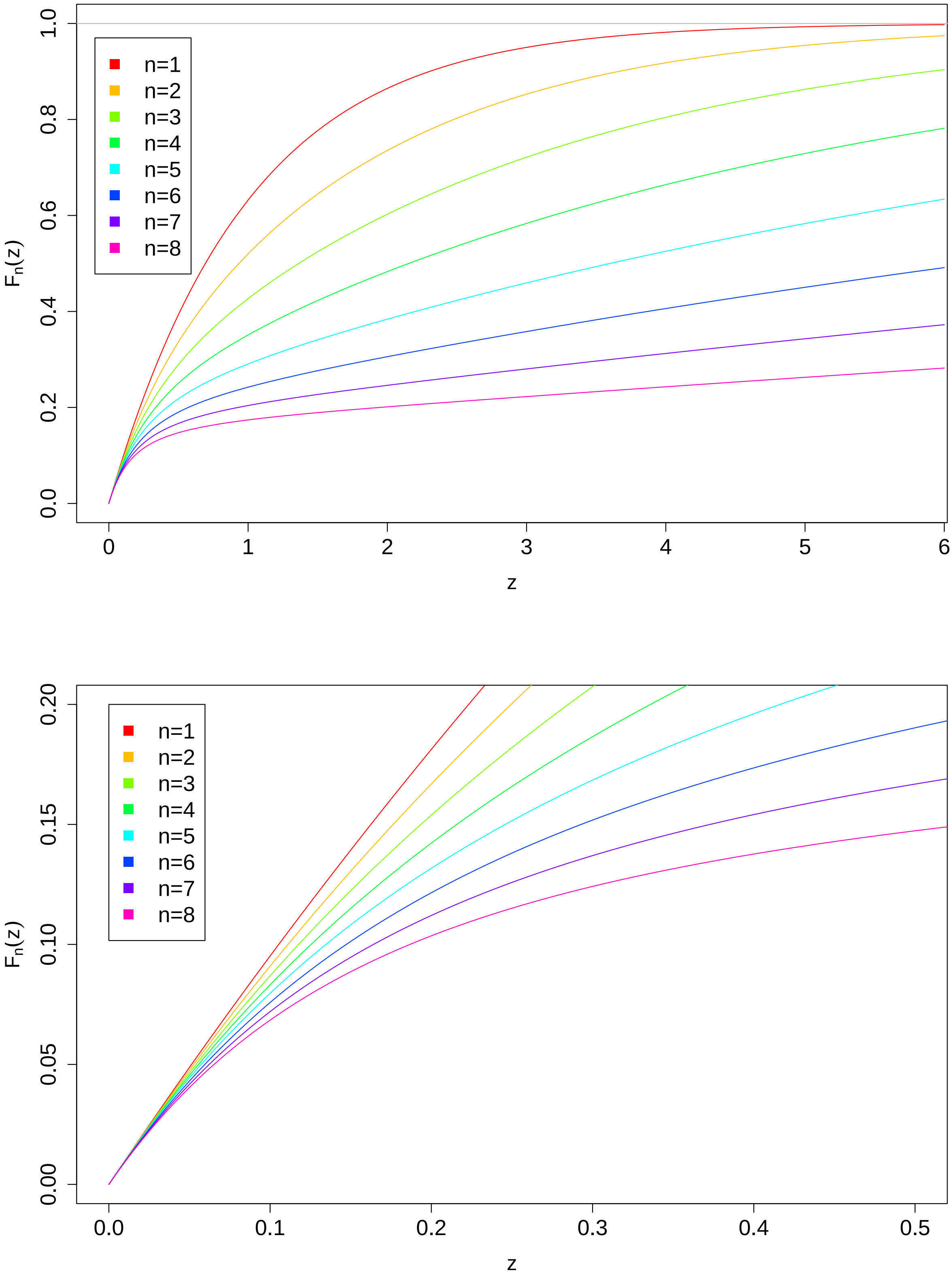

Fig. 2 – Receiver chromosome segment length cumulative distribution function, for different back-cross generation numbers, n, on an unlimited chromosome, without interference. Abscissa is in Morgan units. Lower figure is a zoom near the origin.

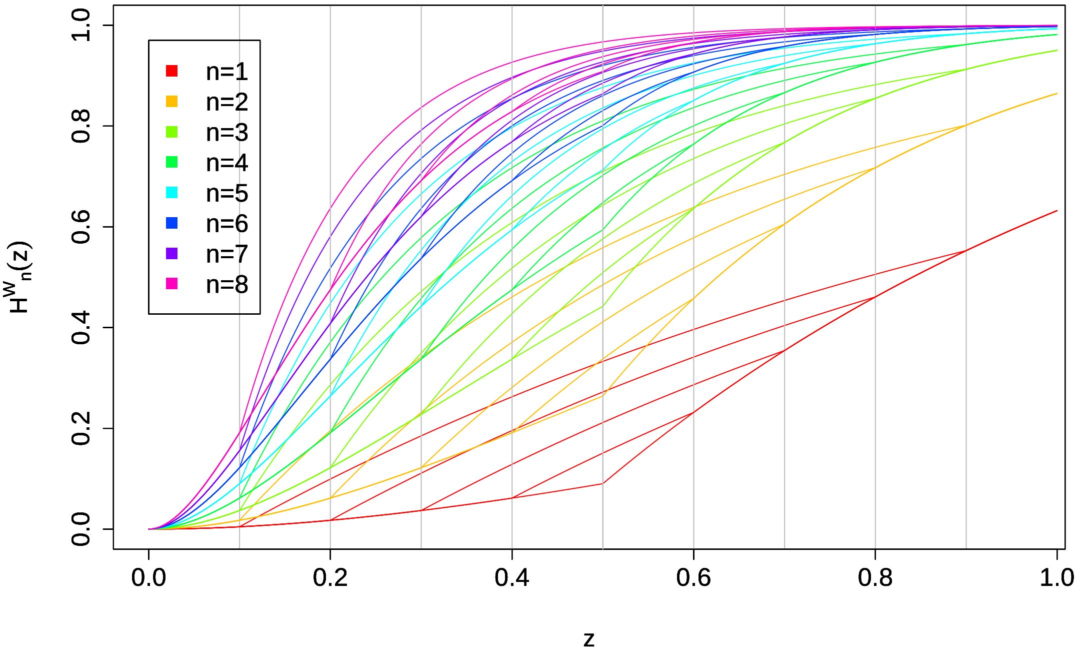

FIG. 3 – Donor chromosome segment length cumulative distribution functions, for different positions of the selected gene, and different back-cross generation numbers, n, on a chromosome of length 1, without interference. Abscissa is in Morgan units.

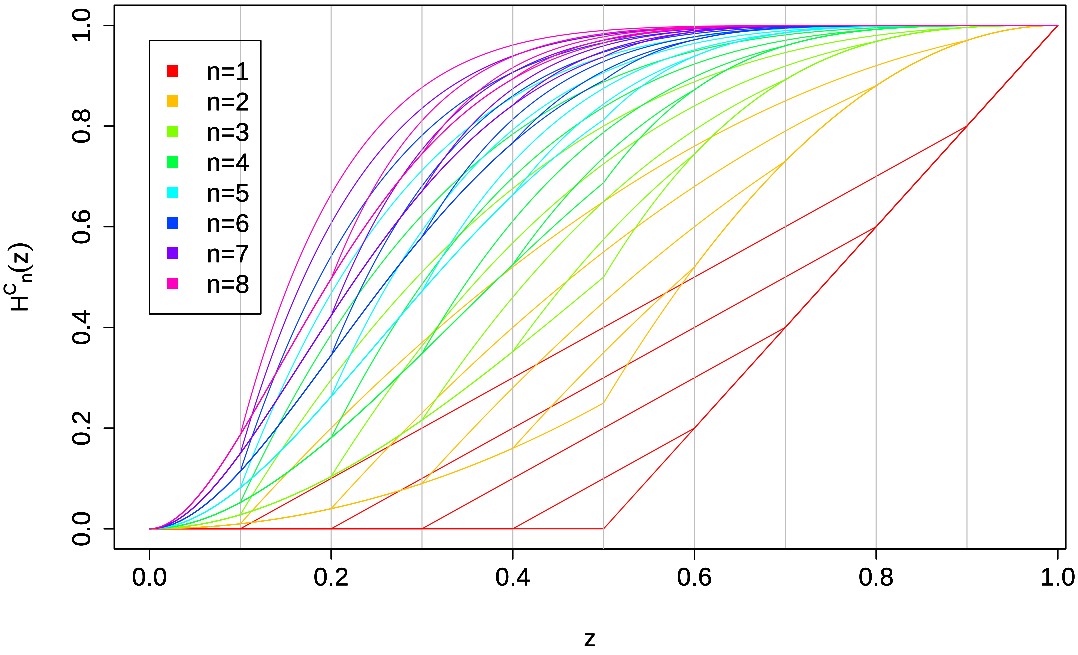

FIG. 4 – Ibidem, with complete interference.

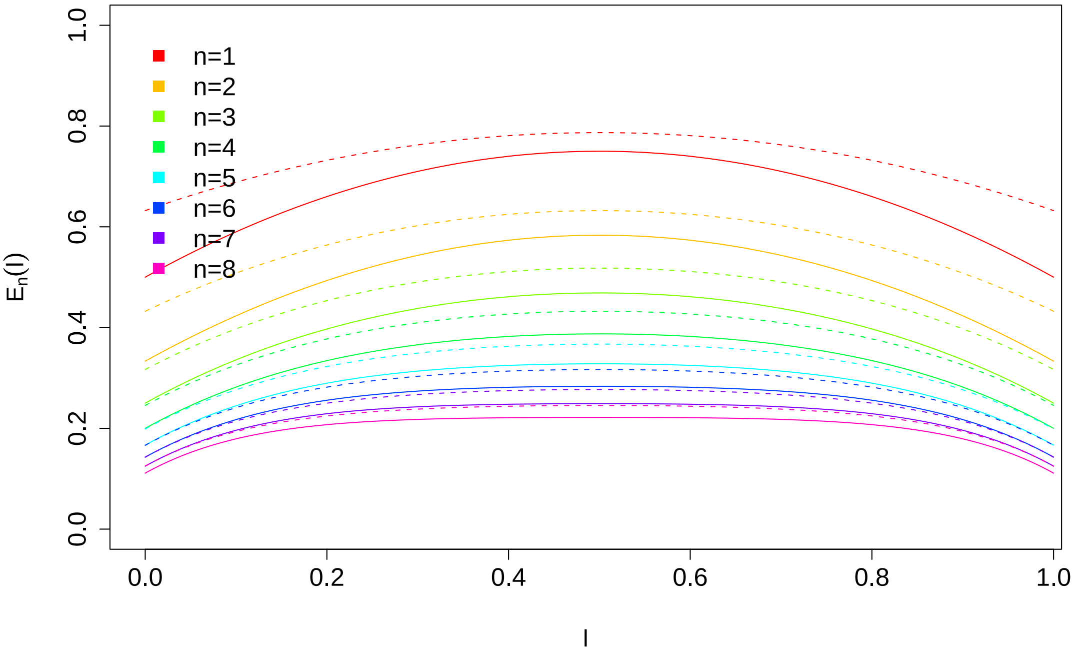

FIG. 5 – Donor chromosome segment length expectation, for different positions of the selected gene, and different back-cross generation numbers, n, on a chromosome of length 1, without interference. Abscissa is in Morgan units. Continuous line : without interference ; interruted line : complete interference.

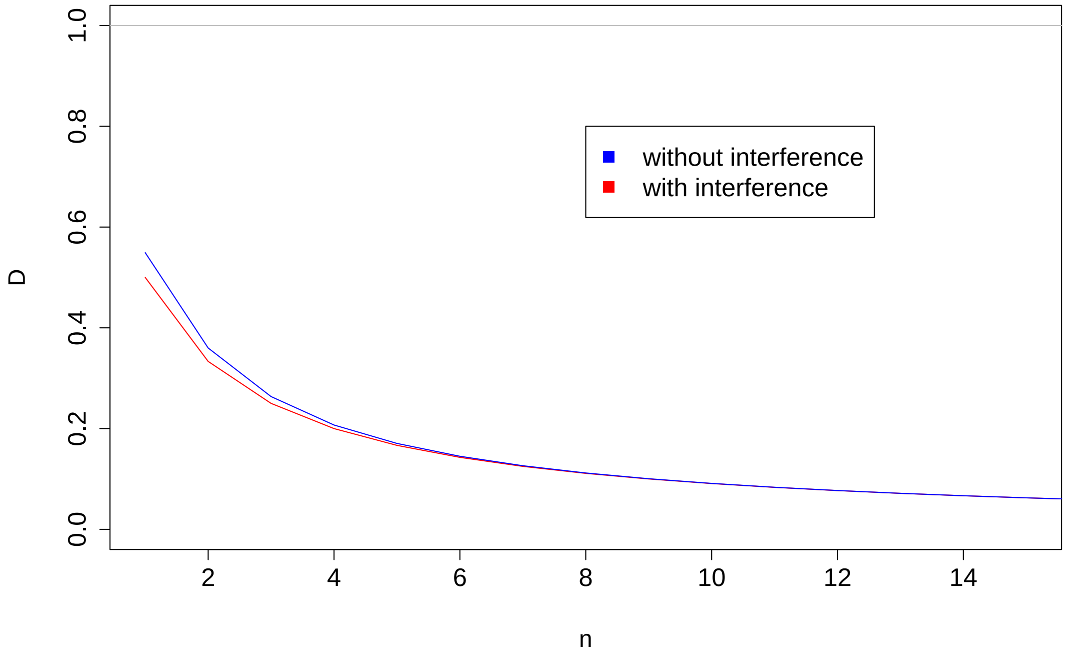

FIG. 6 – Donor chromosome segment length expectation, conditional on its presence, after n back-cross generations without selection.

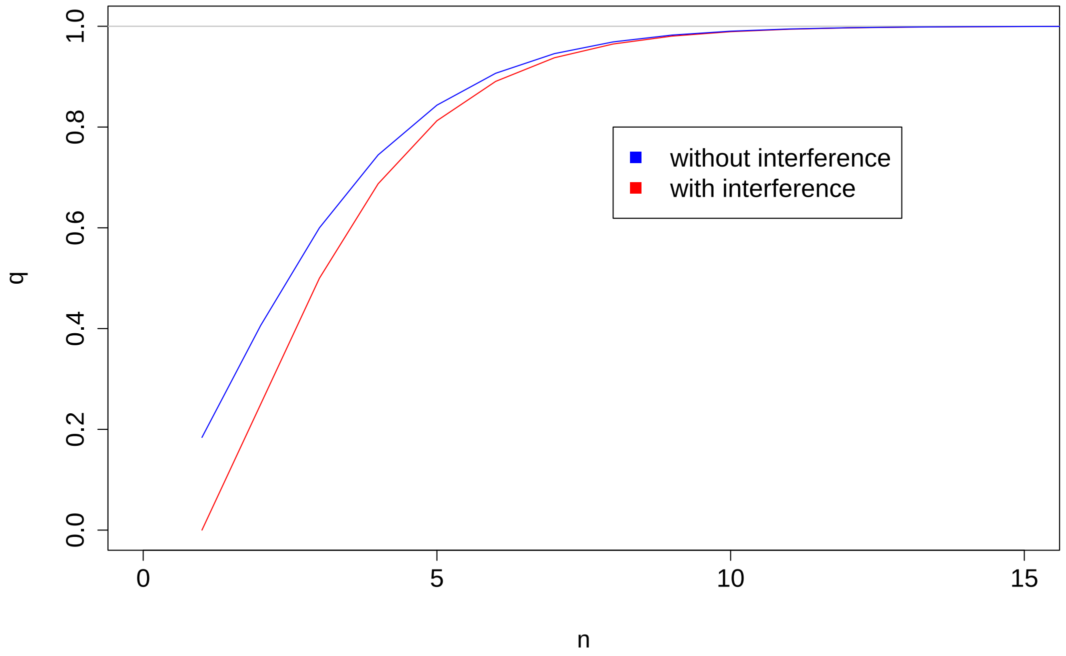

FIG. 7 – Probability of donor disappearence on a chromosome of length 1, in a back-cross without selection, according to generation number, n.

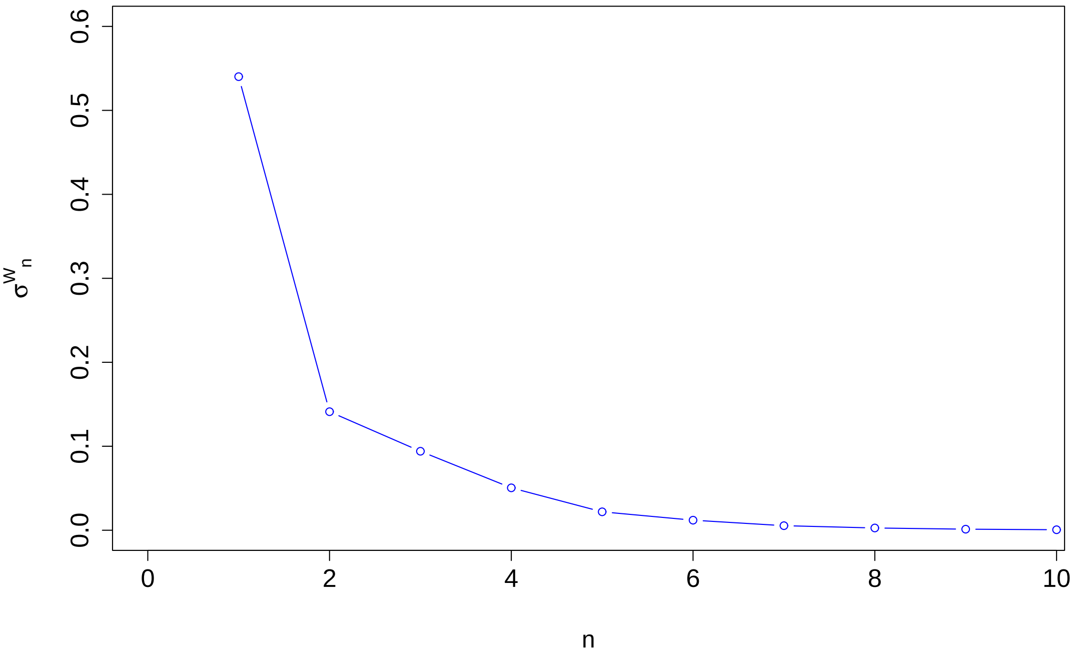

FIG. 8 – Probability of donor multiple occurences on a chromosome of length 1 without interference, in a back-cross without selection, according to generation number, n.

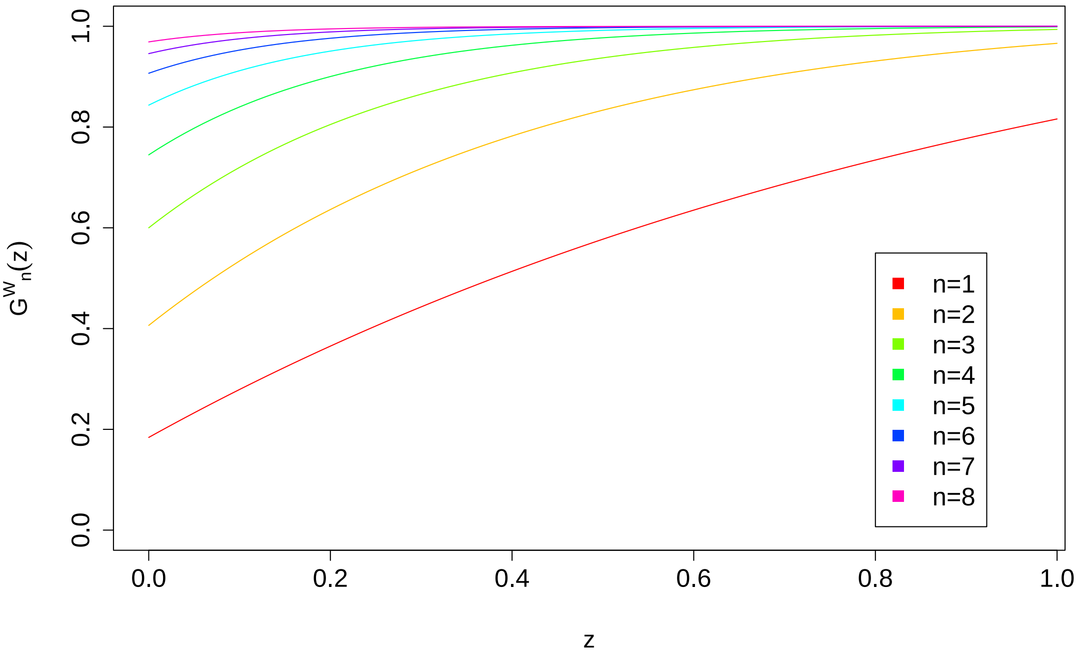

FIG. 9 – Length cumulative distribution functions of the first (or last) donor chromosome segment, without selection, on a chromosome of length 1, without interference. n is the number of generations

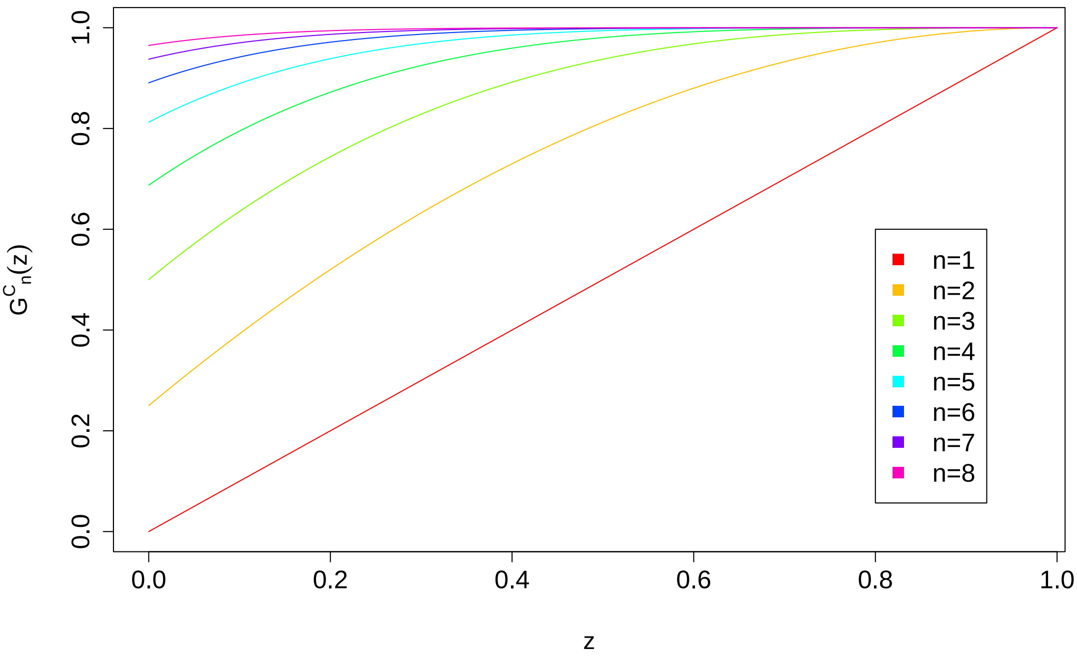

FIG. 10 – Ibidem, with complete interference.