Journal of Computational Biology
Volume 4, Number 1, 1997
Mary Ann Liebert, Inc.
Pp. 61–82

COVERAGE PROCESSES IN PHYSICAL MAPPING BY ANCHORING RANDOM CLONES.

Sophie Schbath[1]

Running head: Coverage processes in physical mapping.

Abstract

The aim of this paper is to provide general results for predicting progress in a physical mapping project by anchoring random clones, when clones and anchors are not homogeneously distributed along the genome. A complete physical map of the DNA of an organism consists of overlapping clones spanning the entire genome. Several schemes can be used to construct such a map, depending on the way that clones overlap. We focus here on the approach consisting of assembling clones sharing a common random short sequence called an anchor. Some mathematical analyses providing statistical properties of anchored clones have been developed in the stationary case. Modeling the clone and anchor processes as non-homogeneous Poisson processes provides such an analysis in a general non-stationary framework. We apply our results to two natural non-homogeneous models to illustrate the effect of inhomogeneity. This study reveals that using homogeneous processes for clones and anchors provides an overly optimistic assessment of the progress of the mapping project.

Key words: Anchoring method, coverage processes, hotspots, non-homogeneous Poisson processes, physical mapping.

1 Introduction

A complete physical map of the DNA of an organism consists of ordered overlapping fragments spanning the entire genome. Such a map is a useful tool for further genetic analyses, like gene location or sequencing of particular parts of the genome. A large number of fragments, called *clones*, are chosen at random from a library in which the entire genome is represented. Several approaches can be used to infer the overlap of these clones. The basic approach is to describe each clone by a "fingerprint" reflecting partial information about the clone, and then to infer the overlaps of clones when the fingerprints are very similar. Physical maps were constructed in that way for *Saccharomyces cerevisiae* (Olson *et al.*, 1986), *Caenorhabditis elegans* (Coulson *et al.*, 1986) and *Escherichia coli* (Kohara *et al.*, 1987). The anchoring approach, efficient for large clones and then for very large genomes, uses an additional random genomic library of *anchors* and is based on the anchor-content of the clones. These anchors consist of any short sequence of DNA

[1]I.N.R.A., Unité de Biométrie, 78352 Jouy-en-Josas, France. E-mail: Sophie.Schbath@jouy.inra.fr

that occurs exactly once in the genome. Thus, clones containing an anchor in common overlap. Clones that overlap are then assembled into *islands* which cover some regions of the genome. Physical maps of the human chromosome 21q (Chumakov *et al.*, 1992), the human Y chromosome (Foote *et al.*, 1992) and recently of the entire human genome (Hudson *et al.*, 1995) have been constructed using this approach. To plan a physical mapping project, it is therefore important to study, for instance, the proportion of the genome covered by islands, the number and the length of islands, with respect to the number of clones and anchors considered. To study the statistical properties of the islands, we must model the clones and anchors processes. This problem is typically part of the theory of coverage processes (Hall, 1988).

Several mathematical analyses have been given of physical mapping by fingerprinting random clones (Lander and Waterman, 1988), by end-characterized random clones (Port *et al.*, 1995) or by anchoring random clones (Arratia *et al.* (1991), Barillot *et al.* (1991), Ewens *et al.* (1991), Torney (1991), Lee (1992), and Marr *et al.* (1992)). Nonrandom anchoring strategies have also been discussed (Palazzolo *et al.* (1991), Zhang and Marr (1993), Nelson and Speed (1994)). In all these works, the process of clone (and anchor) locations is assumed to be homogeneous, except in Port *et al.* (1995) where the results of having different clone distributions along the genome are briefly discussed. It is well-known that the homogeneity assumptions in the genome are not strictly true because of possible cloning bias. This bias could result in less rapid progress and has been pointed out by most of the previous authors. Karlin and Macken (1991) discuss some aspects of the Lander-Waterman case when the clone locations occur as a non-homogeneous Poisson process. But no theoretical studies have yet been done without homogeneity assumptions for the more complex anchoring method involving simultaneously two different processes, the clones and the anchors. Recently, Ewens (1996) provided some simulations describing significant effects of hotspots, both for clones and for anchors, on the island properties. Hotspots here means that the genome contains possible rich and poor regions of clones and/or anchors.

The aim of this paper is to provide a mathematical analysis of physical mapping by anchoring random clones, in a general model. Whereas Arratia *et al.* (1991) use homogeneous Poisson processes, we model the clones and anchors locations as non-homogeneous Poisson processes, allowing a large class of rate functions. The difficulty comes from the non-stationarity and the complicated stochastic rates of the point processes we manipulate. Also, we present two applications considering two different types of rates. In particular, the first model is precisely the hotspot model considered by Ewens (1996). These simple non-homogeneous models can be seen as variations of the homogeneous one, where the rates are constant. Therefore, they allow us to compare the results using an homogeneous model versus a non-homogeneous model. As we will see, the inhomogeneity in the clone and anchor locations along the genome changes the predictions of the physical mapping project.

The paper is organized as follows. Section 2 presents the notation and some preliminary results. The main results are stated in Section 3 following the setup of Arratia *et al.* (1991). These results describe several properties of the islands in term of expected values. For instance, the mean number of islands and the mean proportion of the genome covered by islands are derived, providing important information about the mapping project. Some theorems are valid under specific assumptions. Section 4 briefly discusses the results when the rates both for clones and for anchors are periodic, since in this case those assumptions are satisfied. Then, two examples are presented in Section 5, and Section 6 gives the conclusion of this work. Most of the proofs are gathered in the Appendix.

2 Notation and preliminary results

We naturally consider the genome linearly as a discrete sequence of bases; the genome length, that is the number of bases in the sequence, is denoted by G. We assume that clones have independent and identically distributed lengths L, with mean $\mathbb{E}L$. In order to use a continuous model, we rescale by $\mathbb{E}L$. Defining $g = G/\mathbb{E}L$, the normalized genome then corresponds to the interval $(0, g]$ on the real line $(-\infty, +\infty)$. Let f be the probability density function of the normalized clone length $Q \equiv L/\mathbb{E}L$, and $\mathcal{F}$ the tail function $\mathcal{F}(u) = \mathbb{P}(Q \geq u)$. Since the anchor length is very small compared to the clone length, the anchors will be considered to be points. We assume that right ends of clones occur on the real line according to a non-homogeneous Poisson process of rate $\alpha(t)$, and we label them $\{C_i,\, i \in \mathbb{Z}\}$ such that $\cdots C_{-1} < C_0 \leq 0 < C_1 < \cdots < C_{N(g)} \leq g < C_{N(g)+1} < \cdots$; $N(t) \equiv N((0,t])$ denotes the counting process of clones and represents the number of clones ending in $(0, t]$. Similarly, the anchors occur according to a non-homogeneous Poisson process of rate $\lambda(t)$ and are labeled $\{A_j,\, j \in \mathbb{Z}\}$ such that $\cdots A_{-1} < A_0 \leq 0 < A_1 < \cdots < A_{M(g)} \leq g < A_{M(g)+1} < \cdots$; $M(t) \equiv M((0,t])$ denotes the counting process of anchors and represents the number of anchors in $(0, t]$. We assume the process of anchors and the process of clones are independent, and the rates α and λ are positive functions such that their integrals are finite on bounded sets, but are not finite on unbounded sets.

For all $t \in \mathbb{R}$, let V_t be the distance forward from t to the next anchor, and W_t be the distance back from t to the previous anchor. The probability density functions $g_t(v)$ and $h_t(w)$ respectively of V_t and W_t can be easily obtained since, for $v \geq 0$, $\mathbb{P}(V_t \leq v) = 1 - \mathbb{P}(\text{no anchors in } (t, t+v]) = 1 - \exp\left(-\int_t^{t+v} \lambda(x)dx\right)$, and, for $w \geq 0$, $\mathbb{P}(W_t \leq w) = 1 - \exp\left(-\int_{t-w}^{t} \lambda(x)dx\right)$. By differentiating these quantities, we have

$$g_t(v) = \lambda(t+v)\exp\left(-\int_t^{t+v} \lambda(x)dx\right) \mathbb{1}\{v \geq 0\}, \quad (1)$$

$$h_t(w) = \lambda(t-w)\exp\left(-\int_{t-w}^{t} \lambda(x)dx\right) \mathbb{1}\{w \geq 0\}. \quad (2)$$

Moreover, V_t and W_t are independent since $\mathbb{P}(V_t > v,\, W_t > w) = \exp\left(-\int_{t-w}^{t+v} \lambda(x)dx\right)$.

Proposition 1 *The probability $J(t;x)$ that the points t and $t+x$ ($x>0$) are not covered by a common clone is*

$$J(t;x) = \exp\left(-\int_x^\infty \alpha(t+u)\mathcal{F}(u)du\right).$$

PROOF. A clone of length Q ending at y covers $[t, t+x]$ if and only if $y \geq t+x$ and $y - Q \leq t$. Therefore, the probability that a clone ending at y covers $[t, t+x]$ is

$$\mathbb{E}_Q\left(\mathbb{1}\{y \geq t+x\}\,\mathbb{1}\{y-Q \leq t\}\right) = \mathbb{1}\{y \geq t+x\}\mathcal{F}(y-t).$$

Since the rate of clones is $\alpha(y)$, the process of clones covering $[t, t+x]$ is a Poisson process with rate $\alpha(y)\mathcal{F}(y-t)\mathbb{1}\{y \geq t+x\}$ (Marking Theorem, Kingman (1993)). Therefore, the number of clones covering $[t, t+x]$ is a Poisson variable with mean $\int_{-\infty}^{+\infty} \alpha(y)\mathcal{F}(y-t)\mathbb{1}\{y \geq t+x\}dy$, and $J(t;x)$ is the probability that this Poisson variable is equal to zero:

$$J(t;x) = \exp\left(-\int_{t+x}^{+\infty} \alpha(y)\mathcal{F}(y-t)dy\right).$$

The change of variable $u = y - t$ leads to the result. ∎

Proposition 2 *If $t_1 \leq t_2 \leq t_3 \leq t_4$ then*

(i) The probability that no clones cover $[t_1, t_2]$ and no clones cover $[t_3, t_4]$ is

$$J(t_1; t_2 - t_1)J(t_3; t_4 - t_3)/J(t_1; t_4 - t_1).$$

(ii) The probability that at least one clone covers $[t_1, t_2]$, no clones cover $[t_1, t_3]$ and at least one clone covers $[t_2, t_3]$ is

$$\left(1 - \frac{J(t_1; t_2 - t_1)}{J(t_1; t_3 - t_1)}\right)\left(J(t_1; t_3 - t_1) - J(t_2; t_3 - t_2)\right).$$

PROOF. (i) Let E be the event that no clones cover $[t_1, t_2]$ and no clones cover $[t_3, t_4]$. We can write E as $E_1 \cap E_2$ where E_1 is the event that no clones cover $[t_3, t_4]$ and E_2 is the event that no clones ending in $[t_2, t_4)$ cover t_1. Since the clones occur according to a Poisson process, the events E_1 and E_2 are independent and $\mathbb{P}(E) = \mathbb{P}(E_1)\mathbb{P}(E_2) = J(t_3; t_4 - t_3)\mathbb{P}(E_2)$ from Proposition 1. We now calculate $\mathbb{P}(E_2)$: a clone of length Q, ending at y, ends in $[t_2, t_4)$ and covers t_1 if and only if $t_2 \leq y < t_4$ and $y - Q \leq t_1$. Hence, the probability that a clone, ending at y, ends in $[t_2, t_4)$ and covers t_1 is

$$\mathbb{E}_Q\left(\mathbb{1}\{t_2 \leq y < t_4\}\mathbb{1}\{y - Q \leq t_1\}\right) = \mathcal{F}(y - t_1)\mathbb{1}\{t_2 \leq y < t_4\}.$$

Then, the number of clones ending in $[t_2, t_4)$ and covering t_1 is a Poisson variable with mean

$$\int_{-\infty}^{\infty} \alpha(y)\mathcal{F}(y-t_1)\mathbb{I}\{t_2 \le y < t_4\}dy = \int_{t_2}^{t_4} \alpha(y)\mathcal{F}(y-t_1)dy = \int_{t_2-t_1}^{t_4-t_1} \alpha(u+t_1)\mathcal{F}(u)du.$$

Therefore,

$$\begin{aligned}\mathbb{P}(E_2) &= \exp\left(-\int_{t_2-t_1}^{t_4-t_1} \alpha(u+t_1)\mathcal{F}(u)du\right) \\ &= \exp\left(-\int_{t_2-t_1}^{\infty} \alpha(u+t_1)\mathcal{F}(u)du + \int_{t_4-t_1}^{\infty} \alpha(u+t_1)\mathcal{F}(u)du\right) \\ &= \frac{J(t_1;t_2-t_1)}{J(t_1;t_4-t_1)},\end{aligned}$$

which establishes (i).

(ii) Let E' be the event that at least one clone covers $[t_1, t_2]$, no clones cover $[t_1, t_3]$ and at least one clone covers $[t_2, t_3]$. We can write E' as $E_3 \cap E_4$ where the events E_3 and E_4 are as follows: $E_3 = \{\text{at least one clone ending in } [t_2, t_3) \text{ covers } t_1\}$ and $E_4 = \{\text{at least one clone covers } [t_2, t_3] \text{ but no clones cover } [t_1, t_3]\}$. Since E_3 concerns the clones ending in $[t_2, t_3)$ and E_4 concerns the clones ending after t_3, the events E_3 and E_4 are independent and their probabilities are

$$\begin{aligned}\mathbb{P}(E_3) &= 1 - \mathbb{P}(\text{no clones ending in } [t_2,t_3) \text{ cover } t_1) = 1 - \frac{J(t_1;t_2-t_1)}{J(t_1;t_3-t_1)}. \\ \mathbb{P}(E_4) &= \mathbb{P}(\text{no clones cover } [t_1,t_3]) - \mathbb{P}(\text{no clones cover } [t_2,t_3]) \\ &= J(t_1;t_3-t_1) - J(t_2;t_3-t_2).\end{aligned}$$

Finally, $\mathbb{P}(E') = \mathbb{P}(E_3)\mathbb{P}(E_4)$ which leads to the result. ∎

3 Main results

We introduce the following terminology: a clone containing an anchor is an anchored clone, and the clones containing one or more common anchors are assembled into anchored islands as shown by Figure 1. An anchored island can be composed of a unique anchored clone. A region between two anchored islands is an ocean. The main aim of this section is to

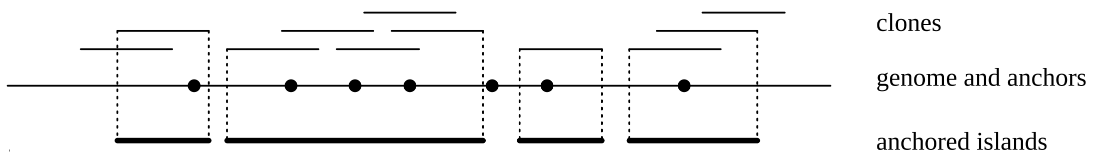

Figure 1: Illustration of the anchored islands

provide results about anchored islands.

The next theorem gives properties of unanchored clones and is proved in Appendix A.1.

Theorem 1 *Let $N_u(t) = N_u((0,t])$ be the counting process associated with the right end of unanchored clones.*

(i) The probability $q_1(t)$ that a clone ending at t contains no anchor is

$$q_1(t) = \int_0^\infty \exp\left(-\int_{t-\ell}^t \lambda(x)dx\right) f(\ell)d\ell\,.$$

(ii) The stochastic rate of the unanchored clones (right ends) process is

$$\phi_u(t) = \alpha(t)(1 - \mathcal{F}(W_t))\,.$$

(iii) The mean number of unanchored clones ending in $(0,g]$ is

$$\mathbb{E}N_u(g) = \int_0^g \alpha(t)q_1(t)dt\,.$$

Remark 1 *It follows easily from Theorem 1 that the rate $\phi(t)$ of the anchored clones (right ends) process is given by*

$$\phi(t) = \alpha(t)\mathcal{F}(W_t) = \alpha(t) - \phi_u(t)\,.$$

If $N_a(g)$ denotes the number of anchored clones ending in $(0,g]$, we have $N_a(g)+N_u(g) = N(g)$ and $\mathbb{E}N_a(g) = \mathbb{E}\int_0^g \phi(t)\,dt = \int_0^g \alpha(t)(1-q_1(t))\,dt$.

The next theorem gives properties of anchored islands process and is proved in Appendix A.2.

Theorem 2 *Let $I_a(t) = I_a((0,t])$ be the counting process associated with the right end of anchored islands.*

(i) The probability $p_1(t)$ that a clone ending at t is the rightmost clone of an anchored island is

$$p_1(t) = \int_0^\infty J(t-w;w)\mathcal{F}(w)\lambda(t-w)\exp\left(-\int_{t-w}^t \lambda(x)dx\right)dw\,.$$

(ii) The stochastic rate of the anchored islands (right ends) process is

$$\psi(t) = \alpha(t)J(t-W_t;W_t)\mathcal{F}(W_t)\,.$$

(iii) The mean number of anchored islands ending in $(0,g]$ is

$$\mathbb{E}I_a(g) = \int_0^g \alpha(t)p_1(t)dt\,.$$

The next theorem gives properties of particular anchored islands: the singleton anchored islands, composed of a unique anchored clone. It is proved in Appendix A.3.

Theorem 3 *Let $I_s(t) = I_s((0,t])$ be the counting process associated with the right end of the singleton anchored islands.*

(i) The probability $p_2(t)$ that a clone ending at t is a singleton anchored island is

$$\begin{aligned} p_2(t) &= \int_{\ell=0}^{\infty}\int_{w=0}^{\ell}\int_{v=0}^{\ell-w} \frac{J(t-w;w)J(t-\ell;v)}{J(t-\ell;\ell)}\,\lambda(t-\ell+v)\lambda(t-w) \\ &\qquad\times\exp\left(-\int_{t-w}^{t}\lambda(x)dx-\int_{t-\ell}^{t-\ell+v}\lambda(x)dx\right)f(\ell)dv\,dw\,d\ell \\ &\quad+\int_{\ell=0}^{\infty}\int_{w=0}^{\ell}\frac{J(t-w;w)J(t-\ell;\ell-w)}{J(t-\ell;\ell)}\,\lambda(t-w)\exp\left(-\int_{t-\ell}^{t}\lambda(x)dx\right)f(\ell)dw\,d\ell\,. \end{aligned}$$

(ii) The stochastic rate of the singleton anchored islands (right ends) process is

$$\gamma(t) = \alpha(t)J(t-W_t;W_t)\mathbb{E}_Q\left(\frac{J(t-Q;V_{t-Q})}{J(t-Q;Q)}\mathbb{1}\{Q \geq W_t + V_{t-Q}\}\right).$$

(iii) The mean number of singleton anchored islands ending in $(0,g]$ is

$$\mathbb{E}I_s(g) = \int_0^g \alpha(t)p_2(t)dt\,.$$

The next theorem is about the proportion of oceans, in other words the proportion of the genome not covered by anchored islands. It is proved in Appendix A.4.

Theorem 4 *The probability $r_0(t)$ that t is not covered by any anchored island is*

$$r_0(t) = \int_0^\infty\int_0^\infty \frac{J(t;v)J(t-w;w)}{J(t-w;v+w)}\,\lambda(t+v)\lambda(t-w)\exp\left(-\int_{t-w}^{t+v}\lambda(x)dx\right)dv\,dw\,.$$

The mean proportion r_0 of the genome not covered by anchored islands is

$$r_0 = \frac{1}{g}\int_0^g r_0(t)dt\,.$$

For Theorems 5-7 that follow, we make the following assumptions on the rates α and λ, via the rate ψ of the anchored islands process and the rate ϕ of the anchored clones process given by Theorem 2 and Remark 1. Let us recall that $\mathbb{E}\int_0^g \psi(t)\,dt = \mathbb{E}I_a(g) = \int_0^g \alpha(t)p_1(t)\,dt$ and $\mathbb{E}\int_0^g \phi(t)\,dt = \mathbb{E}N_a(g) = \int_0^g \alpha(t)(1-q_1(t))\,dt$.

$$(A_1) \qquad \lim_{g\to\infty} \mathbb{E}\int_0^g \psi(t)\,dt = \infty \quad \text{and} \quad \frac{\int_0^g \psi(t)\,dt}{\mathbb{E}\int_0^g \psi(t)\,dt} \xrightarrow{\mathbb{P}} 1\,, \text{ as } g\to\infty\,,$$

$$(A_1') \qquad \frac{g}{\mathbb{E}\int_0^g \psi(t)\,dt} \text{ is bounded as } g\to\infty\,,$$

$$(A_2) \qquad \lim_{g\to\infty} \mathbb{E}\int_0^g \phi(t)\,dt = \infty \quad \text{and} \quad \frac{\int_0^g \phi(t)\,dt}{\mathbb{E}\int_0^g \phi(t)\,dt} \xrightarrow{\mathbb{P}} 1\,, \text{ as } g\to\infty\,.$$

We will see in Section 4 that these assumptions are satisfied for periodic rates α and λ.

The next proposition states the weak law of large numbers for the number of anchored islands and the number of anchored clones. It is proved in Appendix A.4.

Proposition 3 *(i) Under the assumption (A_1), the ratio $I_a(g)/\int_0^g \alpha(t)p_1(t)\,dt$ converges to 1 in probability, as g tends to infinity.*

(ii) Under the assumption (A_2), the ratio $N_a(g)/\int_0^g \alpha(t)(1-q_1(t))\,dt$ converges to 1 in probability, as g tends to infinity.

The next theorem concerns the length of the anchored islands and is proved in Appendix A.5.

Theorem 5 *Under the assumptions (A_1) and (A_1'), the average length of the anchored islands ending in $(0,g]$, denoted by $\overline{S}_g$, is such that*

$$\overline{S}_g - \frac{\int_0^g \Omega(t)\,dt}{\int_0^g \alpha(t)p_1(t)dt} \xrightarrow{\mathbb{P}} 0$$

as $g\to\infty$, where

$$\begin{aligned}\Omega(t) \;=\; 1 + \int_0^\infty \int_0^\infty & [J(t-w;v+w) - J(t;v) - J(t-w;w)] \\ & \times\lambda(t+v)\lambda(t-w)\exp\left(-\int_{t-w}^{t+v}\lambda(x)dx\right) dv dw\,.\end{aligned}$$

$\overline{S}_g\mathbb{E}L$ is then the average length of anchored islands in the genome.

The two following theorems provide the mean number of clones and the mean number of anchors in an anchored island. They are proved in Appendix A.5.

Theorem 6 *Under the assumptions (A_1) and (A_2), the average number of clones per anchored island ending in $(0,g]$, denoted by $\overline{Y}_g$, is such that*

$$\overline{Y}_g - \frac{\int_0^g \alpha(t)(1-q_1(t))dt}{\int_0^g \alpha(t)p_1(t)dt} \xrightarrow{\mathbb{P}} 0 \quad \textit{as } g\to\infty\,.$$

Theorem 7

(i) The process of anchors covered by clones is a Poisson process with rate given by

$$\nu(t) = \lambda(t)(1 - J(t;0))\,.$$

(ii) Under the assumption (A_1), the average number of anchors per anchored island ending in $(0, g]$, denoted by $\overline{H}_g$, is such that

$$\overline{H}_g - \frac{\int_0^g \lambda(t)(1 - J(t;0))\,dt}{\int_0^g \alpha(t) p_1(t)\,dt} \xrightarrow{\mathbb{P}} 0 \qquad as\ g \to \infty\,.$$

Whereas an ocean is a region not covered by anchored islands, an actual ocean is defined to be a region not covered by clones.

Theorem 8 *The probability $p_3(t;x)$ that an anchored island ending at t is followed by an actual ocean of length at least x is*

$$p_3(t;x) = J(t+x;0) \exp\left(-\int_t^{t+x} \alpha(u)du\right) \frac{1 - q_1(t)}{p_1(t)}\,.$$

Proof. It follows from Bayes' formula that $p_3(t;x) = p(t;x)/p_1(t)$ where $p(t;x)$ is the probability that a clone ending at t is the right end of an anchored island followed by an actual ocean of length at least x. We now calculate $p(t;x)$. Suppose that the given clone ending at t has length Q. Let W_t be the distance back from t to the previous anchor. Given Q and W_t, the clone of length Q ending at t is the right end of an anchored island followed by an actual ocean of length at least x if and only if $W_t \le Q$, no clones end in $[t, t+x]$ and no clones cover $t+x$. Therefore,

$$\begin{aligned}
p(t;x) &= \mathbb{E}_{Q,W_t}\left(\mathbb{1}\{W_t \le Q\} J(t+x;0) \exp\left(-\int_t^{t+x} \alpha(u)du\right)\right) \\
&= J(t+x;0) \exp\left(-\int_t^{t+x} \alpha(u)du\right) \int_{\ell=0}^{\infty} \int_{w=0}^{\ell} \lambda(t-w) \exp\left(-\int_{t-w}^{t} \lambda(u)du\right) dw\, f(\ell) d\ell \\
&= J(t+x;0) \exp\left(-\int_t^{t+x} \alpha(u)du\right) \int_{\ell=0}^{\infty} 1 - \exp\left(-\int_{t-\ell}^{t} \lambda(u)du\right) f(\ell) d\ell\,,
\end{aligned}$$

and using Theorem 1 completes the proof. ■

4 Periodic case

In this section, we focus on the case where the rate of clones and the rate of anchors are periodic and have a common period. This particular framework is not only theoretically interesting because Theorems 5, 6 and 7 hold; indeed, the next proposition ensures the assumptions are satisfied. It is also interesting from a practical point of view since it can lead to very simple and natural non-homogeneous models, as we will see in Section 5.

We say that α and λ have a period in common if there exists $\theta > 0$ such that $\alpha(t+\theta) = \alpha(t)$ and $\lambda(t+\theta) = \lambda(t)$ for all $t \in \mathbb{R}$.

Proposition 4 *If the rates α and λ of the clones and anchors processes are periodic and have a period in common, then the assumptions (A_1), (A_1') and (A_2) are satisfied.*

Proof. We intend to prove only that (A_1) and (A_1') are satisfied, since the proof for (A_2) is similar to that for (A_1).
Let $\theta > 0$ such that $\alpha(t+\theta) = \alpha(t)$ and $\lambda(t+\theta) = \lambda(t)$ for all $t \in \mathbb{R}$. It follows from (2) that the two random variables W_t and $W_{t+\theta}$ have the same distribution. Since $\psi(t) = \alpha(t)J(t - W_t; W_t)\mathbb{P}_Q(Q \geq W_t)$ (Theorem 2(ii)) and $J(t;x) = J(t+\theta;x)$, $\psi(t)$ and $\psi(t+\theta)$ are also identically distributed, $\forall t \in \mathbb{R}$. Let us define the nonnegative random variables

$$X_i \equiv \int_{(i-1)\theta}^{i\theta} \psi(t)\, dt\,; \tag{3}$$

the X_i's have the same distribution but are not independent; their expectation is denoted by m, so $m = \mathbb{E}X_i = \mathbb{E}\int_0^\theta \psi(t)\, dt > 0$. Moreover, they satisfy the weak law of large numbers (proved below): $n^{-1}\sum_{i=1}^n X_i \to m$ in probability as $n \to \infty$. Since, for $n \in \mathbb{N}$, $\int_0^{n\theta} \psi(t)\, dt = \sum_{i=1}^n X_i$ and $\mathbb{E}\int_0^{n\theta} \psi(t)\, dt = nm$, we see that $\mathbb{E}\int_0^{n\theta} \psi(t)\, dt \to \infty$ as $n \to \infty$ and

$$\frac{\int_0^{n\theta} \psi(t)\, dt}{\mathbb{E}\int_0^{n\theta} \psi(t)\, dt} \xrightarrow{\mathbb{P}} 1\,.$$

Now, if n is the integer part of g/θ, the assumption (A_1) holds from $\mathbb{E}\int_0^g \psi(t)\, dt \geq \mathbb{E}\int_0^{n\theta} \psi(t)\, dt$ and

$$\frac{n}{n+1}\,\frac{\int_0^{n\theta} \psi(t)\, dt}{\mathbb{E}\int_0^{n\theta} \psi(t)\, dt} \leq \frac{\int_0^{g} \psi(t)\, dt}{\mathbb{E}\int_0^{g} \psi(t)\, dt} \leq \frac{n+1}{n}\,\frac{\int_0^{(n+1)\theta} \psi(t)\, dt}{\mathbb{E}\int_0^{(n+1)\theta} \psi(t)\, dt}\,,$$

since the lower and upper bounds converge both to 1 as $n \to \infty$.
The assumption (A_1') holds in the same way from

$$\frac{n\theta}{(n+1)\mathbb{E}\int_0^{\theta} \psi(t)\, dt} \leq \frac{g}{\mathbb{E}\int_0^{g} \psi(t)\, dt} \leq \frac{(n+1)\theta}{n\mathbb{E}\int_0^{\theta} \psi(t)\, dt}\,.$$

To complete the proof, let us show that the X_i's satisfy the weak law of large numbers. From the Markov's theorem (Gnedenko, 1968, p.248), it holds when $n^{-2}\mathrm{Var}\,(\sum_{i=1}^n X_i) \to$

0 as $n \to \infty$. We introduce the notation $G(t, W_t) \equiv J(t - W_t; W_t)\mathcal{F}(W_t)$, so $\psi(t) = \alpha(t)G(t, W_t)$ and

$$\mathbb{E}X_i^2 = \mathbb{E}X_1^2 = \int_{t=0}^{\theta} \int_{s=0}^{\theta} \alpha(t)\alpha(s)\mathbb{E}_{(W_t,W_s)}(G(t, W_t)G(s, W_s))\, ds\, dt\,.$$

The joint distribution of (W_t, W_s), for $t > s$, can be easily obtained since

$$\begin{aligned}
\mathbb{P}(W_s > x, W_t > y) &= \mathbb{P}(\text{no anchors in } [s-x; s) \text{ and in } [t-y; t)) \\
&= \begin{cases} \exp\left(-\int_{s-x}^{s} \lambda(u)\, du - \int_{t-y}^{t} \lambda(u)\, du\right) & \text{if } y < t - s\,, \\ \exp\left(-\int_{s-x}^{s} \lambda(u)\, du\right) & \text{if } t - s \leq y < t - s + x\,, \\ \exp\left(-\int_{t-y}^{t} \lambda(u)\, du\right) & \text{if } y \geq t - s + x\,. \end{cases}
\end{aligned}$$

We obtain $d\mathbb{P}_{(W_s,W_t)}(x, y) = d\mathbb{P}_{W_s}(x)d\mathbb{P}_{W_t}(y)\mathbb{1}\{y < t - s\}$ if $t > s$. Therefore,

$$\begin{aligned}
\mathbb{E}X_1^2 &= \int_{t=0}^{\theta} \alpha^2(t)\, \mathbb{E}\left(G^2(t, W_t)\right)\, dt \\
&\quad + 2\int_{t=0}^{\theta} \int_{s=0}^{t} \alpha(t)\alpha(s)\, \mathbb{E}(G(s, W_s))\, \mathbb{E}(G(t, W_t)\mathbb{1}\{W_t < t - s\})\, ds\, dt \\
&\leq \int_{t=0}^{\theta} \alpha^2(t)\, \mathbb{E}\left(G^2(t, W_t)\right)\, dt + 2\int_{t=0}^{\theta} \int_{s=0}^{t} \alpha(t)p_1(t)\alpha(s)p_1(s)\, ds\, dt \\
&= \int_{t=0}^{\theta} \alpha^2(t)\, \mathbb{E}\left(G^2(t, W_t)\right)\, dt + m^2\,;
\end{aligned}$$

we used $\mathbb{E}(G(s, W_s)) = p_1(s)$ and $m = \int_0^{\theta} \alpha(t)p_1(t)\, dt$ from Theorem 2. $\mathrm{Var}X_1$ is then bounded and $n^{-2}\sum_{i=1}^{n} \mathrm{Var}X_i = n^{-1}\mathrm{Var}X_1$ converges to 0 as $n \to \infty$.
We now calculate $\mathrm{Cov}(X_i, X_j)$ for $j < i$. By same arguments, we have

$$\mathbb{E}X_iX_j = \int_{t=(i-1)\theta}^{i\theta} \int_{s=(j-1)\theta}^{j\theta} \alpha(s)p_1(s)\alpha(t)\, \mathbb{E}(G(t, W_t)\mathbb{1}\{W_t < t - s\})\, ds\, dt\,,$$

and

$$\mathrm{Cov}(X_i, X_j) = -\int_{t=(i-1)\theta}^{i\theta} \int_{s=(j-1)\theta}^{j\theta} \alpha(s)p_1(s)\alpha(t)\, \mathbb{E}(G(t, W_t)\mathbb{1}\{W_t \geq t - s\})\, ds\, dt \leq 0\,.$$

Therefore, $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \leq \sum_{i=1}^{n} \mathrm{Var}X_i$ leading to $n^{-2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \to 0$ as $n \to \infty$. This completes the proof of the proposition. ■

Remark 2 *Actually, the X_i's defined by (3) obviously satisfy the following conditions: $X_i \geq 0$, $\sup_{i\in\mathbb{N}} \mathbb{E}X_i < \infty$, $\mathbb{E}X_i^2 < \infty$, $Cov(X_i, X_j) \leq 0$ for $j > i$ and $\sum_{i=1}^{\infty} i^{-2}\, VarX_i < \infty$. This ensures from Etemadi (1983) that the X_i's satisfy the strong law of large numbers. Therefore, if the rates α and λ of the clones and anchors processes are periodic and have a common period, then the assumptions (A_1) and (A_2) even hold with almost sure convergence instead of convergence in probability.*

We now give corollaries of Theorems 5, 6 and 7.

Corollary 1 *If the rates α and λ are periodic with period θ in common, the average length of the anchored islands ending in $(0, g]$, denoted by $\overline{S}_g$, is such that*

$$\overline{S}_g \xrightarrow{\mathbb{P}} \frac{\int_0^\theta \Omega(t)\, dt}{\int_0^\theta \alpha(t) p_1(t)\, dt}$$

as $g \to \infty$, where $\Omega(t)$ is given by Theorem 5.

Proof. If the rates α and λ are periodic and have the period θ in common, the assumptions (A_1) and (A_1') are satisfied from Proposition 4. We can then apply Theorem 5. Since p_1 and Ω are also θ-periodic and positive, we can write

$$\frac{n}{n+1} \frac{\int_0^\theta \Omega(t) dt}{\int_0^\theta \alpha(t) p_1(t) dt} \leq \frac{\int_0^g \Omega(t) dt}{\int_0^g \alpha(t) p_1(t) dt} \leq \frac{n+1}{n} \frac{\int_0^\theta \Omega(t) dt}{\int_0^\theta \alpha(t) p_1(t) dt},$$

where n is the integer part of g/θ. Therefore, the ratio $\int_0^g \Omega(t)\, dt / \int_0^g \alpha(t) p_1(t)\, dt$ converges to $\int_0^\theta \Omega(t)\, dt / \int_0^\theta \alpha(t) p_1(t)\, dt$ as g tends to infinity, which establishes the corollary. ■

Corollary 2 *If the rates α and λ are periodic with period θ in common, then the average number of clones per anchored island ending in $(0, g]$, denoted by $\overline{Y}_g$, is such that*

$$\overline{Y}_g \xrightarrow{\mathbb{P}} \frac{\int_0^\theta \alpha(t)(1 - q_1(t)) dt}{\int_0^\theta \alpha(t) p_1(t) dt} \quad \text{as } g \to \infty.$$

Proof. If the rates α and λ are periodic and have the period θ in common, the assumptions (A_1) and (A_2) are satisfied from Proposition 4. We can then apply Theorem 6. Since p_1 and q_1 are also θ-periodic and positive, the ratio $\int_0^g \alpha(t)(1 - q_1(t))\, dt / \int_0^g \alpha(t) p_1(t)\, dt$ converges to $\int_0^\theta \alpha(t)(1 - q_1(t))\, dt / \int_0^\theta \alpha(t) p_1(t)\, dt$ as g tends to infinity, which leads to the result. ■

Corollary 3 *If the rates α and λ are periodic with period θ in common, then the average number of anchors per anchored island ending in $(0, g]$, denoted by $\overline{H}_g$, is such that*

$$\overline{H}_g \xrightarrow{\mathbb{P}} \frac{\int_0^\theta \lambda(t)(1 - J(t; 0))\, dt}{\int_0^\theta \alpha(t) p_1(t)\, dt} \quad \text{as } g \to \infty.$$

Proof. If the rates α and λ are periodic and have the period θ in common, then the assumption (A_1) is satisfied, and $p_1(\cdot)$ and $J(\cdot; 0)$ are θ-periodic. The corollary then follows using Theorem 7(ii). ■

5 Applications

In this section, we present numerical results for a genome of length $G = 100,000$ kb, 2300 clones of fixed length $L = 250$ kb and 500 anchors, that corresponds approximately to the physical mapping project of *Arabidopsis thaliana* genome (Ewens *et al.*, 1991). In our notation, the normalized genome length is $g = 400$, and the mean number of clones and anchors in $(0, g]$ are respectively $\mathbb{E}N \equiv \mathbb{E}N(g) = 2300$ and $\mathbb{E}M \equiv \mathbb{E}M(g) = 500$.
We focus on four quantities, namely the mean number of unanchored clones (Theorem 1), the mean number of anchored islands (Theorem 2), the average length of anchored islands (Theorem 5) and the mean proportion of the genome not covered by anchored islands (Theorem 4). We consider the two following models that represent two simple ways to describe non-homogeneity in the genome. Whereas the stationary model means the rates α and λ are constant, our first non-homogeneous model assumes each rate is piecewise constant and can take two values. It means the genome is composed of alternating rich and poor regions of clones, and of anchors. This is the hotspot model introduced by Ewens (1996). The second model allows some sinusoidal fluctuations around the constant rates.

5.1 Hotspot model

Let $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$; we consider the clone rate $\alpha(t)$ such that it can only take values α_1 and α_2. This rate is piecewise constant, and for simplicity here we define $\alpha(t)$ as the θ_α-periodic function on $\mathbb{R}$:

$$\alpha(t) = \begin{cases} \alpha_1 & \text{if } t \bmod \theta_\alpha \in (0, \theta_\alpha/2] \\ \alpha_2 & \text{if } t \bmod \theta_\alpha \in (\theta_\alpha/2, \theta_\alpha] \end{cases} \tag{4}$$

with $\theta_\alpha \geq 0$. The parameters α_1, α_2 and θ_α are such that $\int_0^g \alpha(t)dt = \mathbb{E}N$. Similarly, let us consider the anchor rate $\lambda(t)$ as the θ_λ-periodic function on $\mathbb{R}$:

$$\lambda(t) = \begin{cases} \lambda_1 & \text{if } t \bmod \theta_\lambda \in (0, \theta_\lambda/2] \\ \lambda_2 & \text{if } t \bmod \theta_\lambda \in (\theta_\lambda/2, \theta_\lambda] \end{cases} \tag{5}$$

where λ_1, λ_2 and θ_λ are positive and such that $\int_0^g \lambda(t)dt = \mathbb{E}M$.
For the numerical application, we take $\theta_\alpha = \theta_\lambda = 500/L = 2$. This implies that $\int_0^g \alpha(t)dt = g(\alpha_1 + \alpha_2)/2$ and $\int_0^g \lambda(t)dt = g(\lambda_1 + \lambda_2)/2$. In other words, we have $\alpha_2 = 11.5 - \alpha_1$ and $\lambda_2 = 2.5 - \lambda_1$. Therefore, the case $\alpha_1 = \alpha_2 = 5.75$ corresponds to the stationary model for the clone locations and $\lambda_1 = \lambda_2 = 1.25$ leads to the stationary model for the anchors. If $\alpha_1 > \alpha_2$, the model suggests the regions where $\alpha(t) = \alpha_1$ are rich regions of (right ends of) clones, whereas the other ones are poor; and conversely if $\alpha_1 < \alpha_2$. The same interpretation holds for the anchors. Therefore, rates defined by (4) and (5) characterize the hotspot phenomena both for clones and for anchors. In our present application, the clone and anchor hotspots are either in phase, if $0 \leq \alpha_1 \leq 5.75$ and $0 \leq \lambda_1 \leq 1.25$ or if

$5.75 < \alpha_1 \leq 11.5$ and $1.25 < \lambda_1 \leq 2.5$, or completely out of phase otherwise. A more general study of the hotspot effect will be presented in another paper. Actually, we obtain the most extreme results (with $\theta_\alpha = \theta_\lambda$) when the lag between the clone hotspot and the anchor hotspot is $-1/2$ or $3/4$. It is rather natural since α is in fact the rate of the clone right ends; the rate of the clone centers would be the shifted α with a $-1/2$ shift.

Tables 1-4 give respectively the mean number of unanchored clones, the mean number Table 1
of anchored islands, the average length of anchored islands and the mean proportion of Table 2
the genome not covered by anchored islands, when α_1 varies from 0 to 11.5 and λ_1 varies Table 3
from 0 to 2.5. Since $\alpha_1 = 5.75$ is the stationary case, increasing $|\alpha_1 - 5.75|$ shows the Table 4
effect of increasing the level of the clone hotspot. In the same way, increasing $|\lambda_1 - 1.25|$ shows the effect of increasing the level of the anchor hotspot. As a consequence of the model, the results for a pair (α_1, λ_1) have to be similar to those obtained for the pair $(11.5 - \alpha_1, 2.5 - \lambda_1)$, for all α_1 and λ_1; the lower rows can then be deduced from the upper rows. However, one can notice some slight differences due to numerical computations. Values boldfaced in the tables correspond to the minimum and the maximum values.

These are the main trends. When the hotspots are in phase, the number of unanchored clones decreases as the level of clone hotspot increases. As both of the hotspot levels increase, the mean number of anchored islands and the mean length of an anchored island decrease, leading to an increase of the mean proportion of the genome not covered by anchored islands.

When the hotspots are completely out of phase, the number of unanchored clones increases as both of the hotspot levels increase. As the level of clone hotspot increases, the mean number of anchored islands increases and then drops, whereas it decreases as the level of anchor hotspot increases. The mean length of anchored islands decreases as the level of clone hotspot increases; it decreases also when the level of anchor hotspot increases, except for small $|\alpha_1 - 5.75|$ when it remains first constant, then increases. The mean proportion of the genome not covered by anchored islands increases as the levels of hotspot increase, rather dramatically when the level of clone hotspot increases.

The most important effect is undeniably that the more the model departs from the stationary model, the greater the mean proportion of genome not covered by anchored islands. In particular, the stationary case ($\alpha_1 = 5.75$ and $\lambda_1 = 1.25$) minimizes the mean proportion of the genome not covered by anchored islands, predicts long anchored islands and a rather large number of anchored islands.

The simulation results given by Ewens (1996) (Tables 7.1, 7.2 and 7.7) and these theoretical calculations are quite close, with a difference between them explained by end effects: whereas we consider anchored islands whose right end falls in $(0, g]$, Ewens considered anchored islands with both ends falling in $(0, g]$.

Table 1: Mean number of unanchored clones

	α_1									
λ_1	0	1.4375	2.875	4.3125	5.75	7.1875	8.625	10.0625	11.5	
0	516.6	648.6	780.5	912.5	1044.4	1176.4	1308.3	1440.3	**1572.2**	$\nearrow$
0.3125	**508.9**	598.0	687.0	776.1	865.1	954.1	1043.2	1132.2	1221.3	$\nearrow$
0.625	527.7	582.6	637.5	692.4	747.3	802.2	857.1	912.0	966.9	$\nearrow$
0.9375	575.9	602.1	628.2	654.4	680.6	706.7	732.9	759.0	785.2	$\nearrow$
1.25	659.0	659.0	659.0	659.0	659.0	659.0	659.0	659.0	659.0	$\rightarrow$
1.5625	785.2	759.0	732.9	706.7	680.6	654.4	628.2	602.1	575.9	$\searrow$
1.875	966.9	912.0	857.1	802.2	747.3	692.4	637.5	582.6	527.7	$\searrow$
2.1875	1221.3	1132.2	1043.2	954.1	865.1	776.1	687.0	598.0	**508.9**	$\searrow$
2.5	**1572.2**	1440.3	1308.3	1176.4	1044.4	912.5	780.5	648.6	516.6	$\searrow$
	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	

Table 2: Mean number of anchored islands

	α_1									
λ_1	0	1.4375	2.875	4.3125	5.75	7.1875	8.625	10.0625	11.5	
0	115.7	122.0	124.2	126.3	130.0	136.0	144.3	147.6	**89.5**	$\nearrow\downarrow$
0.3125	126.7	147.8	151.5	152.9	155.8	161.5	169.7	172.6	119.8	$\nearrow\downarrow$
0.625	135.2	166.1	169.4	169.3	170.9	175.3	182.8	185.3	136.2	$\nearrow\downarrow$
0.9375	141.2	178.7	180.7	178.6	178.6	181.8	187.7	**189.1**	143.3	$\uparrow\rightarrow\downarrow$
1.25	144.3	186.4	186.7	182.6	181.0	182.6	186.7	186.4	144.3	$\uparrow\rightarrow\downarrow$
1.5625	143.3	**189.0**	187.7	181.8	178.6	178.6	180.7	178.7	141.2	$\uparrow\rightarrow\downarrow$
1.875	136.2	185.4	182.8	175.5	170.8	169.3	169.4	166.1	135.2	$\uparrow\searrow$
2.1875	119.8	172.9	169.7	161.3	155.9	153.1	151.5	147.8	126.7	$\uparrow\searrow$
2.5	**89.4**	147.4	144.5	136.0	129.9	126.4	124.1	121.9	115.8	$\uparrow\searrow$
	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	$\nearrow\searrow$	

Table 3: Average length of anchored islands

λ_1	α_1 0	1.4375	2.875	4.3125	5.75	7.1875	8.625	10.0625	11.5	
0	540.7	573.0	589.0	**589.7**	574.4	542.0	484.8	413.6	**383.9**	$\nearrow\searrow$
0.3125	513.3	527.6	550.9	560.8	555.1	530.7	486.9	426.8	403.2	$\nearrow\searrow$
0.625	491.0	501.5	532.0	549.1	549.2	530.9	493.4	440.1	419.5	$\nearrow\searrow$
0.9375	471.6	483.1	520.0	543.1	548.1	534.5	501.4	453.7	436.3	$\nearrow\searrow$
1.25	453.6	467.7	510.4	538.6	547.9	538.6	510.5	467.7	453.6	$\nearrow\searrow$
1.5625	436.3	453.6	501.6	534.5	548.1	543.0	520.0	483.1	471.5	$\nearrow\searrow$
1.875	419.7	440.0	493.3	531.4	549.4	548.7	532.3	501.5	491.3	$\nearrow\searrow$
2.1875	402.5	426.7	486.8	531.4	555.0	560.1	550.9	527.7	513.7	$\nearrow\searrow$
2.5	**384.3**	414.1	486.3	540.4	575.3	**589.9**	589.0	573.2	541.1	$\nearrow\searrow$
	$\searrow$	$\searrow$	$\searrow\rightarrow$	$\searrow\rightarrow\nearrow$	$\searrow\rightarrow\nearrow$	$\searrow\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	

Table 4: Mean proportion of the genome not covered by anchored islands

λ_1	α_1 0	1.4375	2.875	4.3125	5.75	7.1875	8.625	10.0625	11.5	
0	0.402	0.331	0.300	0.288	0.286	0.296	0.326	0.408	**0.656**	$\searrow\nearrow$
0.3125	0.392	0.277	0.233	0.215	0.217	0.217	0.242	0.315	0.543	$\searrow\nearrow$
0.625	0.389	0.242	0.190	0.171	0.166	0.171	0.193	0.260	0.473	$\searrow\nearrow$
0.9375	0.392	0.224	0.167	0.147	0.142	0.147	0.168	0.231	0.430	$\searrow\nearrow$
1.25	0.405	0.220	0.160	0.139	**0.134**	0.139	0.160	0.220	0.405	$\searrow\nearrow$
1.5625	0.430	0.231	0.168	0.147	0.142	0.147	0.167	0.224	0.392	$\searrow\nearrow$
1.875	0.473	0.260	0.193	0.171	0.166	0.171	0.190	0.242	0.389	$\searrow\nearrow$
2.1875	0.543	0.315	0.242	0.217	0.211	0.215	0.233	0.277	0.392	$\searrow\nearrow$
2.5	**0.656**	0.408	0.326	0.296	0.286	0.288	0.300	0.331	0.401	$\searrow\nearrow$
	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	$\searrow\nearrow$	

5.2 Sinusoidal model

Now, we consider the following rates

$$\begin{aligned}\alpha(t) &= \alpha_0 + \alpha_1 \sin(\alpha_2 t)\\ \lambda(t) &= \lambda_0 + \lambda_1 \sin(\lambda_2 t)\end{aligned}$$

where $(\alpha_i)_{i=0,\dots,2}$ and $(\lambda_i)_{i=0,\dots,2}$ are such that α and λ are positive and their integral on $(0, g]$ are respectively equal to $\mathbb{E}N$ and $\mathbb{E}M$. For the numerical application, we chose $\alpha_2 = \lambda_2 = 2\pi L/500 = \pi$. This implies that $\alpha_0 = \mathbb{E}N/g = 5.75$ and $\lambda_0 = \mathbb{E}M/g = 1.25$; moreover, one needs to have $|\alpha_1| \leq 5.75$ and $|\lambda_1| \leq 1.25$ to ensure the rates are positive. The case $\alpha_1 = 0$ corresponds to the stationary model for the clone distribution and $\lambda_1 = 0$ represents the stationary case for anchors. When α_1 and λ_1 have the same sign, the rates are in phase; otherwise they are completely out of phase. Increasing the amplitudes $|\alpha_1|$ and $|\lambda_1|$ shows the effect of departing from the stationary model. Figure 2 presents the mean number of unanchored clones, the mean number of anchored islands, the average length of anchored islands and the mean proportion of the genome not covered by anchored islands, when α_1 varies from 0 to 5.75 and $\lambda_1 \in \{0; 0.2; 0.4; 0.6; 0.8; 1; 1.2\}$. The mean number of unanchored clones seems to not depend on α_1 and increases as λ_1 increases. Increasing either α_1 or λ_1 has the same effect on the average length of anchored islands and the mean proportion of genome not covered by anchored islands: the length decreases whereas the proportion increases. However, as α_1 increases, the mean number of anchored islands increases more rapidly when λ_1 is small. As λ_1 increases, the mean number of anchored islands increases for small α_1, seems to be constant for α_1 close to 4.75, and decreases for large α_1. The stationary case, $\alpha_1 = \lambda_1 = 0$ minimizes the mean number of anchored islands and the mean proportion of genome not covered by anchored islands and maximizes the average length of anchored islands.

Figure 2

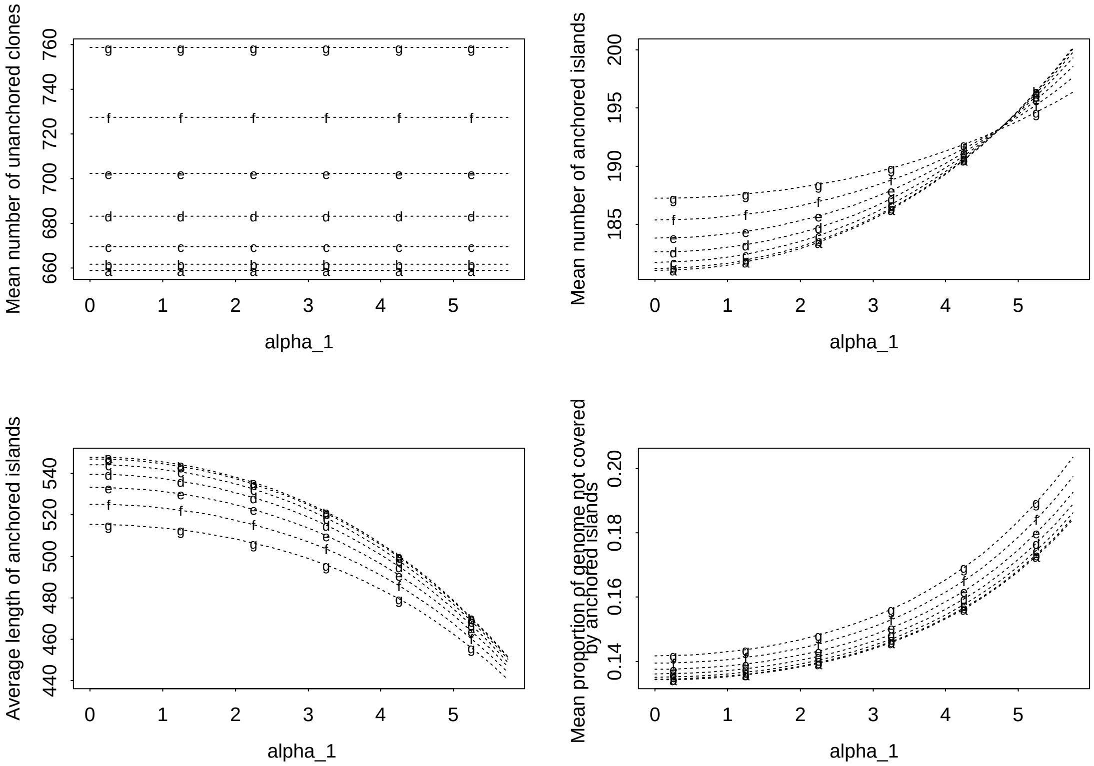

Figure 2: Effects of varying α_1 from 0 to 5.75 for seven values of λ_1: $\lambda_1 = 0$ (a), $\lambda_1 = 0.2$ (b), $\lambda_1 = 0.4$ (c), $\lambda_1 = 0.6$ (d), $\lambda_1 = 0.8$ (e), $\lambda_1 = 1.0$ (f), $\lambda_1 = 1.2$ (g).

6 Conclusion

Undeniably, the inhomogeneity in the clone and anchor locations along the genome substantially changes the predictions in a physical mapping project. Since the goal of a physical mapping project (at least the first step) is to obtain few long anchored islands and a small proportion of genome not covered by anchored islands, the previous applications clearly show that using homogeneous Poisson processes for clones and anchors provides an overly optimistic assessment of the progress of the mapping project. It would be very interesting to see, for instance, if the mean proportion of the genome covered by anchored islands is theoretically maximized, among all rate functions $\alpha(\cdot)$ and $\lambda(\cdot)$ satisfying $\int_0^g \alpha(t)dt = \mathbb{E}N$ and $\int_0^g \lambda(t)dt = \mathbb{E}M$, by the homogeneous rates $\alpha(t) \equiv \mathbb{E}N/g$ and $\lambda(t) \equiv \mathbb{E}M/g$.

The duality of the mean number of anchored islands, noticed by Arratia *et al.* (1991), between the constant clone rate and the constant anchor rate when the clone length is fixed, does not hold in the non-homogeneous case. However, if $I(\alpha, \lambda)$ denotes the mean number of anchored islands ending in $(0, g]$ where the parameters are respectively the rate of the clones and the rate of the anchors, one can show that, if α and λ are g-periodic then $I(\alpha, \lambda) = I(\lambda, \Theta\alpha)$, where $\Theta\alpha(t) = \alpha(t+1)$, $\forall t \in \mathbb{R}$.

The difficulty in practice remains to model the inhomogeneity occurring in the genome. The detection of regions rich or poor in restriction sites involved in the cloning process, for instance, could be a first step in characterizing some hotspots. Since longer DNA sequences are becoming available, modeling heterogeneity has become an important problem in sequence analysis.

A Appendix

A.1 Proof of Theorem 1

Suppose the given clone ending at t has length Q. Given Q and W_t, it is unanchored if and only if $W_t > Q$. Averaging over the distributions of Q and W_t gives

$$\begin{aligned} q_1(t) &= \mathbb{E}_{Q,W_t} \mathbb{I}\{W_t > Q\} = \int_0^\infty \mathbb{P}(W_t > \ell) f(\ell)\, d\ell \\ &= \int_0^\infty \exp\left(-\int_{t-\ell}^t \lambda(x)dx\right) f(\ell) d\ell\,, \end{aligned}$$

which proves statement (i) of Theorem 1.

The stochastic rate of the right end of unanchored clones process is defined by

$$\begin{aligned} \phi_u(t) &= \lim_{h\to 0} \frac{1}{h} \mathbb{P}(N_u[t, t+h] = 1 \mid \mathcal{G}_{t-}) \\ &= \lim_{h\to 0} \frac{1}{h} \sum_{i\in\mathbb{Z}} \mathbb{I}\{C_{i-1} < t \le C_i\} \mathbb{P}(C_i \in [t, t+h], \text{the clone ending at } C_i \text{ is unanchored} \mid \mathcal{G}_{t-}) \end{aligned}$$

where $\mathcal{G}_{t-}$ is the σ-algebra of the past events up to t:

$$\mathcal{G}_{t-} = \sigma((M(k), k < t), (N(u), u < t))\,. \tag{6}$$

We define the following notation: $\mathcal{G}_{t-,s} = \sigma((M(k), k < t), (N(u), u \le s))$ and $\mathcal{G}_s^N = \sigma(N(u), u \le s)$; we then have $\mathcal{G}_{t-} \subset \mathcal{G}_{t-,s}$ $\forall s \ge t$.

Suppose the clone ending at C_i has length Q; it is unanchored if and only if there are no anchors in $[C_i - Q, C_i]$ (see Figure 3). The event $\{C_i \in [t, t+h]$, no anchors in $[C_i - Q, C_i]\}$ can be written as the union of the two following disjoint events:

$$\begin{aligned} E_1 &= \{C_i \in [t, t+h]\} \cap \{\text{no anchors in } [C_i - Q, t)\} \cap \{\text{no anchors in } [t, C_i]\} \cap \{Q \ge C_i - t\} \\ E_2 &= \{C_i \in [t, t+h]\} \cap \{\text{no anchors in } [C_i - Q, C_i]\} \cap \{Q < C_i - t\}\,. \end{aligned}$$

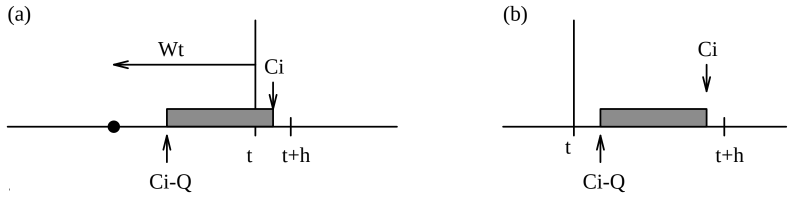

Figure 3: The clone ending at C_i is unanchored if and only if there are no anchors in $[C_i - Q, C_i]$. (a) $Q \geq C_i - t$, (b) $Q < C_i - t$.

We then want to calculate $\mathbb{P}(E_1 \mid \mathcal{G}_{t-}, Q)$ and $\mathbb{P}(E_2 \mid \mathcal{G}_{t-}, Q)$. The conditional probability of E_1 is obtained by conditioning with $\mathcal{G}_{t-,t+h}$. Since the events $\{C_i \in [t, t+h]\}$, $\{\text{no anchors in } [C_i - Q, t)\} = \{W_t > t - C_i + Q\}$ and $\{Q \geq C_i - t\}$ are $\mathcal{G}_{t-,t+h}$-measurable, and $\mathbb{P}(\text{no anchors in } [t, C_i] | \mathcal{G}_{t-,t+h}) = \mathbb{P}(\text{no anchors in } [t, C_i] | \mathcal{G}^N_{t+h}) = \exp\left(-\int_t^{C_i} \lambda(x)\, dx\right)$, we obtain

$$
\begin{aligned}
&\mathbb{P}(E_1 \mid \mathcal{G}_{t-}, Q) \\
&= \mathbb{E}(\mathbb{I}\{C_i \in [t, t+h]\}\mathbb{I}\{Q \geq C_i - t\}\mathbb{I}\{W_t > t - C_i + Q\} \exp\left(-\int_t^{C_i} \lambda(x)\, dx\right) \mid \mathcal{G}_{t-}, Q) \\
&= \int_t^{t+h} \mathbb{I}\{y - t \leq Q < W_t - t + y\} \exp\left(-\int_t^y \lambda(x)\, dx\right) \alpha(y)\, dy\,.
\end{aligned}
$$

In the same way, we calculate the conditional probability of E_2: since $\mathbb{P}(\text{no anchors in } [C_i - Q, C_i] \mid \mathcal{G}_{t-,t+h}) = \exp\left(-\int_{C_i - Q}^{C_i} \lambda(x)\, dx\right)$, we have

$$
\begin{aligned}
&\mathbb{P}(E_2 \mid \mathcal{G}_{t-}, Q) \\
&= \mathbb{E}(\mathbb{I}\{C_i \in [t, t+h]\}\mathbb{I}\{Q < C_i - t\} \exp\left(-\int_{C_i - Q}^{C_i} \lambda(x)\, dx\right) \mid \mathcal{G}_{t-}, Q) \\
&= \int_t^{t+h} \mathbb{I}\{Q < y - t\} \exp\left(-\int_{y-Q}^y \lambda(x)\, dx\right) \alpha(y)\, dy\,.
\end{aligned}
$$

Therefore, it follows that

$$
\begin{aligned}
\mathbb{P}(E_1 \mid \mathcal{G}_{t-}) &= \int_t^{t+h} \mathbb{P}_Q(y - t \leq Q < W_t - t + y) \exp\left(-\int_t^y \lambda(x)\, dx\right) \alpha(y)\, dy \\
\mathbb{P}(E_2 \mid \mathcal{G}_{t-}) &= \int_t^{t+h} \mathbb{E}_Q\left(\mathbb{I}\{Q < y - t\} \exp\left(-\int_{y-Q}^y \lambda(x)\, dx\right)\right) \alpha(y)\, dy\,.
\end{aligned}
$$

Finally, $\lim_{h\to 0} h^{-1}\mathbb{P}(E_1 \mid \mathcal{G}_{t-}) = \mathbb{P}_Q(Q < W_t)\alpha(t)$ and $\lim_{h\to 0} h^{-1}\mathbb{P}(E_2 \mid \mathcal{G}_{t-}) = 0$ leading to $\phi_u(t) = \mathbb{P}_Q(Q < W_t)\alpha(t)$. This establishes statement (ii) of Theorem 1.
We note the component $\{Q < C_i - t\}$ brings no contribution to the limit defining ϕ_u as $C_i \in [t, t+h]$ and $h \to 0$.

By definition, $\mathbb{E}N_u((0,g]) = \mathbb{E}\int_0^g \phi_u(t)\,dt$ where ϕ_u is the rate of the unanchored clones process and is given by (ii). Since ϕ_u is a positive function, we can exchange the expectation and the integral using Tonelli-Fubini's Theorem, and it follows that

$$\mathbb{E}N_u(g) = \mathbb{E}\int_0^g \phi_u(t)\,dt = \int_0^g \alpha(t)\mathbb{E}_{Q,W_t}\mathbb{I}\{Q < W_t\}dt = \int_0^g \alpha(t)q_1(t)dt\,.$$

This establishes statement (iii) of Theorem 1. ■

A.2 Proof of Theorem 2

Suppose the given clone ending at t has length Q. Given Q and W_t, it is the rightmost clone of an anchored island if and only if $W_t \leq Q$ and no other clones cover both t and $t - W_t$. Averaging over the distributions of Q and W_t gives

$$\begin{aligned} p_1(t) &= \mathbb{E}_{Q,W_t}\left(J(t-W_t;W_t)\mathbb{I}\{W_t \leq Q\}\right) \\ &= \int_0^\infty J(t-w;w)h_t(w)\mathcal{F}(w)dw \\ &= \int_0^\infty J(t-w;w)\mathcal{F}(w)\lambda(t-w)\exp\left(-\int_{t-w}^t \lambda(x)dx\right)dw \end{aligned}$$

which proves statement (i) of Theorem 2.

The stochastic rate of the right end of anchored islands process is defined by

$$\begin{aligned} \psi(t) &= \lim_{h\to 0}\frac{1}{h}\mathbb{P}(I_a[t,t+h] = 1 \mid \mathcal{G}_{t-}) \\ &= \lim_{h\to 0}\frac{1}{h}\sum_{i\in\mathbb{Z}}\mathbb{I}\{C_{i-1} < t \leq C_i\}\mathbb{P}(C_i \in [t,t+h], \text{the clone ending at } C_i \text{ is the} \\ &\qquad\qquad \text{rightmost clone of an anchored island} \mid \mathcal{G}_{t-}) \end{aligned}$$

where $\mathcal{G}_{t-}$ is the σ-algebra defined in (6).
Suppose the clone ending at C_i has length Q. The event $\{C_i \in [t,t+h]$, the clone ending at C_i is the rightmost clone of an anchored island$\}$ can be written as the disjoint union of the following events E_1, E_2 and E_3 (see Figure 4):

$$\begin{aligned} E_1 &= \{C_i \in [t,t+h]\} \cap \{Q \geq C_i - t\} \cap \{W_t \leq t - C_i + Q\} \cap \{\text{no anchors in } [t,C_i]\} \\ &\qquad \cap\{\text{no clones cover } [t-W_t, C_i]\} \\ E_2 &= \{C_i \in [t,t+h]\} \cap \{Q \geq C_i - t\} \cap \{\text{at least one anchor in } [t,C_i]\} \\ &\qquad \cap\{\text{no clones cover } [C_i - W_{C_i}, C_i]\} \\ E_3 &= \{C_i \in [t,t+h]\} \cap \{Q < C_i - t\} \cap \{\text{at least one anchor in } [C_i - Q, C_i]\} \\ &\qquad \cap\{\text{no clones cover } [C_i - W_{C_i}, C_i]\} \end{aligned}$$

where W_{C_i} is the distance back from C_i to the previous anchor.

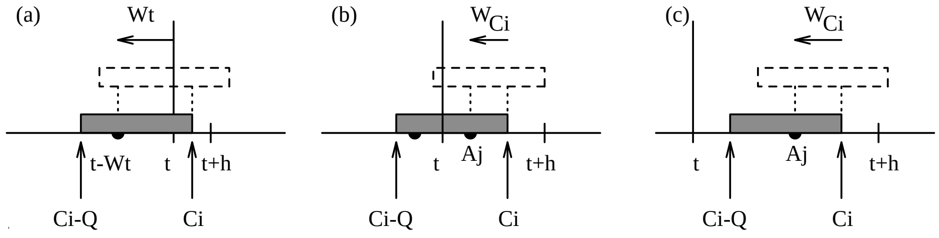

Figure 4: The clone ending at C_i is the rightmost clone of an anchored island if and only if one of the three following cases holds: (a) $Q \geq C_i - t$, $W_t \leq t - C_i + Q$, there are no anchors in $[t, C_i]$ and no clones cover $[t - W_t, C_i]$, (b) $Q \geq C_i - t$, there is at least one anchor in $[t, C_i]$ and no clones cover $[C_i - W_{C_i}, C_i]$, (c) $Q < C_i - t$, there is at least one anchor in $[C_i - Q, C_i]$ and no clones cover $[C_i - W_{C_i}, C_i]$.

Thus, we need to calculate three terms:

$$\begin{aligned}
&\mathbb{P}(C_i \in [t, t+h], \text{the clone ending at } C_i \text{ is the rightmost clone of an anchored island} \mid \mathcal{G}_{t-}) \\
&\quad = \mathbb{E}_Q\left(\mathbb{P}(E_1 \mid \mathcal{G}_{t-}, Q)\right) + \mathbb{E}_Q\left(\mathbb{P}(E_2 \mid \mathcal{G}_{t-}, Q)\right) + \mathbb{E}_Q\left(\mathbb{P}(E_3 \mid \mathcal{G}_{t-}, Q)\right).
\end{aligned}$$

As in Appendix A.1, we expect the third term not to contribute to the limit defining ψ because it involves the events $\{C_i \in [t, t+h]\}$ and $\{Q < C_i - t\}$, and h converges to zero. In the same way, the second term will no contribute either because it involves the events $\{C_i \in [t, t+h]\}$ and $\{\text{at least one anchor in } [t, C_i]\}$. The first term $\mathbb{E}_Q\left(\mathbb{P}(E_1 \mid \mathcal{G}_{t-}, Q)\right)$ appears then to be the most important. We use the σ-algebra notations introduced in Section A.1.
We start calculating the conditional probability of E_1 conditioning by $\mathcal{G}_{t-,t+h}$. Since the events $\{Q \geq C_i - t\}$, $\{C_i \in [t, t+h]\}$ and $\{W_t \leq t - C_i + Q\}$ are $\mathcal{G}_{t-,t+h}$-measurable, and $\mathbb{P}(\text{no anchors in } [t, C_i] \mid \mathcal{G}_{t-,t+h}) = \exp\left(-\int_t^{C_i} \lambda(x)\, dx\right)$ and $\mathbb{P}(\text{no clones cover } [t - W_t, C_i] \mid \mathcal{G}_{t-,t+h}) = J(t - W_t; C_i - t + W_t)$, we have

$$\begin{aligned}
&\mathbb{E}_Q\left(\mathbb{P}(E_1 \mid \mathcal{G}_{t-}, Q)\right) \\
&\quad = \mathbb{E}_Q\left(\mathbb{E}(\mathbb{I}\{C_i \in [t, t+h]\}\mathbb{I}\{W_t \leq t - C_i + Q\} \exp\left(-\int_t^{C_i} \lambda(x)\, dx\right) \times \right. \\
&\qquad\qquad\qquad\qquad \left. J(t - W_t; C_i - t + W_t) \mid \mathcal{G}_{t-}, Q)\right) \\
&\quad = \int_t^{t+h} \mathcal{F}(W_t - t + y) \exp\left(-\int_t^y \lambda(x)\, dx\right) J(t - W_t; y - t + W_t)\alpha(y)\, dy,
\end{aligned}$$

and

$$\lim_{h \to 0} \frac{1}{h} \mathbb{E}_Q\left(\mathbb{P}(E_1 \mid \mathcal{G}_{t-}, Q)\right) = \mathcal{F}(W_t) J(t - W_t; W_t)\alpha(t).$$

We now calculate the conditional probability of E_2. Let $A_j = C_i - W_{C_i}$ be the anchor location such that $A_j < C_i \leq A_{j+1}$ (see Figure 4). We can then rewrite the event E_2

as $E_2 = \{C_i \in [t, t+h]\} \cap \{Q \geq C_i - t\} \cap \{A_j \in [t, C_i)\} \cap \{\text{no clones cover } [A_j, C_i]\}$. Therefore,

$$\begin{aligned}\mathbb{P}(E_2 \mid \mathcal{G}_{t-}, Q) &= \sum_{j \in \mathbb{Z}} \mathbb{I}\{A_j < C_i \leq A_{j+1}\} \mathbb{P}(C_i \in [t, t+h], A_j \in [t, C_i), \\ &\qquad \text{no clones cover } [A_j, C_i], Q \geq C_i - t \mid \mathcal{G}_{t-}, Q)\,.\end{aligned}$$

Now, we use that the events $\{C_i \in [t, t+h]\}$, $\{A_j \in [t, C_i)\}$ and $\{Q \geq C_i - t\}$ are $\mathcal{G}_{t+h}$-measurable, and $\mathbb{P}(\text{no clones cover } [A_j, C_i] \mid \mathcal{G}_{t+h}) = J(A_j; C_i - A_j)$. Conditioning by $\mathcal{G}_{t+h}$, it follows

$$\begin{aligned}&\mathbb{P}(C_i \in [t, t+h], A_j \in [t, C_i), \text{no clones cover } [A_j, C_i], Q \geq C_i - t \mid \mathcal{G}_{t-}) \\ &\quad = \mathbb{E}_Q\left(\mathbb{E}(\mathbb{I}\{C_i \in [t, t+h]\} \mathbb{I}\{A_j \in [t, C_i)\} J(A_j; C_i - A_j) \mathbb{I}\{Q \geq C_i - t\} \mid \mathcal{G}_{t-}, Q)\right) \\ &\quad = \int_{y=t}^{t+h} \int_{x=t}^{y} J(x; y-x) \mathcal{F}(y-t) \alpha(y) \lambda(x)\, dx\, dy\end{aligned}$$

and

$$\lim_{h \to 0} \frac{1}{h} \mathbb{E}_Q\left(\mathbb{P}(E_2 \mid \mathcal{G}_{t-}, Q)\right) = 0\,.$$

Similarly, we calculate the conditional probability of E_3. We can then rewrite the event E_3 as $E_3 = \{C_i \in [t, t+h]\} \cap \{Q < C_i - t\} \cap \{A_j \in [C_i - Q, C_i)\} \cap \{\text{no clones cover } [A_j, C_i]\}$. Therefore,

$$\begin{aligned}\mathbb{P}(E_3 \mid \mathcal{G}_{t-}, Q) &= \sum_{j \in \mathbb{Z}} \mathbb{I}\{A_j < C_i \leq A_{j+1}\} \mathbb{P}(C_i \in [t, t+h], A_j \in [C_i - Q, C_i), \\ &\qquad \text{no clones cover } [A_j, C_i], Q < C_i - t \mid \mathcal{G}_{t-}, Q)\,.\end{aligned}$$

Now, we use that the events $\{C_i \in [t, t+h]\}$, $\{A_j \in [C_i - Q, C_i)\}$ and $\{Q < C_i - t\}$ are $\mathcal{G}_{t+h}$-measurable, and $\mathbb{P}(\text{no clones cover } [A_j, C_i] \mid \mathcal{G}_{t+h}) = J(A_j; C_i - A_j)$. It follows

$$\begin{aligned}&\mathbb{P}(C_i \in [t, t+h], A_j \in [C_i - Q, C_i), \text{no clones cover } [A_j, C_i], Q < C_i - t \mid \mathcal{G}_{t-}, Q) \\ &\quad = \mathbb{E}(\mathbb{I}\{C_i \in [t, t+h]\} \mathbb{I}\{A_j \in [C_i - Q, C_i)\} J(A_j; C_i - A_j) \mathbb{I}\{Q < C_i - t\} \mid \mathcal{G}_{t-}, Q) \\ &\quad = \int_{y=t}^{t+h} \int_{x=y-Q}^{y} J(x; y-x) \mathbb{I}\{Q < y-t\} \alpha(y) \lambda(x)\, dx\, dy\,.\end{aligned}$$

Averaging over the distribution of Q gives

$$\mathbb{E}_Q\left(\mathbb{P}(E_3 \mid \mathcal{G}_{t-}, Q)\right) = \int_{y=t}^{t+h} \alpha(y) \mathbb{E}_Q\left(\mathbb{I}\{Q < y-t\} \int_{x=y-Q}^{y} J(x; y-x) \lambda(x)\, dx\right) dy\,.$$

and

$$\lim_{h \to 0} \frac{1}{h} \mathbb{E}_Q\left(\mathbb{P}(E_3 \mid \mathcal{G}_{t-}, Q)\right) = 0\,.$$

Finally, $\psi(t) = \alpha(t) J(t - W_t; W_t) \mathcal{F}(W_t)$, establishing statement (ii) of Theorem 2.

Since $\mathbb{E}I_a(g) = \mathbb{E}\int_0^g \psi(t)\,dt$ and ψ is a positive function, we can exchange the integral and the expectation leading to

$$\mathbb{E}I_a(g) = \int_0^g \mathbb{E}_{Q,W_t}\left(\alpha(t)J(t-W_t;W_t)\,\mathbb{1}\{Q \geq W_t\}\right)dt = \int_0^g \alpha(t)p_1(t)\,dt\,.$$

This completes the proof of the theorem. ∎

A.3 Proof of Theorem 3

Suppose the given clone ending at t has length Q. V_{t-Q} is the distance forward from $t-Q$ to the next anchor and W_t is the distance back from t to the previous anchor. Given Q, V_{t-Q} and W_t, the given clone is a singleton anchored island if and only if $V_{t-Q}+W_t \leq Q$, no clones cover $[t-Q, t-Q+V_{t-Q}]$ and no clones cover $[t-W_t, t]$. Using Proposition 2 gives

$$\begin{aligned}
p_2(t) &= \mathbb{E}_{Q,V_{t-Q},W_t}\left(\frac{J(t-W_t;W_t)J(t-Q;V_{t-Q})}{J(t-Q;Q)}\,\mathbb{1}\{V_{t-Q}+W_t \leq Q\}\right)\\
&= \mathbb{E}_{Q,V_{t-Q},W_t}\left(\frac{J(t-W_t;W_t)J(t-Q;V_{t-Q})}{J(t-Q;Q)}\,\mathbb{1}\{V_{t-Q}+W_t < Q\}\right)\\
&\quad + \mathbb{E}_{Q,V_{t-Q},W_t}\left(\frac{J(t-W_t;W_t)J(t-Q;Q-W_t)}{J(t-Q;Q)}\,\mathbb{1}\{V_{t-Q}+W_t = Q\}\right).
\end{aligned}$$

Given Q, we need to determine the joint distribution of (V_{t-Q}, W_t).
If $0 \leq v+w < \ell$ then the intervals $(t-\ell, t-\ell+v]$ and $[t-w, t)$ are disjoint, and it follows that

$$\begin{aligned}
\mathbb{P}(W_t > w,\, V_{t-\ell} > v) &= \mathbb{P}(\text{no anchors in } [t-w,t) \text{ and no anchors in } (t-\ell, t-\ell+v])\\
&= \mathbb{P}(\text{no anchors in } [t-w,t))\,\mathbb{P}(\text{no anchors in } (t-\ell, t-\ell+v])\\
&= \mathbb{P}(W_t > w)\,\mathbb{P}(V_{t-\ell} > v)\,.
\end{aligned}$$

Therefore, given Q and $V_{t-Q}+W_t < Q$, the conditional joint density of (V_{t-Q}, W_t) is $g_{t-Q}(v)h_t(w)$, where the density functions g and h are defined by (1) and (2).
If $v+w = \ell$,

$$\begin{aligned}
\mathbb{P}(W_t \in [w, w+dw[,\, V_{t-\ell} = v) &= \mathbb{P}(V_{t-\ell} = \ell - w \mid W_t = w)\,\mathbb{P}(W_t \in [w, w+dw[)\\
&= \mathbb{P}(\text{no anchors in } (t-\ell, t-w))\,h_t(w)dw\\
&= \exp\left(-\int_{t-\ell}^{t-w}\lambda(x)dx\right)h_t(w)dw\,.
\end{aligned}$$

Therefore, given Q and $V_{t-Q}+W_t=Q$, the conditional joint density of (V_{t-Q},W_t) is $\lambda(t-w)\exp\left(-\int_{t-Q}^{t}\lambda(x)dx\right)$. Hence,

$$\begin{aligned}
p_2(t) &= \int_{\ell=0}^{\infty}\int_{w=0}^{\ell}\int_{v=0}^{\ell-w}\frac{J(t-w;w)J(t-\ell;v)}{J(t-\ell;\ell)}\,g_{t-\ell}(v)h_t(w)f(\ell)\,dvdwd\ell \\
&\quad+\int_{\ell=0}^{\infty}\int_{w=0}^{\ell}\frac{J(t-w;w)J(t-\ell;\ell-w)}{J(t-\ell;\ell)}\,\lambda(t-w)\exp\left(-\int_{t-\ell}^{t}\lambda(x)dx\right)f(\ell)dwd\ell\,.
\end{aligned}$$

This establishes statement (ii) of Theorem 3.

The stochastic rate $\gamma(t)$ of the right end of singleton anchored islands process is defined by

$$\begin{aligned}
\gamma(t) &= \lim_{h\to 0}\frac{1}{h}\sum_{i\in\mathbb{Z}}\mathbb{1}\{C_{i-1}<t\le C_i\}\mathbb{P}(C_i\in[t,t+h],\text{ the clone ending at }C_i\text{ is a} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{singleton anchored clone}\mid\mathcal{G}_{t-})
\end{aligned}$$

where $\mathcal{G}_{t-}$ has been defined in (6).

To proceed, the main idea is to decompose the event $\{C_i\in[t,t+h],$ the clone ending at C_i is a singleton anchored clone$\}$ into an union of disjoint events and to use same techniques we utilized to calculate the rates $\phi_u(t)$ and $\psi(t)$ in Appendices A.1 and A.2. Suppose the clone ending at C_i has length Q. We have to distinguish the case $Q\ge C_i-t$ from $Q<C_i-t$, and then the case where there is at least one anchor in $[t,C_i]$ from the case where there are no anchors in $[t,C_i]$. Since $C_i\in[t,t+h]$ and h will tend to zero, we see that the components $\{Q<C_i-t\}$ and $\{$at least one anchor in $[t,C_i]\}$ provide a null contribution to the limit term defining $\gamma(t)$ (see proofs of Theorems 1 and 2 for more details).

We then focus on the case where $Q\ge C_i-t$ and there are no anchors in $[t,C_i]$. In this case, for the clone ending at C_i to be anchored, one needs to have at least one anchor in $[C_i-Q,t)$, in other words one needs either $W_t<t-C_i+Q$ either $W_t=t-C_i+Q$. It is not difficult to convince oneself that the event $\{W_t=t-C_i+Q\}$ will also lead to a null contribution since $W_t\in\mathcal{G}_{t-}$ and Q is a continuous random variable. Moreover, $W_t<t-C_i+Q$ is equivalent to $W_t+V_{C_i-Q}\le t-C_i+Q$. It allows us to write that

$$\gamma(t)=\lim_{h\to 0}\frac{1}{h}\sum_{i\in\mathbb{Z}}\mathbb{1}\{C_{i-1}<t\le C_i\}\mathbb{P}(C_i\in[t,t+h],\,E\mid\mathcal{G}_{t-})$$

where E is the following event:

$$\begin{aligned}
E &= \{Q\ge C_i-t\}\cap\{\text{no anchors in }[t,C_i]\}\cap\{W_t+V_{C_i-Q}\le t-C_i+Q\} \\
&\quad\cap\{\text{no clones cover }[t-W_t,C_i]\}\cap\{\text{no clones cover }[C_i-Q,C_i-Q+V_{C_i-Q}]\}\,.
\end{aligned}$$

Since $C_i\in[t,t+h]$, the events $\{Q\ge C_i-t\}$ and $\{W_t+V_{C_i-Q}\le t-C_i+Q\}$ are

$\mathcal{G}_{t-,t+h}$-measurable, $\mathbb{P}(\text{no anchors in } [t, C_i] | \mathcal{G}_{t-,t+h}) = \exp(-\int_t^{C_i} \lambda(x)\, dx)$ and, using Proposition 2,

$$\begin{aligned}&\mathbb{P}(\text{no clones cover } [t - W_t, C_i], \text{no clones cover } [C_i - Q, C_i - Q + V_{C_i - Q}] \mid \mathcal{G}_{t-,t+h}) \\ &\quad = \frac{J(C_i - Q; V_{C_i-Q}) J(t - W_t; C_i - t + W_t)}{J(C_i - Q; Q)} .\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{P}(C_i \in [t, t+h],\, E \mid \mathcal{G}_{t-}) = \mathbb{E}_Q \left(\mathbb{P}(C_i \in [t, t+h],\, E \mid \mathcal{G}_{t-}, Q)\right) \\
&= \mathbb{E}_Q \Bigg(\mathbb{E}\Bigg(\mathbb{1}\{C_i \in [t, t+h]\} \mathbb{1}\{Q \geq C_i - t\} \mathbb{1}\{W_t + V_{C_i-Q} \leq t - C_i + Q\} \exp\left(-\int_t^{C_i} \lambda(x)\, dx\right) \\
&\qquad \frac{J(C_i - Q; V_{C_i-Q}) J(t - W_t; C_i - t + W_t)}{J(C_i - Q; Q)} \mid \mathcal{G}_{t-}, Q \Bigg)\Bigg) \\
&= \int_t^{t+h} \exp\left(-\int_t^y \lambda(x)\, dx\right) J(t - W_t; y - t + W_t) \\
&\qquad \mathbb{E}_Q \left(\frac{J(y - Q; V_{y-Q})}{J(y - Q; Q)} \mathbb{1}\{Q \geq y - t + W_t + V_{y-Q}\}\right) \alpha(y)\, dy .
\end{aligned}$$

Therefore,

$$\lim_{h\to 0} \frac{1}{h} \mathbb{P}(C_i \in [t, t+h],\, E | \mathcal{G}_{t-}) = \alpha(t) J(t - W_t; W_t) \mathbb{E}_Q \left(\frac{J(t - Q; V_{t-Q})}{J(t - Q; Q)} \mathbb{1}\{Q \geq W_t + V_{t-Q}\}\right)$$

which establishes statement (ii) of Theorem 3.

Finally, $\mathbb{E} I_s(g) = \mathbb{E}\int_0^g \gamma(t)\, dt = \int_0^g \mathbb{E}\gamma(t)\, dt = \int_0^g \alpha(t) p_2(t) dt$. ∎

A.4 Proofs of Theorem 4 and Proposition 3

Proof of Theorem 4. Given V_t, the distance forward from t to the next anchor, and W_t, the distance back from t to the previous anchor, the point t is not covered by an anchored island if and only if no clones cover $[t - W_t, t]$ and no clones cover $[t, t + V_t]$. Using Proposition 2 gives

$$r_0(t) = \mathbb{E}_{V_t, W_t} \left(\frac{J(t - W_t; W_t) J(t; V_t)}{J(t - W_t; V_t + W_t)} \right) .$$

Now, V_t and W_t are independent, so we have

$$r_0(t) = \int_0^\infty \int_0^\infty \frac{J(t; v) J(t - w; w)}{J(t - w; v + w)} g_t(v) h_t(w) dv dw ,$$

where the density functions g and h are given by (1) and (2).

The mean proportion of genome not covered by anchored islands is then

$$\frac{1}{g} \mathbb{E} \int_0^g \mathbb{1}\{t \text{ is not covered by an anchored island}\}\, dt = \frac{1}{g} \int_0^g r_0(t)\, dt ,$$

using Tonelli-Fubini's Theorem, which completes the proof of the theorem. ∎

Proof of Proposition 3. We intend to prove only the first statement of the proposition since the proof is the same to prove statement (ii).
Provided $(a_g)_{g\in\mathbb{R}}$ is a sequence of reals converging to infinity as g tends to infinity, we have from Lenglart (1977)

$$\forall g > 0, \forall \delta > 0, \forall \eta > 0, \mathbb{P}\left(\sup_{s\in[0,g]} \frac{1}{|a_s|}\left|I_a(s) - \int_0^s \psi(t)\,dt\right| > \eta\right) \leq \frac{\delta}{\eta^2} + \mathbb{P}\left(\frac{\int_0^g \psi(t)\,dt}{a_g^2} > \eta\right).$$

From (A_1), $\mathbb{E}\int_0^g \psi(t)\,dt \to \infty$ as $g \to \infty$; using Lenglart's result with $a_g = \mathbb{E}\int_0^g \psi(t)\,dt$ leads directly to the convergence in probability of the ratio $I_a(g)/\mathbb{E}\int_0^g \psi(t)\,dt$ to 1 since (A_1) ensures that $\int_0^g \psi(t)\,dt/\mathbb{E}\int_0^g \psi(t)\,dt \to 1$ in probability as $g \to \infty$.
Moreover, $\mathbb{E}\int_0^g \psi(t)\,dt = \mathbb{E}I_a(g) = \int_0^g \alpha(t)p_1(t)\,dt$, which establishes the proposition. ∎

A.5 Proofs of Theorems 5, 6 and 7

Proof of Theorem 5. Let $I_a(g)$ be the number of anchored islands whose right end fall in $(0,g]$ and S_j be the length of the jth anchored island ending in $(0,g]$. The average length of the anchored islands in $(0,g]$ is defined by

$$\overline{S}_g = \frac{1}{I_a(g)}\sum_{j=1}^{I_a(g)} S_j\,.$$

Let $r_2(t)$ be the probability that t is covered by exactly two anchored islands. The quantity $\int_0^g 1 - r_0(t) + r_2(t)\,dt$ represents then the sum of the mean length of the genome covered by at least one anchored island and the mean length of the genome covered by exactly two anchored islands. Since a point can fall at most in two anchored islands, it follows that

$$\frac{1}{g}\sum_{j=1}^{I_a(g)} S_j - \frac{1}{g}\int_0^g 1 - r_0(t) + r_2(t)dt \xrightarrow{\mathbb{P}} 0\,,$$

as $g \to \infty$. Hence, writing $\overline{S}_g$ as

$$\begin{aligned}\overline{S}_g &= \frac{g}{\int_0^g \alpha(t)p_1(t)dt}\,\frac{\int_0^g \alpha(t)p_1(t)dt}{I_a(g)}\left(\frac{1}{g}\sum_{j=1}^{I_a(g)} S_j - \frac{1}{g}\int_0^g 1 - r_0(t) + r_2(t)dt\right)\\ &\quad + \frac{\int_0^g \alpha(t)p_1(t)dt}{I_a(g)}\,\frac{\int_0^g 1 - r_0(t) + r_2(t)dt}{\int_0^g \alpha(t)p_1(t)dt}\,,\end{aligned}$$

and using the assumptions (A_1), (A_1') and Proposition 3, we have

$$\overline{S}_g - \frac{\int_0^g 1 - r_0(t) + r_2(t)dt}{\int_0^g \alpha(t)p_1(t)dt} \xrightarrow{\mathbb{P}} 0\,.$$

We now calculate the probability $r_2(t)$ that the point t is covered by two anchored islands.

V_t is the distance forward from t to the next anchor and W_t is the distance back from t to the previous anchor. Given V_t and W_t, the point t is covered by two anchored islands if and only if at least one clone covers $[t-W_t, t]$, no clones cover $[t-W_t, t+V_t]$ and at least one clone covers $[t, t+V_t]$. From Proposition 2(ii) and averaging over the joint distribution of (V_t, W_t), we obtain

$$\begin{aligned} r_2(t) &= \mathbb{E}_{(V_t,W_t)}\left(\left(1-\frac{J(t-W_t;W_t)}{J(t-W_t;V_t+W_t)}\right)(J(t-W_t;V_t+W_t)-J(t;V_t))\right) \\ &= \mathbb{E}_{(V_t,W_t)}\left(\frac{(J(t-W_t;V_t+W_t)-J(t-W_t;W_t))(J(t-W_t;V_t+W_t)-J(t;V_t))}{J(t-W_t;V_t+W_t)}\right). \end{aligned}$$

Therefore, because of the independence of V_t and W_t, we have

$$\begin{aligned} r_2(t)-r_0(t) &= \mathbb{E}_{V_t,W_t}\left(J(t-W_t;V_t+W_t)-J(t;V_t)-J(t-W_t;W_t)\right) \\ &= \int_0^\infty \int_0^\infty [J(t-w;v+w)-J(t;v)-J(t-w;w)]g_t(v)h_t(w)dvdw \end{aligned}$$

which establishes the theorem. ■

Proof of Theorem 6. Let $I_a(g)$ be the number of anchored islands whose right-hand ends fall in the genome $(0,g]$, and Y_j be the number of clones in the jth anchored island, $j=1,\ldots,I_a(g)$. The average number of clones per anchored island in $(0,g]$ is

$$\overline{Y}_g = \frac{1}{I_a(g)}\sum_{j=1}^{I_a(g)} Y_j.$$

Let $N_a(g)$ be the number of anchored clones whose right-hand ends fall in $(0,g]$, and Y'_g be the difference between $\sum_{j=1}^{I_a(g)} Y_j$ and $N_a(g)$. Then,

$$\begin{aligned} \overline{Y}_g &= \frac{1}{I_a(g)}(N_a(g)+Y'_g), \\ &= \frac{\int_0^g \alpha(t)p_1(t)dt}{I_a(g)}\,\frac{\int_0^g \alpha(t)(1-q_1(t))dt}{\int_0^g \alpha(t)p_1(t)dt}\,\frac{1}{\int_0^g \alpha(t)(1-q_1(t))dt}(N_a(g)+Y'_g). \end{aligned}$$

The absolute value of Y'_g is less than the number of clones in the anchored islands that cover 0 or g. As g increases to infinity, the mean number of anchored island in $(0,g]$ tends to infinity from (A_1). Since a point can be covered by at most two anchored islands, necessarily the mean length of the anchored island(s) covering 0 or g is finite. Using that the integral of α on a finite set is finite, we have that $|Y'_g|$ has finite mean. Therefore, the assumption (A_2) ensures that $\left(\int_0^g \alpha(t)(1-q_1(t))dt\right)^{-1} Y'_g \to 0$ almost surely as $g\to\infty$, and using Proposition 3 gives

$$\overline{Y}_g - \frac{\int_0^g \alpha(t)(1-q_1(t))dt}{\int_0^g \alpha(t)p_1(t)dt} \xrightarrow{\mathbb{P}} 0 \quad \text{as } g\to\infty.$$

This completes the proof of the theorem. ■

PROOF OF THEOREM 7. (i) The stochastic rate ν of the process of anchors covered by clones is defined by

$$\nu(t) \;=\; \lim_{h\to 0}\frac{1}{h}\sum_{j\in\mathbb{Z}} \mathbb{1}\{A_{j-1} < t \le A_j\}\mathbb{P}(A_j \in [t,t+h], \text{the anchor at } A_j \text{ is covered by a clone} \mid \mathcal{G}_{t-})$$

where $\mathcal{G}_{t-}$ is defined in (6).

The event $\{A_j \in [t,t+h]$, the anchor at A_j is covered by a clone$\}$ is independent of the σ-algebra $\mathcal{G}_{t-}$ because "the anchor at A_j is covered by a clone" involves only clones ending after A_j, and since $A_j \in [t,t+h]$, the independence follows from the Poisson property of the clones and anchors processes. Moreover,

$$\mathbb{P}(A_j \in [t,t+h], \text{the anchor at } A_j \text{ is covered by a clone}) = \int_t^{t+h} (1-J(x;0))\lambda(x)\,dx\,.$$

Therefore,

$$\nu(t) = (1-J(t;0))\lambda(t)\,.$$

Since $\nu(t)$ is deterministic, the process of anchors covered by clones is a Poisson process.

(ii) Let $I_a(g)$ be the number of anchored islands ending in $(0,g]$ and H_j be the number of anchors in the jth anchored island whose right end fall in $(0,g]$. The average number of anchors per anchored island is

$$\overline{H}_g = \frac{1}{I_a(g)}\sum_{j=1}^{I_a(g)} H_j\,.$$

Let $M_c(g)$ be the number of anchors in $(0,g]$ covered by clones, and H'_g be the difference between $\sum_{j=1}^{I_a(g)} H_j$ and $M_c(g)$. By a similar argument used to prove Theorem 6, the absolute value of H'_g has a finite mean since it is less than the number of anchors in the anchored islands that cover 0 or g. Therefore, assumption (A_1) ensures that $H'_g/\int_0^g \alpha(t)p_1(t)\,dt \to 0$ as $g\to\infty$. From (i), the process of anchors covered by clones is a Poisson process with rate $\nu(t) = \lambda(t)(1-J(t;0))$; using the strong law of large numbers (Kingman, 1993) we obtain, with probability 1

$$\lim_{g\to\infty}\frac{M_c(g)}{\int_0^g \lambda(t)(1-J(t;0))\,dt} = 1\,.$$

Now, writing $\overline{H}_g$ like

$$\overline{H}_g = \frac{\int_0^g \alpha(t)p_1(t)\,dt}{I_a(g)}\left(\frac{\int_0^g \lambda(t)(1-J(t;0))\,dt}{\int_0^g \alpha(t)p_1(t)\,dt}\,\frac{M_c(g)}{\int_0^g \lambda(t)(1-J(t;0))\,dt} + \frac{H'_g}{\int_0^g \alpha(t)p_1(t)\,dt}\right)$$

and using (A_1) leads to

$$\overline{H}_g - \frac{\int_0^g \lambda(t)(1 - J(t;0))\, dt}{\int_0^g \alpha(t) p_1(t)\, dt} \xrightarrow{\mathbb{P}} 0\,.$$

■

Acknowledgment: I wish to thank Simon Tavaré and Warren Ewens for calling my attention to hotspot effects in physical mapping by anchoring random clones and for their encouragement to study non-homogeneous models. Warren Ewens generously made his unpublished report available. I thank Odile Pons for helpful discussions. This work was carried out while visiting the Center for Computational Biology, University of Southern California, and was supported in part by NSF Grant BIR 95-04393 and the NATO.

REFERENCES

ARRATIA, R., LANDER, E. S., TAVARÉ, S. and WATERMAN, M. S. (1991). Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics* **11** 806–827.

BARILLOT, E., DAUSSET, J. and COHEN, D. (1991). Theoretical analysis of a physical mapping strategy using random single-copy landmarks. *Proc. Natl. Acad. Sci.* **88** 3917–3921.

CHUMAKOV, I., RIGAULT, P., GUILLOU, S., OUGEN, P., BILLAUT, A., GUASCONI, G., GERVY, P., LE GALL, I., SOULARUE, P. and GRINAS, L. (1992). Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359** 380–387.

COULSON, A., SULSTON, J., BRENNER, S. and KARN, J. (1986). Toward a physical map of the genome of the nematode, *Caenorhabditis elegans. Proc. Natl. Acad. Sci.* **83** 7821–7825.

ETEMADI, N. (1983). On the laws of large numbers for nonnegative random variables. *J. Mult. Analysis* **13** 187–193.

EWENS, W. J. (1996). Simulation results for anchored clones. *Research report number 252*, Department of Mathematics, Monash University, Australia.

EWENS, W. J., BELL, C. J., DONNELLY, P. J., DUNN, P., MATALLANA, E. and ECKER, J. R. (1991). Genome mapping with anchored clones: Theoretical aspects. *Genomics* **11** 799–805.

FOOTE, S., VOLLRATH, D., HILTON, A. and PAGE, D. C. (1992). The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258** 60–66.

GNEDENKO, B.V. (1968). *Theory of Probability.* Chelsea Publishing Company.

HALL, P. (1988). *Introduction to the Theory of Coverage Processes.* Wiley, New York.

HUDSON, T., STEIN, L., GERETY, S., MA, J., CASTLE, A., SILVA, J., SLONIM, D., BAPTISTA, R., KRUGLYAK, L., XU, S., HU, X., COLBERT, A., ROSENBERG, C., REEVE-DALY, M., ROZEN, S., HUI, L., WU, X., VESTERGAARD, C., WILSON, K., BAE, J., MAITRA, S., GANIATSAS, S., EVANS, C., DEANGELIS, M., INGALLS, K., NAHF, R., HORTON, L., OSKIN ANDERSON, M., COLLYMORE, A., YE, W., KOUYOUMJIAN, V., ZEMSTEVA, I., TAM, J., DEVINE, R., COURTNEY, D., RENAUD, M., NGUYEN, H., O'CONNOR, T., FIZAMES, C., FAURÉ, S., GYAPAY, G., DIB, C., MORISSETTE, J., ORLIN, J., BIRREN, B., GOODMAN, N., WEISSENBACH, J., HAWKINS, T., FOOTE, S., PAGE, D. and LANDER E. S. (1995). An STS-Based Map of the Human Genome. *Science* **270** 1945–1954.

KINGMAN, J. F. C. (1993). *Poisson processes.* Oxford-University Press.

KARLIN, S. and MACKEN, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Am. Stat. Assoc.* **86** 27–35.

KOHARA, Y., AKIYAMA, K. and ISONO, K. (1987). The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50** 495–508.

LANDER, E. S. and WATERMAN, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2** 231–239.

LEE, W. (1992). "A mathematical analysis of genome physical mapping", Masters Thesis, Department of Mathematics, University of Southern California.

LENGLART, E. (1977). Relation de dominance entre deux processus. *Ann. Inst. H. Poincaré* **13** 171–179.

MARR, T. G., YAN, X. and YU, Q. (1992). Genomic mapping by single copy landmark detection: A predictive model with a discrete mathematical approach. *Mamm. Genome* **3** 644–649.

NELSON, D. O. and SPEED, T. P. (1994). Predicting progress in directed mapping projects. *Genomics* **24** 41–52.

OLSON, M. V., DUTCHIK, J. E., GRAHAM, M. Y., BRODEUR, G. M., HELMS, C., FRANK, M., MACCOLLIN, M., SCHEINMAN, R. and FRANK, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci.* **83** 7826–7830.

PALAZZOLO, M. J., SAWYER, S. A., MARTIN, C. H., SMOLLER, D. A. and HARTL, D. L. (1991). Optimized strategies for sequence-tagged-site selection in genome mapping. *Proc. Natl. Acad. Sci.* **88** 8034–8038.

PORT, E., SUN, F., MARTIN, D. and WATERMAN, M. S. (1995). Genomic mapping by end-characterized random clones: A mathematical analysis. *Genomics* **26** 84–100.

TORNEY, D. C. (1991). Mapping using unique sequences. *J. Mol. Biol.* **217** 259–264.

ZHANG, M. Q. and MARR, T. G. (1993). Genome mapping by nonrandom anchoring: A discrete theoretical analysis. *Proc. Natl. Acad. Sci.* **90** 600–604.