The Annals of Applied Probability
2005, Vol. 15, No. 4, 2553–2574
DOI: 10.1214/105051605000000584

INVARIANCE PRINCIPLE FOR THE COVERAGE RATE OF GENOMIC PHYSICAL MAPPINGS

BY DIDIER PIAU

Université Lyon 1

We study some stochastic models of physical mapping of genomic sequences. Our starting point is a global construction of the process of the clones and of the process of the anchors which are used to map the sequence. This yields explicit formulas for the moments of the proportion occupied by the anchored clones, even in inhomogeneous models. This also allows to compare, in this respect, inhomogeneous models to homogeneous ones. Finally, for homogeneous models, we provide nonasymptotic bounds of the variance and we prove functional invariance results.

0. Introduction. The goal of the projects of genomic physical mapping is to reconstruct almost completely the sequence of a genome, starting from a multitude of exactly sequenced fragments, which are called clones. One approach to the reconstruction of the overall positions of these clones in the complete genomic sequence uses so-called anchors. These are short, exactly sequenced, portions of the genome which are assumed to appear only once in the full genomic sequence. An anchored clone is a clone which contains an anchor. In this paper we assume that the positions of the anchors, hence of the anchored clones, are exactly known. Maximal connected unions of anchored clones are called islands or, more exactly, anchored islands, aka contigs. The complement of the islands is called the ocean. When suitably rescaled, the full genomic sequence is identified with (a portion of) the real line, the anchors are identified with points, and the clones and the islands are identified with intervals.

The overall quality of the reconstitution of a given genomic sequence depends obviously on the number of islands, on their length and on the proportion of the sequence which is occupied by the ocean, among other characteristics of the project. One hopes that the islands are as few and as long as possible, and that the proportion occupied by the ocean is as low as possible. Arratia, Lander, Tavaré and Waterman [1] introduced a stochastic model of physical mapping, where the positions of the right ends of the clones and the positions of the anchors are distributed according to independent homogeneous Poisson processes on the real line, and where the lengths of the clones are random, i.i.d. and independent of everything

Received June 2004; revised April 2005.

AMS 2000 subject classifications. Primary 60G55; secondary 92D20, 60F17.

Key words and phrases. DNA sequences, physical mapping, anchored islands, coverage processes, FKG inequality, genomic sequences, inhomogeneous Poisson processes, invariance principle.

else. For this model, Arratia et al. computed the mean values of the three quantities of interest that we mentioned above. For related studies, see [3–5].

Motivated by the fact that actual genomic sequences do not fulfill the homogeneity hypotheses which underlie the stochastic model introduced by Arratia, Lander, Tavaré and Waterman [1], Schbath [7] and Schbath, Bossard and Tavaré [8] extended this setting in two directions. In both papers the independence properties of the model remain, but Schbath [7] studied the case when the intensities of the Poisson processes which generate the positions of the clones and the positions of the anchors may depend on their respective positions along the genome, and Schbath, Bossard and Tavaré [8] studied the case when the distributions of the lengths of the clones may depend on their respective positions along the genome. In these two wider contexts, these papers provide expressions of the mean value of the number of islands, of the mean value of the proportion occupied by the ocean and, under an additional technical hypothesis, of the mean value of the length of the islands.

In the present paper we pursue the study of this class of models. As a first contribution, we consider the class of models where the Poisson process of the clones, the Poisson process of the anchors and the distributions of the lengths of the clones can all be inhomogeneous simultaneously. To give a flavor of our results in this direction, we state Proposition 1 below, which extends formulas of the papers mentioned above, for the mean value of the number of clones and for the mean value of the number of anchored clones which cover a point.

To state Proposition 1, we introduce the measure $c(dx)$ on the real line as the intensity measure of the Poisson process $\mathcal{C}$ of the (right ends of the) clones, the measure $a(dx)$ on the real line as the intensity measure of the Poisson process $\mathcal{A}$ of the anchors and, for every x on the real line, the random variable L_x as the length of a clone whose right end is at position x, and we refer to Section 1 for more precise definitions of these objects.

For every x on the real line, $n_{\mathcal{C}}(x)$ denotes the number of clones which contain the point x, and $n_{\mathcal{A}}(x)$ denotes the number of anchored clones which contain the point x.

PROPOSITION 1 (General case). (i) *The random variable $n_{\mathcal{C}}(x)$ follows the Poisson distribution whose mean value is given by the expression*

$$\mathbb{E}(n_{\mathcal{C}}(x)) = \int_x^{+\infty} c(dz)\mathbb{P}(L_z \geq z - x).$$

(ii) *The mean value of $n_{\mathcal{A}}(x)$ is given by the expression*

$$\mathbb{E}(n_{\mathcal{A}}(x)) = \int_x^{+\infty} c(dz) \int_{z-x}^{+\infty} \mathbb{P}(L_z \in dt)\big(1 - e^{-a([z-t,z])}\big).$$

Here $a([z-t,z])$ denotes the measure of the interval $[z-t,z]$ with respect to the measure $a(dx)$.

(iii) *The distribution of the random variable $n_{\mathcal{A}}(x)$ is not Poisson. More specifically, either $n_{\mathcal{A}}(x) = 0$ almost surely, or the variance of $n_{\mathcal{A}}(x)$ is strictly greater than its mean value.*

In actual physical mapping projects, the condition that $n_{\mathcal{A}}(x) = 0$ almost surely is never fulfilled. On the mathematical side, this would correspond to degeneracies such as the fact that $L_z \le z - x$ almost surely for every $z \ge x$ and/or the fact that the intensity of the anchors is zero on a suitable neighborhood of x.

The homogeneous case is when $c(dx) = \kappa\, dx$ and $a(dx) = \alpha\, dx$ for two given positive constants κ and α, and when every L_x is distributed like a given random variable L. The specialization of Proposition 1 to the homogeneous case is as follows.

COROLLARY 1 (Homogeneous case). *In the homogeneous case with parameters κ, α and L, the mean values of $n_{\mathcal{C}}(x)$ and $n_{\mathcal{A}}(x)$ do not depend on the point x and are given by the expressions*

$$\mathbb{E}(n_{\mathcal{C}}) = \kappa \mathbb{E}(L), \qquad \mathbb{E}(n_{\mathcal{A}}) = \kappa \mathbb{E}\big(L(1 - e^{-\alpha L})\big).$$

More importantly than the slight generalizations above, our second contribution is to provide explicit formulas for the higher moments of these quantities in the general model with variable intensities. In the homogeneous case, our results imply, for instance, that the proportion of a large genomic sequence occupied by the ocean is asymptotically Gaussian; see Theorem 1 below.

THEOREM 1 (Homogeneous case). *Consider the homogeneous case with parameters κ, α and L, and assume that L is square integrable. For any positive G, let the random variable $\mathcal{O}_G$ denote the measure of the intersection of the ocean with any interval of length G, for instance the interval $[0, G]$, and let $\sigma^2(\mathcal{O}_G)$ denote the variance of $\mathcal{O}_G$.*

(i) *There exists a positive constant $\varrho < 1$ such that $\mathbb{E}(\mathcal{O}_G) = \varrho G$ for every nonnegative G.*

(ii) *There exist finite positive constants ν and λ such that, for every nonnegative G,*

$$\nu G - \lambda \le \sigma^2(\mathcal{O}_G) \le \nu G.$$

Hence $\sigma^2(\mathcal{O}_G) \sim \nu G$ when $G \to \infty$. Furthermore, the function $G \mapsto \sigma^2(\mathcal{O}_G)$ is convex, and $\sigma^2(\mathcal{O}_G) - \nu G + \lambda \to 0$ when $G \to \infty$.

(iii) *For every positive G, let Θ_G denote the random process, indexed by the real numbers $0 \le t \le 1$, and defined by*

$$\Theta_G(t) := (\mathcal{O}_{Gt} - \varrho G t)/\sqrt{\nu G}.$$

When $G \to \infty$, the process Θ_G converges in distribution to a standard Wiener process on the space of continuous functions on $[0, 1]$, equipped with the metric of the uniform convergence.

(iv) *The constants ϱ, ν and λ above can be written explicitly as integrals which involve the parameters κ, α, and the distribution of L.*

The starting point of our results is a global construction of the clones, the anchors and the islands, using a single Poisson process. We expose this global construction in Section 1. We provide alternative descriptions of this process, locating for instance the clones by their left ends instead of their right ends. A natural conjecture in this setting is that the homogeneous model would be the only one invariant by the symmetry of the real line, but we disprove this. In Section 2 we rewrite in our general setting various formulas due to Arratia, Lander, Tavaré and Waterman [1] or to Schbath [7] or to Schbath, Bossard and Tavaré [8]. Section 3 provides explicit formulas for every moment of the proportion of the real line which is occupied by the ocean in the general case and provides rather sharp bounds of the variance in the homogeneous case. Finally, Section 4 proves the invariance result stated in Theorem 1 above, in the homogeneous case. On our way, we provide asymptotics of the variance when the number of clones is vanishingly small and we build comparison tools that yield effective upper and lower bounds in some inhomogeneous cases.

1. Global model. In this section we build the clones, the anchors and the islands from a single Poisson process. Sections 1.2 and 1.3 are not used in the rest of the paper and may be omitted on a first reading.

1.1. *Clones.* Let $\mathbb{R}$ denote the real line and $\mathbb{R}^+ := [0, +\infty)$ the nonnegative half line. Let $c(dx)$ denote the intensity measure of the Poisson process of the right ends of the clones. Assume that, when the right end of a clone is located at x, its length follows the distribution of a given random variable L_x. We represent the clone which covers exactly the interval $[x - t, x]$ of length $t \geq 0$ by the point (x, t) in $\mathbb{R} \times \mathbb{R}^+$. The distribution of the clones is described by a Poisson process $\mathcal{C}$ on $\mathbb{R} \times \mathbb{R}^+$ of intensity measure m, with

$$m(dx\,dt) := c(dx)\mathbb{P}(L_x \in dt).$$

In other words, $\mathcal{C}$ is a random subset of $\mathbb{R} \times \mathbb{R}^+$, which is almost surely locally finite, and such that the following holds. For all Borel subsets D and D' of $\mathbb{R} \times \mathbb{R}^+$ such that $D \cap D'$ is empty, the random number of points of $\mathcal{C}$ in D and the random number of points of $\mathcal{C}$ in D' are independent. Furthermore, for every Borel subset D of $\mathbb{R} \times \mathbb{R}^+$, the number of points of $\mathcal{C}$ in D is a Poisson random variable of mean value $m(D)$.

In fact, the intensity measure m can be any Borel measure on $\mathbb{R} \times \mathbb{R}^+$ with a locally finite first marginal c, given by

$$c(dx) := \int_{t \geq 0} m(dx\,dt).$$

That is, one assumes that $c([-G, G]) = m([-G, G] \times \mathbb{R}^+)$ is finite for every finite positive G. The assumption that c is locally finite ensures that the distribution of L_x is well defined, and given by the Radon–Nikodym derivative

$$\mathbb{P}(L_x \in dt) := m(dx\,dt)/c(dx).$$

1.2. *Alternative descriptions of the clones.* At first sight, it may seem rather arbitrary to locate the position of a clone by its right endpoint, rather than by its midpoint or by its left endpoint. In fact, these alternative descriptions are also characterized by Poisson processes, albeit possibly with different intensities. For instance, using the couple (y, t) to describe the clone $[y, y + t]$ yields a Poisson process on $\mathbb{R} \times \mathbb{R}^+$ of intensity measure m', with

$$m'(dy\,dt) =: c'(dy)\mathbb{P}(L'_y \in dt).$$

One obtains m' from m, or rather, one obtains $c'(dy)$ and the distributions of the random variables L'_y from $c(dx)$ and from the distributions of the random variables L_x, as follows. For any nonnegative test function Φ, the expected value of the sum over every clone $[y, x]$ of $\Phi(y, x)$ reads

$$\begin{aligned}\mathbb{E}\Biggl(\sum_{[y,x]\text{ clone}} \Phi(y, x)\Biggr) &= \iint m(dx\,dt)\Phi(x - t, x)\\ &= \iint m'(dy\,dt)\Phi(y, y + t).\end{aligned}$$

In other words, one asks that

$$\int c(dx)\mathbb{E}\bigl(\Phi(x - L_x, x)\bigr) = \int c'(dy)\mathbb{E}\bigl(\Phi(y, y + L'_y)\bigr).$$

Since this equality holds for every test function Φ, this implies that $c'(dy)$ and the distributions of the random variables L'_y are given by

$$c'(dy) = \int_y^{+\infty} c(dx)\mathbb{P}(L_x \in x - dy),$$

$$\mathbb{P}(L'_y \in dt) = \mathbb{P}(L_{y+t} \in dt)c(t + dy)/c'(dy).$$

Similar formulas give the intensity measures associated to the description of a clone by its midpoint and by its length, or by its two endpoints.

1.3. *On the (non)specificity of the homogeneous clones.* Based upon the preceding section, the reader might be led to believe that the homogeneous model is privileged with respect to the transformations of the intensity measure $m(dx\,dt)$ into the intensity measure $m'(dy\,dt)$ and of $m'(dy\,dt)$ into $m(dx\,dt)$. To wit, if the intensity $c(dx)$ and the distributions of the random variables L_x are invariant by the translations of the real line, so are the intensity $c'(dy)$ and the distributions of the random variables L'_y. Thus, in the homogeneous case, $c(dx) = c'(dx) = \kappa\, dx$ and the distributions of every L_x and every L'_y do coincide.

Our goal in this section is to point out that there are other cases where the two intensity measures m and m' coincide. To build such examples, we need to introduce, for every x on the real line, the unit interval U_x which is centered at x, that is,

$$U_x := [x - 1/2, x + 1/2).$$

Let B_0 denote the union of the intervals U_{2k} for every integer k, and let B_1 denote its complement. Let $u_0(dx)$ denote a finite measure on U_0, and $u_1(dx)$ a finite measure on U_1. Let $c(dx)$ denote the unique measure on $\mathbb{R}$ which is invariant by the translation $x \mapsto x + 2$ and whose restrictions to U_0 and to U_1 are $u_0(dx)$ and $u_1(dx)$, respectively. Thus, $c = c_0 + c_1$ with, for $i = 0$ and for $i = 1$,

$$c_i(dx) := \sum_k u_i(2k + dx),$$

where both sums run over every integer k. In other words, $c(dx)$ can be any locally finite measure on $\mathbb{R}$, invariant by the translation $x \mapsto x + 2$, and the measure $c_0(dx)$, respectively the measure $c_1(dx)$, denotes the restriction of $c(dx)$ to B_0, respectively the restriction of $c(dx)$ to B_1.

Assume finally that $L_x = 2$ with full probability when x is in B_0, and that $L_x = 4$ with full probability when x is in B_1. Since L_x is always an even integer, the endpoints of a given clone are either both in B_0 or both in B_1. Using this remark, one can check that $m = m'$. Besides, the process which locates the clones by their midpoint is given by a similar intensity measure, choosing with full probability the length 4 when the midpoint belongs to B_0, and choosing with full probability the length 2 when the midpoint belongs to B_1.

In the example above, the distributions of the lengths are discrete, hence the measure $m(dx\, dt)$ is singular with respect to the Lebesgue measure. However, the same idea can be adapted to produce examples where $m(dx\, dt)$ is absolutely continuous. To see this, introduce the Poisson process which describes a clone $[y, x]$ by its endpoints (y, x), and assume that the intensity measure m^* of this Poisson process is

$$m^*(dy\, dx) := dy\, dx \sum_k \mathbf{1}\{(y, x) \in U_{2k} \times U_{2k+2}\} + \mathbf{1}\{(y, x) \in U_{2k-1} \times U_{2k+3}\},$$

where the sum runs over every integer k. In words, the left endpoints and the right endpoints of the clones both have homogeneous intensity measures, and both endpoints of a clone belong to B_0 or both endpoints belong to B_1. Furthermore, given that the left endpoint y belongs to B_0, the right endpoint x is uniformly distributed over the next unit interval of B_0 to the right of y, that is, over the connected component of B_0 which contains $y + 2$. Given that the left endpoint y belongs to B_1, the right endpoint x is uniformly distributed over the second next unit interval of B_1 to the right of y, that is, over the connected component of B_1 which contains $y + 4$.

In this new example, the measure $m(dx\,dt)$ is as follows. The intensity $c(dx)$ is the Lebesgue measure. The length L_x is uniformly distributed over U_{x-2k} when x is in U_{2k+2}, and L_x is uniformly distributed over U_{x-2k+1} when x is in U_{2k+3}. The support of the distribution of L_x is a unit subinterval of the interval $[1, 3]$ when x is in B_0, and it is a unit subinterval of the interval $[3, 5]$ when x is in B_1, hence the distribution of L_x cannot be the same for every x on the real line. Finally, $m = m'$ because m^* is invariant by the symmetries of the real line, since these exchange the left endpoints and the right endpoints of the clones while leaving their lengths unchanged.

1.4. *Anchors.* In this section and in the rest of the paper, we come back to the (x, t) Poisson process of intensity m, which represents the clones by their right endpoint and by their length.

The anchors are described by a Poisson process $\mathcal{A}$ on the real line, with intensity $a(dx)$, independent of the Poisson process $\mathcal{C}$ of the clones which we defined in Section 1.1. Thus, for every Borel subset D of the real line, the number of anchors in D is a random variable whose distribution is Poisson with mean value $a(D)$, and the number of anchors in the Borel sets D and D' are independent random variables as soon as $D \cap D'$ is empty.

For every subset D of the real line, let $I(D)$ denote the cone of influence of D in $\mathbb{R} \times \mathbb{R}^+$. This is the set of clones (x, t) which become anchored clones when every point of D becomes an anchor. Thus,

$$I(D) := \{(x, t) \in \mathbb{R} \times \mathbb{R}^+; [x - t, x] \cap D \neq \varnothing\}.$$

For every measurable D, the process $\mathcal{C}_D := \mathcal{C} \cap I(D)$ of the clones that are anchored by D is deduced from $\mathcal{C}$ by erasing some clones, hence each $\mathcal{C}_D$ is indeed a Poisson process whose intensity measure m_D on $\mathbb{R} \times \mathbb{R}^+$ is the restriction of the original intensity measure m to the set $I(D)$, that is,

$$m_D(dx\,dt) := \mathbf{1}\{(x, t) \in I(D)\} m(dx\,dt).$$

For every locally finite subset D of the real line, let $\mathbb{P}^D$ denote the conditioning of $\mathbb{P}$ by the event $\{\mathcal{C} = D\}$. Finally, let $\mathcal{C}_{\mathcal{A}}$ denote the process of the anchored clones, that is,

$$\mathcal{C}_{\mathcal{A}} := \{(x, t) \in \mathcal{C}; [x - t, x] \cap \mathcal{A} \neq \varnothing\} = \mathcal{C} \cap I(\mathcal{A}).$$

1.5. *Clones + anchors.* One can, and we shall, simultaneously generate the processes $\mathcal{C}$, $\mathcal{A}$ and $\mathcal{C}_{\mathcal{A}}$ from a unique Poisson process, as follows. Let $M := \mathbb{R}^+ \cup \{*\}$, where $*$ denotes any point which is not in $\mathbb{R}^+$. We endow the set M with the smallest σ-algebra which contains the Borel sets of $\mathbb{R}^+$ and the singleton $\{*\}$. We endow the set $\mathbb{R} \times M$ with the product σ-algebra of the Borel σ-algebra of $\mathbb{R}$

and of this σ-algebra of M. Finally, we introduce a Poisson process on $\mathbb{R} \times M$ with intensity

$$g(dx\,dt) := m(dx\,dt) + a(dx)\delta_*(dt).$$

We call this Poisson process the global process. The point (x, t) with t in $\mathbb{R}^+$ represents the clone $[x - t, x]$ and the point $(x, *)$ represents the anchor at x. The restriction of the global process to the domain $\mathbb{R} \times \mathbb{R}^+$ yields the process of the clones described in Section 1.1, since its intensity, which is the restriction of g to $\mathbb{R} \times \mathbb{R}^+$, is $m(dx\,dt)$. Likewise, the projection $(x, *) \mapsto x$ on the real coordinate of the restriction of the global process to the domain $\mathbb{R} \times \{*\}$ yields the process of the anchors described in Section 1.4, since its intensity is $a(dx)$. Finally, the process of the clones and the process of the anchors are indeed independent since they are realized as the restrictions of the global Poisson process to the domains $\mathbb{R} \times \mathbb{R}^+$ and $\mathbb{R} \times \{*\}$, which are disjoint subsets of $\mathbb{R} \times M$.

Proposition 2 below and Proposition 1 and Corollary 1 in our Introduction follow from the construction above. The proofs are simple adaptations of the proofs given by Arratia, Lander, Tavaré and Waterman [1], Schbath [7] and Schbath, Bossard and Tavaré [8], hence we omit them.

PROPOSITION 2 (General case). *With respect to $\mathbb{P}$, $\mathcal{A}$ and $\mathcal{C}$ are independent Poisson processes. For every locally finite D, with respect to $\mathbb{P}^D$, $\mathcal{C}_D$ is a Poisson process. With respect to $\mathbb{P}$, $\mathcal{C}_{\mathcal{A}}$ is not a Poisson process.*

1.6. *Ocean.* Recall that the ocean $\mathcal{O}$ is the complement of the union of the anchored islands. For every Borel set D of the real line, let $\mathcal{O}(D)$ denote the measure of $\mathcal{O} \cap D$. For every positive real number G, let $\mathcal{O}_G := \mathcal{O}([0, G])$. For every Borel set D of the real line, let

$$r(D) := \mathbb{P}(D \subset \mathcal{O}).$$

For every $n \geq 1$ and all real numbers $z_1, \dots, z_n$, let

$$r(z_1, \dots, z_n) := r(\{z_1, \dots, z_n\}) = \mathbb{P}(z_1 \in \mathcal{O}, \dots, z_n \in \mathcal{O}).$$

For instance, $r(z)$ is the probability that z belongs to no anchored clone. Hence $r(z)$ may depend on z but $r(z)$ corresponds to $r(0)$ if the process of the clones and the process of the anchors are both shifted by z. Lemma 1 below stems from the definitions.

LEMMA 1. *For every Borel set D of the real line and every integer $n \geq 1$,*

$$\mathbb{E}(\mathcal{O}(D)^n) = \int_{D^n} r(z_1, \dots, z_n)\, dz_1 \cdots dz_n.$$

For instance,

$$\mathbb{E}(\mathcal{O}_G) = \int_0^G r(z)\, dz, \qquad \mathbb{E}(\mathcal{O}_G^2) = \int_0^G \int_0^G r(z, z')\, dz\, dz'.$$

2. First moments. This section is mainly a rephrasing of results of Arratia, Lander, Tavaré and Waterman [1] and Schbath, Bossard and Tavaré [8]. Our only contribution here is to include both inhomogeneities simultaneously in the results, namely, the inhomogeneities of the lengths of the clones on the one hand, and the inhomogeneities of the positions of the right ends of the clones and of the anchors on the other hand. We are interested in $r(z)$, which describes locally the mean value of the proportion of the real line which is occupied by the ocean.

LEMMA 2. *Let $J(x, y)$ denote the probability of the event that two points x and y such that $x \le y$ belong to no common clone. Then*

$$J(x, y) := \exp\left(-\int_y^{+\infty} \mathbb{P}(L_t \ge t - x)c(dt)\right).$$

Caution: we renamed $J(x, x+t)$ the expression $J(x, t)$ of the papers mentioned above.

LEMMA 3. *For every z, the joint law of the positions x and y of the anchors which are the closest of z to the left and to the right, respectively, is $A^-(z, dx)A^+(z, dy)$, where*

$$A^-(z, dx) := A(x, z)a(dx) \quad \text{and} \quad A^+(z, dy) := A(z, y)a(dy).$$

For all points $x \le y$, we use the notation

$$A(x, y) := \exp\left(-\int_x^y a(dt)\right).$$

THEOREM 2 (Schbath, Bossard and Tavaré [8]). *For every z,*

$$r(z) = \int_{x \le z \le y} \frac{J(x, z)J(z, y)}{J(x, y)} A(x, y)a(dx)a(dy).$$

The contribution of the intensity measure a in $r(z)$ corresponds to the product $A^-(z, dx)A^+(z, dy)$.

A quick look at the ratio of the functions J in the integral above could lead to the erroneous conclusion that $r(z)$ is not well defined when $J(x, y)$ is not always positive. [One knows that $J(x, y)$ is positive when, e.g., the random variables L_t are uniformly integrable, and $c(dt)$ is uniformly bounded, i.e., when there exists a finite κ_+ such that $c(dx) \le \kappa_+ \, dx$.] In fact, one can show that this ratio is at most 1 for any intensity $c(dt)$ and any distributions of the random variables L_t, hence the formula for $r(z)$ in Theorem 2 is always valid.

We recall that, in the homogeneous case, the process of the clones has constant intensity $c(dx) = \kappa \, dx$, the lengths of the clones are i.i.d. and distributed like a random variable L, and the process of the anchors has constant intensity $a(dx) = \alpha \, dx$.

COROLLARY 2 (Arratia, Lander, Tavaré and Waterman [1]). *In the homogeneous case with parameters κ, α and L, $r(z) = \varrho$ does not depend on z and its value is*

$$\varrho := \int_0^{+\infty}\int_0^{+\infty} \alpha^2 e^{-\alpha(u+v)} \frac{J(u)J(v)}{J(u+v)}\,du\,dv.$$

Here, $J(u)$ is the probability that an interval of length u is not covered by any unique clone, hence

$$J(u) := \exp\biggl(-\kappa \int_u^{+\infty} \mathbb{P}(L \geq t)\,dt\biggr).$$

When, furthermore, $L = \ell$ with full probability for a given positive real number ℓ, Arratia et al. deduce from this the value of ϱ as a function of ℓ, κ and α.

One gets the expression of ϱ in Corollary 2 from $r(z)$ in Theorem 2, using the change of variables $u = z - x$, $v = y - z$.

3. Higher moments. Higher moments of the quantities introduced above involve functionals of the processes that depend on more than one point. We first describe the computation of the variance of the proportion of the real line which is occupied by the ocean in the general case, then we consider the higher moments in the general case, and finally we prove precise asymptotics of the variance in the homogeneous case.

3.1. *Variance of the ocean proportion.* Recall that $r(z, z')$ is the probability that neither z nor z' is covered by anchored clones. Let $r_0(z, z')$, respectively $r_1(z, z')$, respectively $r_2(z, z')$, denote the probability of the same event, when the number of anchors between z and z' is 0, respectively 1, respectively 2 or more. One can decompose each of these events, according to the position of the first anchor to the left of the interval (z, z'), which we call x in the integrals below, to the position of the first anchor to the right of (z, z'), which we call y in the integrals below, and to the positions of the leftmost and rightmost anchors, if any, in the interval (z, z'), which we call s and t in the integrals below.

Thus $r(z, z') = r_0(z, z') + r_1(z, z') + r_2(z, z')$ with, for $z \leq z'$,

$$r_0(z, z') := \int_{x \leq z \leq z' \leq y} J(x \mid z, z' \mid y) B(dx, dy),$$

$$r_1(z, z') := \int_{x \leq z \leq s \leq z' \leq y} J(x \mid z \mid s) J(s \mid z' \mid y) a(ds) B(dx, dy),$$

$$r_2(z, z') := \int_{x \leq z \leq s \leq t \leq z' \leq y} J(x \mid z \mid s) J(t \mid z' \mid y) B(dx, ds) B(dt, dy).$$

We mention that $r_i(z, z')$ is defined as an integral of dimension $i + 2$, for $i = 0$, 1 or 2. We used the following notation. The two-dimensional measure B is defined on the subset $x \le y$ of $\mathbb{R} \times \mathbb{R}$ by the formula

$$B(dx, dy) := A(x, y)a(dx)a(dy).$$

For any $x \le z \le z' \le y$,

$$J(x \mid z, z' \mid y) := \frac{J(x, z)J(z', y)}{J(x, y)},$$

$$J(x \mid z \mid y) := \frac{J(x, z)J(z, y)}{J(x, y)}.$$

The quantities involved in the definitions above have the following interpretations. First, $\mathbf{1}\{x \le z \le y\}B(dx, dy)$ is the distribution of the couple formed by the positions of the rightmost anchor to the left of z and of the leftmost anchor to the right of z. Second, $J(x \mid z \mid y)$ is the probability that z is not covered by an anchored clone when the closest anchor to the left of z is at x and the closest anchor to the right of z is at y. Finally, $J(x \mid z, z' \mid y)$ is the probability that z and z' are not covered by anchored clones when the closest anchor to the left of z is at x, the closest anchor to the right of z' is at y, and when there is no anchor between z and z'. Schbath's formula in our Theorem 2 reads

$$r(z) = \int_{x \le z \le y} J(x \mid z \mid y)B(dx, dy).$$

If one forgets the condition that $s \le t$ in the definition of $r_2(z, z')$, one gets the product of the integrals over (x, s) and over (t, y), which are $r(z)$ and $r(z')$, respectively. This implies our Lemma 4 below.

LEMMA 4. *For any $z \le z'$, $r_2(z, z') = r(z)r(z') - r_3(z, z')$ where the term $r_3(z, z')$ is nonnegative and is*

$$r_3(z, z') := \int_{x \le z \le s, t \le z' \le y, s \ge t} J(x \mid z \mid s)J(t \mid z' \mid y)B(dx, ds)B(dt, dy).$$

As a consequence, the variance $\sigma^2(\mathcal{O}_G)$ of $\mathcal{O}_G$ is

$$\sigma^2(\mathcal{O}_G) = \int_0^G \int_0^G (r_0 + r_1 - r_3)(z, z')\, dz\, dz'.$$

3.2. *Higher moments of the ocean proportion.* As mentioned above, one can adapt the technique used in the last section to study the mean value of any power of $\mathcal{O}_G$. For instance,

$$\mathbb{E}(\mathcal{O}_G^3) = \int_0^G \int_0^G \int_0^G r(z, z', z'')\, dz\, dz'\, dz''.$$

Thus, assuming for instance that $n = 3$, one has to compute the n-point function $r(z, z', z'')$. First, one can assume by symmetry that $z \le z' \le z''$. Let x denote the position of the rightmost anchor to the left of z, and y the position of the leftmost anchor to the right of z''. Let s and t denote the positions of the leftmost and rightmost anchors in the interval (z, z'), and s' and t' the positions of the leftmost and rightmost anchors in the interval (z', z''), if these exist.

Then $r(z, z', z'')$ is $n! = 6$ times the sum of $3^{n-1} = 9$ terms $r_{i,i'}(z, z', z'')$. Each term $r_{i,i'}(z, z', z'')$ corresponds to the number $i = 0, 1$ or 2 of anchors to be considered in the interval (z, z') and to the number $i' = 0, 1$ or 2 of anchors to be considered in the interval (z', z''), namely, no anchor at all, or a unique anchor, denoted by s or by s', or two extremal anchors, denoted by s and t, or by s' and t'.

To take an example, consider the case $i = 2$ and $i' = 1$. This yields $r_{2,1}(z, z', z'')$ as the integral

$$\int_{D_{2,1}} J(x \mid z \mid s) J(t \mid z' \mid s') J(s' \mid z'' \mid y) a(ds') B(dx, ds) B(dt, dy),$$

where the domain of integration $D_{2,1}$ has dimension 5 and is defined by the inequalities

$$x \le z \le s \le t \le z' \le s' \le z'' \le y.$$

Likewise, if $i = 0$ and $i' = 2$, $r_{0,2}(z, z', z'')$ is the integral

$$\int_{D_{0,2}} J(x \mid z, z' \mid s') J(t' \mid z'' \mid y) B(dx, ds') B(dt', dy),$$

where the domain of integration $D_{0,2}$ has dimension 4 and is defined by the inequalities

$$x \le z, \qquad z' \le s' \le t' \le z'' \le y.$$

More generally, $\mathbb{E}(\mathcal{O}_G^n)$ is the integral of the n-point function $r(z_1, \ldots, z_n)$ on the domain $[0, G]^n$ with respect to the Lebesgue measure. For every n-tuple $z_1 \le \cdots \le z_n$, $r(z_1, \ldots, z_n)$ can be decomposed as a sum of 3^{n-1} contributions. Each of these contributions corresponds to the event that each interval $[z_k, z_{k+1}]$ contains no anchor at all, or a unique anchor, or at least two anchors.

3.3. *Variance in the homogeneous case.* In this section we study the homogeneous case, when the intensity measures are $a(dt) = \alpha\, dt$ and $c(dx) = \kappa\, dx$, and the distribution of the length L_x of a clone does not depend on its position x and is the distribution of a random variable L. We recall that the distribution of the global process is left invariant by the action of the translations. This implies that $r(z) = \varrho$ for every z, where the value of ϱ is given in Corollary 2. Hence,

$$\mathbb{E}(\mathcal{O}_G) = G\varrho.$$

Since $(z, z') \mapsto r(z, z') - r(z)r(z')$ is a symmetric function, $\sigma^2(\mathcal{O}_G)$ is twice an integral over $z' \ge z$. Likewise, the invariance by the translations implies that

$r(z, z') = r(0, z' - z)$ for every z and z'. Introducing $\bar{r}_i(z) := r_i(0, z)$, one is left with twice some integrals of the functions $\bar{r}_i(z)$ over z in $[0, G]$, namely,

$$\sigma^2(\mathcal{O}_G) = 2\int_0^G (G - z)\big(\bar{r}_0(z) + \bar{r}_1(z) - \bar{r}_3(z)\big)\, dz.$$

The values of the quantities $\bar{r}_i(z)$ for every nonnegative z are

$$\bar{r}_0(z) = \int_{x,y\geq 0} \alpha^2 e^{-\alpha(x+y+z)} \frac{J(x)J(y)}{J(x+y+z)}\, dx\, dy,$$
$$\bar{r}_1(z) = \int_{x,y\geq 0, 0\leq t\leq z} \alpha^3 e^{-\alpha(x+y+z)} \frac{J(x)J(t)J(z-t)J(y)}{J(x+t)J(z-t+y)}\, dx\, dy\, dt,$$
$$\bar{r}_3(z) = \int_{x,y,s,t\geq 0, s+t\geq z} \alpha^4 e^{-\alpha(x+y+s+t)} \frac{J(x)J(t)J(s)J(y)}{J(x+t)J(s+y)}\, dx\, dy\, ds\, dt.$$

We mention that $\bar{r}_0(z)$, respectively $\bar{r}_1(z)$, respectively $\bar{r}_3(z)$, is defined as an integral of dimension 2, respectively 3, respectively 4.

Using the fact that the function $x \mapsto J(x)$ is nondecreasing, one can bound each $\bar{r}_i(z)$ as follows:

$$\bar{r}_0(z) \leq e^{-\alpha z} j(\alpha),$$
$$\bar{r}_1(z) \leq \alpha z e^{-\alpha z} j(\alpha)^2,$$
$$\bar{r}_3(z) \leq (1 + \alpha z)e^{-\alpha z} j(\alpha)^2,$$

with the notation

$$j(\alpha) := \int_0^{+\infty} \alpha e^{-\alpha x} J(x)\, dx.$$

To prove the upper bound of $\bar{r}_0(z)$, one uses the fact that $J(y) \leq J(x + y + z)$, and one performs the integration of the upper bound. Likewise, to prove the upper bound of $\bar{r}_1(z)$, one uses the facts that $J(t) \leq J(x+t)$ and $J(z-t) \leq J(z-t+y)$, and one performs the integration of the upper bound. Finally, to prove the upper bound of $\bar{r}_3(z)$, one uses the facts that $J(t) \leq J(x + t)$ and $J(s) \leq J(s + y)$, and one performs the integration of the upper bound. In this last case, this yields

$$\bar{r}_3(z) \leq j(\alpha)^2 \int_{s,t\geq 0, s+t\geq z} \alpha^2 e^{-\alpha(s+t)}\, ds\, dt,$$

and the last double integral is indeed $(1 + \alpha z)e^{-\alpha z}$.

Since $J(x) \leq 1$, $j(\alpha) \leq 1$. Furthermore, the limit of J at infinity is 1, hence $\bar{r}_0(z) \sim e^{-\alpha z} j(\alpha)^2$ at infinity. Let

$$\sigma_i^2(G) := \int_0^G 2(G - z)\bar{r}_i(z)\, dz.$$

From the bounds on the three functions $\bar{r}_i$ which are stated above, it is not difficult to prove that, when $G \to \infty$,

$$\sigma_i^2(G) = \nu_i G - \lambda_i + \tau_i(G),$$

where $\tau_i(G) = o(1)$ for $i = 0, 1$ and 3. More specifically, these bounds imply that the numbers ν_i and λ_i, defined as

$$\nu_i := \int_0^{+\infty} 2\bar{r}_i(z)\,dz, \qquad \lambda_i := \int_0^{+\infty} 2z\bar{r}_i(z)\,dz,$$

are indeed finite and positive, and simple computations show that

$$\tau_i(G) := \int_G^{+\infty} 2(z - G)\bar{r}_i(z)\,dz.$$

Introduce $\tau(G) := \tau_0(G) + \tau_1(G) - \tau_3(G)$. Since each $\tau_i(G)$ is nonnegative, $|\tau(G)|$ is at most the maximum of $\tau_0(G) + \tau_1(G)$ and $\tau_3(G)$. Since $j(\alpha) \le 1$, our bounds on the three functions $\bar{r}_i$ imply that

$$|\tau(G)| \le \int_G^{+\infty} 2(z - G)(1 + \alpha z)e^{-\alpha z}\,dz.$$

Performing the integration, one gets

$$|\tau(G)| \le 2\alpha^{-2}(3 + \alpha G)e^{-\alpha G}.$$

Finally, when $G \to \infty$, $\tau(G) = O(Ge^{-\alpha G})$.

Assume now that $L \le \ell$ almost surely, for a finite ℓ. This means that the intensity measure of the global Poisson process on $\mathbb{R} \times \mathbb{R}^+$ puts no mass on the set $\mathbb{R} \times (\ell, +\infty)$. Assume that z and z' are such that $|z - z'| > \ell$. Then $I(z) \cap I(z')$ contains only clones (x, t) such that both points z and z' belong to $[x - t, x]$, hence in particular, such that $t > \ell$. Since $I(z) \cap I(z')$ is a subset of $\mathbb{R} \times (\ell, +\infty)$, its intensity measure must be zero. Thus, the events $\{z \in \mathcal{O}\}$ and $\{z' \in \mathcal{O}\}$ are in fact measurable with respect to the truncated cones of influence $I(z) \cap (\mathbb{R} \times [0, \ell])$ and $I(z') \cap (\mathbb{R} \times [0, \ell])$, respectively. Since these two subsets of $\mathbb{R} \times \mathbb{R}^+$ are disjoint, $\{z \in \mathcal{O}\}$ and $\{z' \in \mathcal{O}\}$ are independent events.

Finally, if $L \le \ell$ almost surely, $r(z, z') = r(z)r(z')$ as soon as z and z' are such that $|z - z'| > \ell$, hence $\bar{r}_0(z) + \bar{r}_1(z) - \bar{r}_3(z) = 0$ for every $z > \ell$, and $\tau(G) = 0$ for every $G \ge \ell$.

Proposition 3 below summarizes the results of this section.

PROPOSITION 3. (i) *Let* $\nu := \nu_0 + \nu_1 - \nu_3$ *and* $\lambda := \lambda_0 + \lambda_1 - \lambda_3$. *Then, when* $G \to \infty$,

$$\sigma^2(\mathcal{O}_G) = \nu G - \lambda + o(1).$$

(ii) *Assume that* $L \le \ell$ *almost surely for a finite* ℓ. *Then, for every* $G \ge \ell$,

$$\sigma^2(\mathcal{O}_G) = \nu G - \lambda.$$

4. Functional invariance in the homogeneous case. Our main task in this section is to prove that ν is positive, that is, not zero. We do this, first, in the limit $\kappa \to 0$ of a vanishing number of clones, then in the general case. Our techniques also yield upper and lower bounds of the mean value and of the variance of $\mathcal{O}_G$ when the intensities are not constant. Finally, we prove the functional invariance result of Theorem 1.

4.1. *Variance for vanishing clones.*

PROPOSITION 4 (Homogeneous case). *Fix the distribution of L and the value of α. Then, if κ is small enough, ν is positive. More precisely, when $\kappa \to 0$,*

$$\nu = \alpha^{-2}\mathbb{E}(\varphi(\alpha L))\kappa + o(\kappa),$$

where the function $x \mapsto \varphi(x)$ is explicit, positive on $x > 0$, and given by the formula

$$\varphi(x) := x - 1 + e^{-x}(1 - x^2/2).$$

PROOF. If $\kappa = 0$, $j(\alpha) = 1$ and $\bar{r}_i(z) = r_i^*(z)$, with

$$\begin{aligned} r_0^*(z) &:= e^{-\alpha z}, \\ r_1^*(z) &:= \alpha z e^{-\alpha z}, \\ r_3^*(z) &:= (1 + \alpha z)e^{-\alpha z}, \end{aligned}$$

hence $r_0^* + r_1^* - r_3^*$ is identically zero. (Besides, when $\kappa = 0$, $\mathcal{O}_G$ is almost surely zero.) We now show that the first derivative of ν with respect to κ at $\kappa = 0^+$ is positive.

When $\kappa = o(1)$, $J(x) = 1 - \kappa H(x) + o(\kappa)$ with

$$H(x) := \int_x^{+\infty} \mathbb{P}(L \geq t)\,dt.$$

This implies that $\bar{r}_i(z) = r_i^*(z) + \kappa s_i(z) + o(\kappa)$, for some explicit functions $s_i(z)$. Introducing $w_i := \int_0^{+\infty} s_i(z)\,dz$ and $w := w_0 + w_1 - w_3$, one gets $\nu = \kappa w + o(\kappa)$. For instance,

$$w_0 = \int_{x,y,z \geq 0} \alpha^2 e^{-\alpha(x+y+z)}\{H(x + y + z) - H(x) - H(y)\}\,dx\,dy\,dz,$$

and similar expressions of w_1 and w_3 obtain. After some tedious but simple computations, one gets

$$w_0 = h_2 - 2h_0, \qquad w_1 = 2h_1 - 4h_0, \qquad w_3 = 2h_2 - 6h_0,$$

where, for every nonnegative integer n, the value of h_n is given by

$$h_n := \int_0^{+\infty} \frac{(\alpha x)^n}{n!} e^{-\alpha x} H(x)\,dx.$$

Summing up these three contributions yields $w = 2h_1 - h_2$. Converting everything back in terms of the distribution of L, one finally gets

$$w = \alpha^{-2}\mathbb{E}(\varphi(\alpha L)),$$

where φ is given in the statement of the proposition above. It happens that $\psi(x) := e^x \varphi(x)$ defines a function ψ such that $\psi(0) = 0$ and whose derivative $\psi'(x) = x(e^x - 1)$ is obviously positive for every positive x. Thus $\varphi(x)$ is positive for every positive x, and w is positive for every distribution of L, except in the degenerate case when $L = 0$ almost surely. This proves that ν is positive for small values of κ. □

Other limiting cases than the one in Proposition 4 are possible. Recall that $\mathbb{E}(\mathcal{O}_G) = \varrho G$ for every nonnegative G, and that $\sigma^2(\mathcal{O}_G) \sim \nu G$ when $G \to \infty$.

1. If $\mathbb{E}(L^3) = o(1)$, then $\nu \sim \frac{1}{3}\alpha\kappa\mathbb{E}(L^3)$.
2. If $\kappa = o(1)$, then $(1 - \varrho) \sim \kappa\mathbb{E}(Le^{-\alpha L})$.
3. If $\mathbb{E}(L) = o(1)$, then $(1 - \varrho) \sim \kappa\mathbb{E}(L)$.

One can note that this last result does not depend on the value of α.

4.2. *Positive dependence.* Proposition 5 below deals with possibly inhomogeneous processes.

PROPOSITION 5 (General case). *For all Borel sets Z and Z',*

$$\mathbb{P}(Z \cup Z' \subset \mathcal{O}) \geq \mathbb{P}(Z \subset \mathcal{O})\mathbb{P}(Z' \subset \mathcal{O}).$$

In particular, $r(z, z') \geq r(z)r(z')$ for every z and z'.

Corollary 3 is a direct consequence of this proposition and of the expression of $\sigma^2(\mathcal{O}_G)$ in Section 3.3.

COROLLARY 3 (Homogeneous case). *For all nonzero intensities κ and α and every nonzero L, the constants ν and λ are positive and the function $G \mapsto \tau(G)$ is nonnegative. In particular, for every G,*

$$\nu G - \lambda \leq \sigma^2(\mathcal{O}_G) \leq \nu G.$$

Hence $\sigma^2(\mathcal{O}_G) \sim \nu G$ when $G \to \infty$. Furthermore, the following properties hold. The function $G \mapsto \sigma^2(\mathcal{O}_G)$ is increasing and convex. When $G \to 0$, $\sigma^2(\mathcal{O}_G) \sim \varrho(1 - \varrho)G^2$. When $G \to \infty$, $\sigma^2(\mathcal{O}_G) = \nu G - \lambda + o(1)$.

PROOF. As regards ν, recall from Section 3.3 that, in the homogeneous case,

$$\nu = \int_0^{+\infty} 2\big(r(0, z) - \varrho^2\big)\, dz.$$

Since $0 < \varrho < 1$, $r(0,0) = r(0) = \varrho > \varrho^2$. Furthermore, one can deduce from Section 3 an expression of $r(0,z)$ from the formulas which give $r_i(z,z')$ for $i = 0$, 1 and 2. The integrals involved are continuous with respect to z and z' because the functions J involved in these integrals are, and because obvious domination properties hold. Finally, $r(0,z) > \varrho^2$ for every nonnegative z in a neighborhood of 0, and $r(0,z) \geq \varrho^2$ for every nonnegative z. This implies that $\nu > 0$.

The proofs that λ is positive and that $\tau(G)$ is nonnegative are similar.

The equivalent of $\sigma^2(\mathcal{O}_G)$ when $G \to 0$ stems from the fact that $r(0,z) \to \varrho$ when $z \to 0$ and from the exact formula

$$\sigma^2(\mathcal{O}_G) = \int_0^G 2(G-z)\big(r(0,z) - \varrho^2\big)\,dz.$$

Finally, this formula and the fact that $r(0,z) \geq \varrho^2$ also yield the fact that the function $G \mapsto \sigma^2(\mathcal{O}_G)$ is increasing and convex, since the derivative of this function is

$$\int_0^G 2\big(r(0,z) - \varrho^2\big)\,dz. \qquad \square$$

PROOF OF PROPOSITION 5. For any Borel set Z, $\{Z \subset \mathcal{O}\}$ is a nonincreasing event, with respect to the global Poisson process introduced in Section 1.5. To see this, note that, if one adds some anchors and/or some clones to a given configuration, the union $\mathbb{R} \setminus \mathcal{O}$ of the anchored islands does not decrease, hence the indicator function of the event $\{Z \subset \mathcal{O}\}$ does not increase. Thus, our proposition is a direct consequence of the Fortuin–Kasteleyn–Ginibre (FKG) inequality

$$\mathbb{P}(D \cap D') \geq \mathbb{P}(D)\mathbb{P}(D'),$$

applied to the nonincreasing events $D := \{Z \subset \mathcal{O}\}$ and $D' := \{Z' \subset \mathcal{O}\}$; see [6], for instance. $\square$

4.3. *Bounds in the general case.* In the inhomogeneous case, minimal assumptions on $c(dx)$ and $a(dx)$ yield upper and lower bounds on $\mathbb{E}(\mathcal{O}_G)$ and $\sigma^2(\mathcal{O}_G)$, as we now show. In this section we assume that the intensities of the processes of the clones and of the anchors are uniformly bounded. Hence, $a(dx)$ and $c(dx)$ are absolutely continuous with respect to the Lebesgue measure and there exist finite positive constants $\alpha_\pm$ and $\kappa_\pm$ such that

$$\alpha_-\,dx \leq a(dx) \leq \alpha_+\,dx,$$
$$\kappa_-\,dx \leq c(dx) \leq \kappa_+\,dx.$$

We assume furthermore that the lengths L_x of the clones are uniformly stochastically bounded from above and from below. This means that there exist nonnegative random variables $L_\pm$ such that L_+ is integrable, such that L_- is not almost surely zero and such that, for every x and t,

$$\mathbb{P}(L_- \geq t) \leq \mathbb{P}(L_x \geq t) \leq \mathbb{P}(L_+ \geq t).$$

In particular, the family $(L_x)_x$ must be uniformly integrable.

PROPOSITION 6. *The assumptions above imply that there exist positive constants $\varrho_\pm < 1$ and finite positive constants $\nu_\pm$ such that, for every G,*

$$\varrho_- G \le \mathbb{E}(\mathcal{O}_G) \le \varrho_+ G, \qquad \nu_- G \le \sigma^2(\mathcal{O}_G) \le \nu_+ G.$$

In these inequalities, ϱ_- corresponds to the homogeneous case of parameters κ_+, α_+ and L_+, and ϱ_+ to the homogeneous case of parameters κ_-, α_- and L_-. As regards the variance, the dependence is not so straightforward, at least the dependence that our techniques yield. The parameter ν_+ that we exhibit depends on α_- alone, a result which may seem surprising, and the parameter ν_- depends on κ_+, ϱ_+ and ϱ_-.

PROOF OF PROPOSITION 6. The bounds on $\mathbb{E}(\mathcal{O}_G)$ would follow from the fact that

$$\varrho_- \le r(z) \le \varrho_+,$$

for any z and for positive $\varrho_\pm < 1$. Such bounds on $r(z)$ themselves stem from the fact that the distribution of the ocean, as a random subset of the real line, is nonincreasing with respect to the intensities of the processes of the clones and of the anchors. Hence, by a coupling argument, the value of $\mathbb{E}(\mathcal{O}_G)$ lies between its value for the homogeneous processes of densities α_+ and κ_+ on the one hand, and α_- and κ_- on the other hand, the distributions of the lengths L_x being fixed.

We now examine the influence of the distributions of the lengths. Once again by a coupling argument, the uniform replacement of the distributions of the lengths L_x by the distribution of L_+ yields longer clones, hence longer islands, hence a stochastically smaller ocean. This proves the lower bound of $\mathbb{E}(\mathcal{O}_G)$. Comparison with L_- yields the upper bound.

Our proof of the lower bound of $\sigma^2(\mathcal{O}_G)$ goes as follows. One knows that

$$\sigma^2(\mathcal{O}_G) = \int_0^G \int_0^G \bigl(r(z,z') - r(z)r(z')\bigr)\,dz\,dz',$$

and that the expression $r(z,z') - r(z)r(z')$ is nonnegative for every z and z'. Assume that there exist positive δ and ε such that, for every z and z' such that $|z - z'| \le \varepsilon$,

$$r(z,z') - r(z)r(z') \ge \delta.$$

The lower bound of $\sigma^2(\mathcal{O}_G)$ would follow. Now, for every $z \le z'$, if z' is in $\mathcal{O}$ and if there is no right end of clone in $[z,z']$, then z is in $\mathcal{O}$. Hence,

$$r(z') = r(z,z') + \mathbb{P}(z \notin O, z' \in O) \le r(z,z') + \mathbb{P}(D),$$

with $D := \{\mathcal{C} \cap ([z, z'] \times \mathbb{R}^+) \neq \varnothing\}$. By definition of the intensity of the Poisson process $\mathcal{C}$,

$$\mathbb{P}(D) = 1 - e^{-c([z,z'])} \leq c([z, z']) \leq \kappa_+(z' - z).$$

Since $r(z') \geq \varrho_-$ and $r(z) \leq \varrho_+$, this proves the lower bound

$$r(z, z') - r(z)r(z') \geq (1 - \varrho_+)\varrho_- - \kappa_+(z' - z).$$

This in turn shows the desired inequality for $z \leq z'$ and $z' - z$ small enough.

As regards the upper bound, it is enough to bound from above the integrals of $r_0(z, z')$ and $r_1(z, z')$, since $r_3(z, z')$ is nonnegative. In the expression of $r_0(z, z')$, for all fixed values of x and y, $J(x \mid z, z' \mid y)$ is a nonincreasing function of the distributions of the lengths L_x and of the intensity of the clones, since having more clones and longer clones only makes the ocean smaller. Thus $r_0(z, z')$ is bounded from above by its value when one replaces $c(dt)$ by $\kappa_- \, dt$ and the distribution of every L_x by the distribution of L_-. Likewise, the interpretation of $B(dx, dy)$ as the joint distribution of the positions of the rightmost anchor to the left of z and of the leftmost anchor to the right of z, and a coupling between two processes of anchors with comparable intensities, show that the anchors become stochastically more distant from z when one replaces $a(dt)$ by the smaller intensity $\alpha_- \, dt$. Hence the probability that z is not covered by an anchored clone cannot decrease. Thus, replacing $a(dt)$ by $\alpha_- \, dt$ cannot make $r_0(z, z')$ decrease.

Finally, the contribution of r_0 in the value of $\sigma^2(\mathcal{O}_G)$ is bounded from above by its value in the homogeneous case which uses the values α_-, κ_- and L_-, that is, for instance, by $2G/\alpha_-$. Likewise, the contribution of r_1 to the value of $\sigma^2(\mathcal{O}_G)$ is at most $2G/\alpha_-$. This yields the desired upper bound with $\nu_+ := 4/\alpha_-$. $\square$

Alternatively, when $L_z \leq \ell$ almost surely and for every z, recall from the end of Section 3.3 that $r(z, z') = r(z)r(z')$ as soon as $|z - z'| > \ell$, hence $\sigma^2(\mathcal{O}_G)$ is at most the area of the part of the square $[0, G]^2$ inside the diagonal strip $|z - z'| \leq \ell$, that is, at most $2\ell G - \ell^2$ when $G \geq \ell$, and $\sigma^2(\mathcal{O}_G)$ is at most G^2 for every G. Hence $\sigma^2(\mathcal{O}_G) \leq 2\ell G$ for every G.

Finally, we mention that one can adapt the proofs in this section to some cases when the intensities of the clones and of the anchors are zero in some places, as long as the intensities stay bounded from below on regions which are spread out enough.

4.4. *Convergence in distribution.* We first explain how one could prove the convergence of the moments by elementary techniques, then we show that general invariance results apply, which yield directly the desired convergence.

4.4.1. *Method of moments.* Assume first that $L \le \ell$ almost surely. Then, a crucial remark from the end of Section 3.3 is that the events $\{Z \subset \mathcal{O}\}$ and $\{Z' \subset \mathcal{O}\}$ are independent as soon as the distance between every point in Z and every point in Z' is at least ℓ. Furthermore,

$$\mathbb{E}\big((\mathcal{O}_G - \varrho G)^n\big) = \int_{Z \in [0,G]^n} \pi(Z)\,dZ, \qquad \pi(Z) := \prod_{z \in Z} (\mathbf{1}\{z \in \mathcal{O}\} - \varrho).$$

If $Z = Z' \cup Z''$ with $|z' - z''| \ge \ell$ for every $z' \in Z'$ and every $z'' \in Z''$, one gets $\mathbb{E}(\pi(Z)) = \mathbb{E}(\pi(Z'))\mathbb{E}(\pi(Z''))$.

For instance, if $n = 3$, every nontrivial partition of Z includes at least one singleton, hence $\mathbb{E}(\pi(Z))$ is zero except when all the distances between the nonempty subsets of Z are at most ℓ. Ordering the points z, z' and z'', we are left with the domain

$$z \le z' \le z + \ell, \qquad z' \le z'' \le z' + \ell,$$

whose volume is at most $\ell^2 G$. Hence $\mathbb{E}((\mathcal{O}_G - \varrho G)^3)$ is of order at most G.

If $n = 4$, the only difference with the $n = 3$ case is due to the partitions of Z into two pairs Z' and Z''. These contribute to the result even when the distance from Z' to Z'' is large. Every such $\mathbb{E}(\pi(Z'))$ and $\mathbb{E}(\pi(Z''))$ is of order at most G, hence $\mathbb{E}((\mathcal{O}_G - \varrho G)^4)$ is of order at most G^2.

Likewise, for every positive integer k, the moments $\mathbb{E}((\mathcal{O}_G - \varrho G)^{2k})$ and $\mathbb{E}((\mathcal{O}_G - \varrho G)^{2k+1})$ are both of order at most G^k.

One can also compute the asymptotics of the moments of $\mathcal{O}_G$ as $G \to \infty$. To do this, one starts from the expression of $\mathbb{E}((\mathcal{O}_G - \varrho G)^{2n})$ as the integral of $\mathbb{E}(\pi(Z))$ over the points Z in $[0, G]^{2n}$. When there exists a partition of Z into two parts Z' and Z'' at a distance at least ℓ, $\mathbb{E}(\pi(Z))$ is the product $\mathbb{E}(\pi(Z'))\mathbb{E}(\pi(Z''))$. The remaining points Z span a volume in $[0, G]^{2n}$ which is $o(G^n)$, hence they contribute to a vanishing part of the asymptotics.

This yields recursions between the asymptotic moment of degree $2n$ and the asymptotic moments of even degrees at most $2n - 2$. One can deduce from these recursions the convergence of the moments of $(\mathcal{O}_G - \varrho G)/\sqrt{G}$ to the moments of a Gaussian random variable.

Finally, one could adapt this strategy to the case where L is unbounded, thus reaching the same conclusion.

4.5. *Direct method.* A stronger conclusion obtains directly from classical results by Doukhan, Massart and Rio [2], for every square integrable L. To see this, introduce for every integer n, the random variable

$$X_n := \mathcal{O}([n, n+1]) - \varrho.$$

Let $\mathcal{F}_n$ denote the σ-algebra generated by the collection $(X_i)_{i \le n}$, and let $\mathcal{G}_n$ denote the σ-algebra generated by the collection $(X_i)_{i \ge n}$. The sequence $(X_n)_n$ is generated by the action of the shift

$$\vartheta : (x, t) \mapsto (x + 1, t),$$

on $\mathbb{R} \times M$, since $X_n = X_0 \circ \vartheta^n$ for every integer n. The strong mixing coefficients α_n associated to the stationary sequence $(X_n)_n$ are defined, for any integer $n \geq 0$, by

$$\alpha_n := \sup\{\mathbb{P}(D \cap D') - \mathbb{P}(D)\mathbb{P}(D'); D \in \mathcal{F}_0, D' \in \mathcal{G}_n\}.$$

Since $|X_0| \leq 1$ almost surely, the condition in Doukhan, Massart and Rio [2] reduces to the summability of the series of general term α_n. Neglecting the influence of the anchors does not decrease the value of α_n. Thus $\alpha_n \leq \mathbb{P}(D_n)$, where D_n is the event that at least one clone covers both points 0 and n. One can bound each $\mathbb{P}(D_n)$ as follows:

$$\mathbb{P}(D_n) = 1 - J(n) \leq \int_n^{+\infty} \kappa \mathbb{P}(L \geq t)\, dt.$$

This shows that the sequence of general term $\mathbb{P}(D_n)$ is summable as soon as L is square integrable. (In fact, this sequence is summable if and only if L is square integrable; we omit the proof.) This shows that the functional invariance stated in Theorem 1 holds, at least for the processes Θ_G such that G is an integer. The general case is an easy consequence, since $\mathcal{O}_G$ depends on G in a monotone way.

Equivalently, one can write directly $\mathcal{O}_G$ as

$$\mathcal{O}_G = \varrho G + \int_0^G Y_x\, dx,$$

where the stationary centered family $Y_x := \mathbf{1}\{x \in \mathcal{O}\} - \varrho$ is indexed by the real numbers x. The same conclusion obtains.

Acknowledgments. We wish to thank Sophie Schbath for an introduction to this subject and for instructive discussions, Julien Michel for the positive association argument used in Section 4.2 and an anonymous referee for a careful reading of the first version of this paper.

REFERENCES

[1] 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.

[2] DOUKHAN, P., MASSART, P. and RIO, E. (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.* **30** 63–82. MR1262892

[3] 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.

[4] GRIGORIEV, A. V. (1993). Theoretical predictions and experimental observations of genomic mapping by anchoring random clones. *Genomics* **15** 311–316.

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

[6] ROY, R. (1991). Percolation of Poisson sticks on the plane. *Probab. Theory Related Fields* **89** 503–517. MR1118561

[7] SCHBATH, S. (1997). Coverage processes in physical mapping by anchoring random clones. *J. Comput. Biol.* **4** 61–82.

[8] SCHBATH, S., BOSSARD, N. and TAVARÉ, S. (2000). The effect of non-homogeneous clone length distribution on the progress of an STS mapping project. *J. Comput. Biol.* **7** 47–58.

UNIVERSITÉ CLAUDE BERNARD LYON 1
INSTITUT CAMILLE JORDAN UMR 5208
DOMAINE DE GERLAND
50, AVENUE TONY-GARNIER
69366 LYON CEDEX 07
FRANCE
E-MAIL: Didier.Piau@univ-lyon1.fr
URL: http://lapcs.univ-lyon1.fr