Title
Predicting Progress in Shotgun Sequencing with Paired Ends

Authors
Yeh, Ru-Fang
Speed, Terence P
Waterman, Michael S
et al.

Publication Date
2002-10-11

Predicting Progress in Shotgun Sequencing with Paired Ends

Ru-Fang Yeh[1*], Terence P Speed[2,3], Michael S Waterman[4,5] and Xiaoman Li[5,6]

[1] Center for Bioinformatics and Molecular Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560

[2] Department of Statistics, University of California, Berkeley, CA 94720-3860

[3] Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3050, Australia

[4] Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

[5] Celera Genomics, Rockville, MD 20878

[6] Department of Mathematics, University of Southern California, Los Angeles, CA 90089

October 11, 2002

Abstract

Paired-end shotgun sequencing has become widely used for large-scale sequencing projects in recent years, including whole genome shotgun sequencing and map-based BAC clone sequencing. Under this scheme, sequences from both ends of random clones are determined and assembled into sequence contigs. The sequence data and their linking information are used to construct clone maps in the form of scaffolds. In order to plan a cost-effective sequencing project utilizing such an approach, it is crucial to have knowledge of the expected project progress in relation to parameters such as insert size, clone

*To whom correspondence should be addressed. Email: *rufang@biostat.ucsf.edu* Tel: (415) 514-2664 Fax: (415) 476-6014

lengths and redundancy. There has been a lack of theoretical analysis for the paired-end sequencing strategy due to the difficulty of correlated ends. Here we present a mathematical analysis for the progress of a sequencing project employing such a scheme. Formulae for various measures of the expected progress such as expected number and size of scaffolds are derived and assessed by Monte Carlo simulations for parameter sets used in the human genome project.

Keywords. paired-end sequencing; Poisson process; physical mapping; DNA sequencing; scaffolds; Lander-Waterman analysis

1 Introduction

Random cloning is the basis of genomic mapping and sequencing. It enables the manipulation of DNA from specific regions of interest, and the allows the possibility of characterizing a mega-base-pair genome using clones and sequences of manageable size. Physical mapping and sequencing has the goal of recovering the positional information on cloned DNA fragments and their associated characteristics such as markers along the chromosome, fingerprints and sequences, so that overlapping clones can be treated as contiguous equivalents. Most physical mapping and sequencing projects can be summarized into three steps: library creation, overlap detection and layout, and finishing. Strategies and projects differ in the scales of their target DNA, types of clone library, methods for detecting clone overlaps and layout, and number of iterations required to complete a project.

Large-scale sequencing projects require accurate, efficient, and automated DNA sequencing methods, and usually involve one or more stages of mapping followed by sequencing. Traditionally physical mapping including the mapping phase of directed sequencing strategies is done separately, prior to sequencing, using information such as restriction fingerprints or sequence-tagged sites. The level of mapping detail determines the amount of sequencing effort and customized gap-closure experiments required to complete a project. Different strategies try to balance the effort in mapping and sequencing, and the choice of methods usually depends on the project size and available laboratory resources. As automated sequencing technology advances and rapidly scales up, large-scale sequencing projects are moving towards methods with minimal effort in the initial mapping phase, with

a popular extreme being whole-genome shotguning [10]. A key element of whole-genome shotgun sequencing is the use of a paired-end sequencing scheme [3] which provides mapping information from the sequence data derived from both ends of clones. Such an approach combines the acquisition of sequence with the acquisition of positional information. It also provides crucial positional information permitting walking across the many repetitive structures in an eukaryote genome.

In paired-end shotgun sequencing, sequences of both ends of random clones of known approximate length are determined. Identical and sufficiently similar overlapping sequences are assembled into sequence contigs. Using positional information inferred from end-sequence overlaps and approximate clone lengths, clones can be organized into ordered and oriented sets, termed *scaffolds*. The target genomic DNA can then be obtained by sequencing the unsequenced parts of the map, including actual gaps, regions left uncovered by the clones, and *sequence-mapped gaps* (abbreviated SMGs) [2] internal to some clones. The mapping objective of a paired-end shotgun sequencing project is typically to obtain one scaffold spanning the entire target DNA when possible, since gaps between scaffolds are rather more difficult to fill in comparison with SMGs.

Mathematical analysis of the progress of a mapping project such as that of Lander and Waterman [5] is important for planning and monitoring a cost-effective project. The progress of a sequencing project can be described by a number of measures, such as the number and size of *islands* and gaps, total island and sequence coverage, and probability of project closure. The three types of islands in the context of paired-end sequencing are illustrated in Figure 1. They are (1) actual islands, two clones being in the same actual island if they are connected by a chain of pairwise overlapping clones; (2) scaffolds (apparent islands), two clones being in the same scaffold if they are connected by a chain of clones whose end sequences overlap; and (3) sequence islands, where two basepairs in clone ends are in the same sequence island if they are connected by a chain of overlapping sequenced clone ends. All three types of islands are of interest, but in practice only scaffolds and sequence islands are direct products. Actual islands can be analyzed approximately by the Lander-Waterman theory [5] and by the exact analysis as described in Roach et al [7]. Sequence islands can be analyzed similarly, and have the same expected values but larger variances for most quantities of interest as in a shotgun project consisting of same number of independent sequence reads [6]. However, there

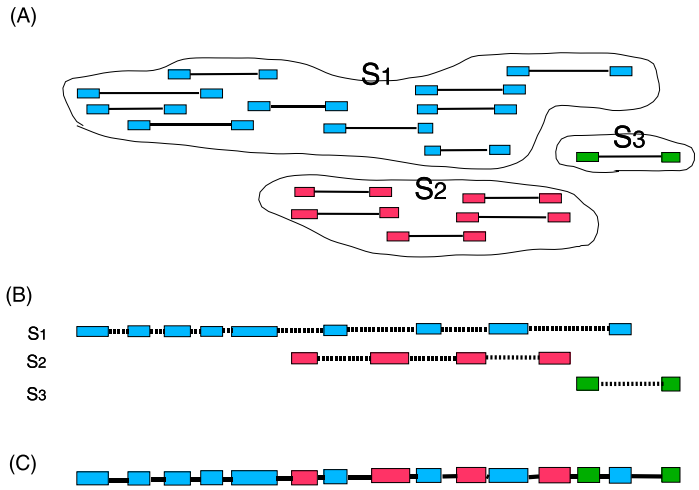

Figure 1: Three types of islands. (A) All sixteen overlapping clones belong to the same actual island, but in three different scaffolds by their end sequence (represented by blocks) overlaps as circled. (B) The thirty-two end sequences form nine sequence islands in scaffold S_1, four in S_2 and two singleton sequence islands in S_3 as indicated. The apparent SMGs are represented by dashed lines within scaffolds. (C) The actual SMGs are shown in solid lines between neighboring sequence islands in the actual island.

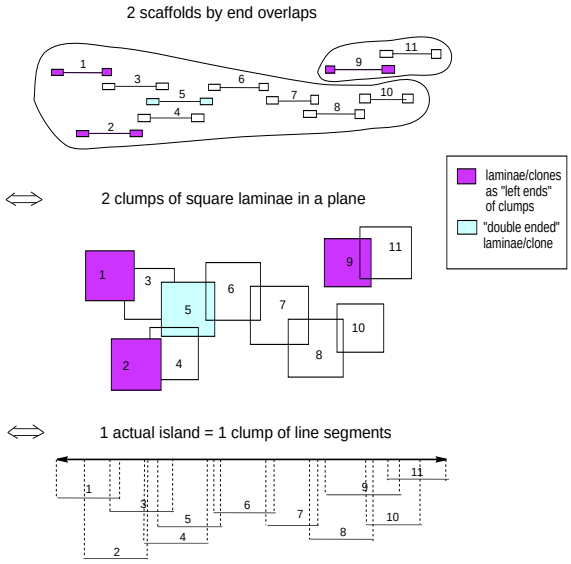

Figure 2: Analogy of scaffold structure to two-dimensional clumps of square laminae. Lamina/clone 5 merges two non-overlapping sub-clumps (1,3), (2,4) specified by the left ends 1 and 3 into the same clump. Actual islands can also be viewed as the projection of the square laminae clumps onto the line.

has been lack of theoretical analysis for the observable scaffolds due to their typically complicated layout as illustrated in Figure 1. If we view actual islands as clumps formed by line segments in the real line, counting scaffolds by end overlaps is analogous to counting clumps formed by square laminae in the plane (Figure 2). Members of a one-dimensional clump can only be joined by consecutively overlapping neighbors, whereas in a two-dimensional clump, two non-overlapping neighbors can be linked by members further apart. This dependence by ends creates complex scaffold structures and requires a more sophisticated analysis [4]. Port *et al.* [6] attempted to approximate scaffolds by a conceptual greedy island approach, but this yielded unsatisfying results. Roach et al [8] provided some insights by Monte Carlo simulations. The aim of this paper is to present a useful and practical theoretical analysis of the properties of scaffolds in relation to relevant parameters (such as target size, clone-length distribution, and clone redundancy) using the same Poisson process framework which was used in the Lander-Waterman analysis of Arratia et al [1]. Monte Carlo simulations are also used to assess the performance of our formulae with the parameter sets used in the Drosophila and human genome projects.

2 Results

Notation and assumptions Based on the nature of clone library creation, the target genomic DNA of size G can be modeled as a long interval $(0, G)$ on the real line $\mathbf{R}$, and the N clones as random intervals along the line. We model the left ends of clones $\cdots < \xi_{-2} < \xi_{-1} \leq 0 < \xi_1 < \cdots < \xi_N < G \leq \xi_{N+1} < \xi_{N+2} < \cdots$ with a homogenous Poisson point process on the real line with intensity $\lambda = \frac{N}{G}$, which essentially mark the uniformity of the random intervals. For a clone starting at t, the left end point $\xi_i = t$ is associated with three lengths (L_i, E_{iL}, E_{iR}), where $\{L_i\}$ are independently and identically distributed (abbreviated *i.i.d.*) samples of a clone length distribution $F_L(\cdot)$ with mean L and variance σ_L^2; the left and right end-sequence lengths E_{iL}, E_{iR} are *i.i.d* samples of a sequence read length distribution $F_E(\cdot)$ with mean E and variance σ_E^2. Note that we also use L_i, E_{iL}, E_{iR} to label the clone starting at $\xi_i = t$, its left and right end respectively where appropriate. Define a *scaffold* as a set of line segments (clones) connected via overlapping end-sequences, assuming

all sequence overlaps are detected. Let $p(t)$ denote the probability that a clone L_t with its left end at t begins a scaffold. In addition, define a *actual sequence-mapped gap* (abbreviated *actual SMG*) to be a target segment covered by some clones but not covered by any end-sequence, whereas an *apparent SMG* (observable sequence gaps under the paired-end sequencing scheme) is defined as a sequence gap internal to a scaffold. The notation is summarized below.

G: length of target genomic DNA;

N: number of clones;

$\{\xi_t\}$: the clone process;

$\lambda = \frac{N}{G}$: intensity of the clone process;

L_t: clone (length) starting at t, $t \in (0, G)$;

$F_L(\cdot)$: distribution function of clone length;

L: average clone length;

σ_L^2: variance of clone length;

E_{tL}: left end-sequence (length) of the clone L_t;

E_{tR}: right end-sequence (length) of the clone L_t;

$F_E(\cdot)$: distribution function of end-sequence length;

E: average end-sequence length;

σ_E^2: variance of end-sequence length;

$p(t)$: scaffold retention probability for L_t;

Finite correction for edge effects The above assumptions ignore the boundaries of the target DNA, which could be taken into account by using the *effective* genome size $G_e = G - L + 1$ in place of G for the clone process intensity $\lambda_e = \frac{N}{G_e}$ to ensure that all clones fall in $(0, G)$. This correction becomes relevant in practice when L/G is not negligible (typically $L \ll G$). Similarly, the analysis follows by replacing the sequence read length E by the *effective* sequence read length $E_e = E - T$ when only sequence overlaps $> T$ bp are detected. Furthermore, an additional scaffold covering the origin should be added to the count of scaffolds in $(0, G)$ to account for the boundary case.

Counting scaffolds for typical parameters With the above assumptions and notations, the scaffold process can be derived by removing points $\{\xi_i\}$ that do not begin a scaffold, and hence *thinning* the clone Poisson process back to $-\infty$ with the scaffold retention probability $p(t)$. Note that the homogeneity of the clone process implies that $p(t) = p$, independent of location t. More precisely, the thinned process represents clones beginning scaffolds, and has intensity λp. The retention probability p strongly depends on the range of parameters (the ratio L/E in particular), and can be calculated as in following proposition. By clone inclusion we mean that one clone is entirely contained in another, longer clone.

Proposition 1 (The scaffold retention probability $p(t)$)

(i) *For $L/E \leq 3$, $p(t) = e^{-\lambda L}$.*

(ii) *For $L/E > 3$, suppose that clone inclusions are negligible (valid when σ_L is not too large, for instance, $\sigma_L \leq L$), then*

$$p(t) \approx e^{-\lambda L} + e^{-\beta\lambda E}(e^{-\alpha\lambda E} - e^{-\lambda(L-3E)})(2 - e^{-\beta\lambda E}). \qquad (1)$$

α and β are functions of length distributions $F_L(\cdot)$ and $F_E(\cdot)$ defined below. Let $H(\cdot)$ denote the distribution of $\Delta_L = L_r - L_l$ for L_r, L_l i.i.d. $\sim F_L(\cdot)$, and $K(\cdot)$ denote the distribution of $\Delta_E = E_i - E_j$ for E_i, E_j i.i.d. $\sim F_E(\cdot)$.

$$\begin{aligned}
\alpha &\equiv \frac{1}{E}\int\int_0^\infty F_E(u)(1 - F_E(|u-d|))\,du\,dH(d) \approx min(2, \frac{\sqrt{(\sigma_L^2 + \sigma_E^2)/2}}{E});\\
\beta &\equiv 2 - (E + \frac{\sigma_E^2}{E})/(L - 3E - \mathbf{E}[(\Delta_L - \Delta_E)\mathbf{1}(\Delta_L > \Delta_E)])\\
&\approx min(2, max(0, 2 - \frac{E + \sigma_E^2/E}{L - 3E - \sqrt{(\sigma_L^2 + \sigma_E^2)/2}})).
\end{aligned}$$

(iii) *For $L/E > 3$, allowing the possibility of clone inclusion,*

$$\begin{aligned}
p(t) = \text{Formula } (1) + (1 - p_1)\,(&e^{-(2\,|\,\alpha)\lambda E} - e^{-\lambda L})(e^{-\lambda E} - e^{-2\lambda E}\\
&+ e^{-(2\,|\,\beta)\lambda E} - e^{-2\lambda E\,|\,\int_{z+\Delta_z}^{z+\Delta_r} - \Delta_{zr}|1\;\; F_\Delta(s\,|\,L_z\;\; E_{zr}\;\; z)dzdV(z\;\; s)|})
\end{aligned} \qquad (2)$$

where p_1 is defined as the conditional probability that some clone interlaces L_t given t is covered by some clone not overlapping L_t at ends.

$$p_1 = \frac{2\int_0^\infty F_\Delta(u)|F_\Delta(u\,|\,L_z\;\; E_{zr})\;\; F_\Delta(u\,|\,E_{zr}\,|\,E_{zl})|du}{2\int_0^\infty F_\Delta(u)|F_\Delta(u\,|\,L_z\;\; E_{zr})\;\; F_\Delta(u\,|\,E_{zr}\,|\,E_{zl})\,|\,1\;\; F_\Delta(u\,|\,L_z\,|\,E_{zr})|du}.$$

When $L \leq 3E$, all clone overlaps involve some degree of end overlaps and vice versa, implying the equivalence of actual islands and scaffolds. Hence $p(t) = \Pr\{t \text{ is not covered by any clones}\} = e^{-\lambda L}$. For $L > 3E$, however, the event that L_i initiates a scaffold includes not only the above event of beginning an actual island, but also includes the other complex event that t is covered by some clone L_r interlacing or encompassing L_i. Typically $L >> 3E$: E is the length of a sequencing read ≈ 500 bp with small variance $\sigma_E \approx 50$ bp; L depends on the type of clone library. For instance, $L \approx 3$ kb with $\sigma_L \approx 600$ bp for the plasmid library used in the Berkeley Drosophila Genome Project for BAC sequencing, and L is a mixture of 2 kb, 10 kb, and 50 kb libraries in Celera's whole-genome shotgun Human Genome sequencing [9]. For most common parameter sets, the explicit formula (1) is sufficient, and the contribution of clone inclusion as in the more complicated formula (2) is negligible. However, formula (2) should be used when the lengths of the clones vary significantly (such as with the mixture clone sizes in Celera's human genome project) and clone inclusion occurs frequently. See Methods (Section 4) for the details of the calculation of $p(t)$ for $L > 3E$. With the scaffold retention probability $p(t)$ described above, we have

Theorem 1 (Formula of scaffolds)

(i) *The expected proportion of the genome covered by scaffolds* $= 1 - e^{-\lambda L}$.

(ii) *The expected number of scaffolds* $= 1 - e^{-\lambda L} + Np$.

(iii) *Define an effective scaffold to be a scaffold that is not included in any other scaffolds and hence contributes at least partially to the total coverage. Let* $p_c(t) = p_c$ *denote the probability that a clone with left end at* t *begins an effective scaffold. Then*

$$\Pr\{\text{project closure}\} = e^{-Np_c}(1 + e^{-\lambda L}(Np_c - 1)) \approx e^{-Np_c}.$$

p_c can be approximated by one of the following formula:

$$e^{-\lambda L} + (e^{-(2-\alpha)\lambda E} - e^{-\lambda L})(2e^{-2\lambda E} - e^{-(2-\beta)\lambda E})e^{-2\lambda E} \quad (3)$$

$$e^{-\lambda L} + (e^{-(2-\alpha)\lambda E} - e^{-\lambda L})e^{-2\lambda E} \quad (4)$$

$$e^{-\lambda L} + (e^{-(2-\alpha)\lambda E} - e^{-\lambda L})(e^{-2\lambda E} - e^{-(2-\beta)\lambda E}) \quad (5)$$

$$e^{-\lambda L} + (e^{-(2-\alpha)\lambda E} - e^{-\lambda L})(2e^{-2\lambda E} - e^{-(2-\beta)\lambda E}) \cdot (2e^{-\lambda E} - e^{-(2-\beta)\lambda E} - e^{-(1-\alpha)\lambda E}) \quad (6)$$

(iv) Assume that scaffold lengths and orders (number of clones per scaffold) are i.i.d.. Then

$$\text{the expected order of a scaffold} = 1/p\,;$$
$$\text{the expected length of a scaffold} \approx \frac{G\cdot(1-e^{-\lambda L})}{1-e^{-\lambda L}+Np_e}.$$

(v) Suppose that scaffold sizes and numbers of sequence islands per scaffold are i.i.d. respectively.

$$\mathrm{E}\{\text{number of apparent SMGs in a scaffold}\} = \frac{1+(2N-1)e^{-2\lambda L}}{1-e^{-\lambda L}+Np} - 1.$$
$$\mathrm{E}\{\text{size of an apparent SMG}\} = \frac{\left(\frac{G\cdot(1-e^{-\lambda L})}{1-e^{-\lambda L}+Np_e} - \frac{G(1-e^{-2\lambda L})}{1-e^{-\lambda L}+Np}\right)}{\frac{1+(2N-1)e^{-2\lambda L}}{1-e^{-\lambda L}+Np} - 1}.$$

Proof.

(i) The total coverage by scaffolds depends only on the underlying configuration of clones. Hence

$$\begin{aligned} & \mathrm{E}\{\text{total coverage by scaffolds}\} \\ = \; & \mathrm{E}\{\text{total coverage by actual islands with the same set of clones}\} \\ = \; & 1-e^{-\lambda L}. \end{aligned}$$

(ii) Since the scaffold process has intensity $\lambda p(t)$, the number of scaffolds starting in $(0, G)$ is a Poisson random variable N_s with mean $\int_0^G \lambda p(t)\,dt = Np$. To account for the edge effect of the finite case, an additional scaffold should be added when the origin 0 is covered by some scaffold beginning before 0. So $\mathrm{E}\{\text{number of scaffolds}\} = \mathrm{E}[\mathbf{1}(0 \text{ is covered}) + N_s] = 1-e^{-\lambda L}+Np$.

(iii) See Methods (Section 4) for the derivation of p_e. The concept of effective scaffolds allows the translation of complete coverage by one scaffold to the computationally simpler event of having a unique (effective) scaffold. In other words, $p_e = \Pr\{\text{Number of effective scaffold} = 1\}$. This distinction is typically significant since there are usually small scaffolds or singleton clones "embedded" in the larger scaffold of complete coverage. Consider the finite correction for the edges similar to counting scaffolds, and

let N_e be the number of effective scaffolds beginning in $(0, G)$. $N_e \sim Poisson(Np_e)$.

$$\begin{aligned} & \Pr\{\text{Number of effective scaffolds} = 1\} \\ = \; & \Pr\{0 \text{ is covered and } N_e = 0\} + \Pr\{0 \text{ is not covered and } N_e = 1\} \\ = \; & (1 - e^{-\lambda L})e^{-Np_e} + e^{-\lambda L} N p_e e^{-Np_e} \approx e^{-Np_e}. \end{aligned}$$

(iv) The expected scaffold order can be computed easily by the ratio of clone and scaffold processes $\lambda/(\lambda p) = 1/p$. We approximate the expected scaffold length by the expected effective scaffold length. By definition, effective scaffolds are either disjoint or with actual overlaps mostly limited to $L - E$ since further overlaps require successive clone interlacing with negligible probability. So

$$\mathrm{E}\{\text{total length of scaffold coverage}\} \approx \mathrm{E}\{\textstyle\sum \text{effective scaffold lengths}\}$$

Assuming scaffold sizes are *i.i.d.*,

$$\mathrm{E}\{\text{scaffold length}\} \approx \frac{G \cdot (\text{total coverage})}{\mathrm{E}\{\text{number of effective scaffolds}\}} = \frac{G \cdot (1 - e^{-\lambda L})}{1 - e^{-\lambda L} + Np_e}.$$

(v) By the definition of apparent SMGs and the *i.i.d.* assumption on the number of sequence islands in each scaffold,

$$\begin{aligned} \mathrm{E}\{\text{No. of apparent SMGs per scaffold}\} &= \frac{\mathrm{E}\{\text{No. of sequence islands}\}}{\mathrm{E}\{\text{No. of scaffolds}\}} - 1 \\ &= \frac{1 + (2N - 1)e^{-2\lambda E}}{1 - e^{-\lambda L} + Np} - 1. \end{aligned}$$

The *i.i.d.* scaffold assumptions and apparent SMG sizes imply that

$$\begin{aligned} & \mathrm{E}\{\text{apparent SMG size}\}\,\mathrm{E}\{\text{No. of apparent SMGs per scaffold}\} \\ = \; & \mathrm{E}\{\text{scaffold size} - \text{total length of sequence islands per scaffold}\}. \end{aligned}$$

$$\begin{aligned} \text{So} \qquad \mathrm{E}\{\text{apparent SMG size}\} &= \frac{\mathrm{E}\{\text{scaffold length}\} - \frac{G(1 - e^{-2\lambda E})}{\mathrm{E}\{\text{No. of scaffolds}\}}}{\mathrm{E}\{\text{No. of apparent SMGs in a scaffold}\}} \\ &= \frac{\frac{G \cdot (1 - e^{-\lambda L})}{1 - e^{-\lambda L} + Np_e} - \frac{G(1 - e^{-2\lambda E})}{1 - e^{-\lambda L} + Np}}{\frac{1 + (2N - 1)e^{-2\lambda E}}{1 - e^{-\lambda L} + Np} - 1}. \end{aligned}$$

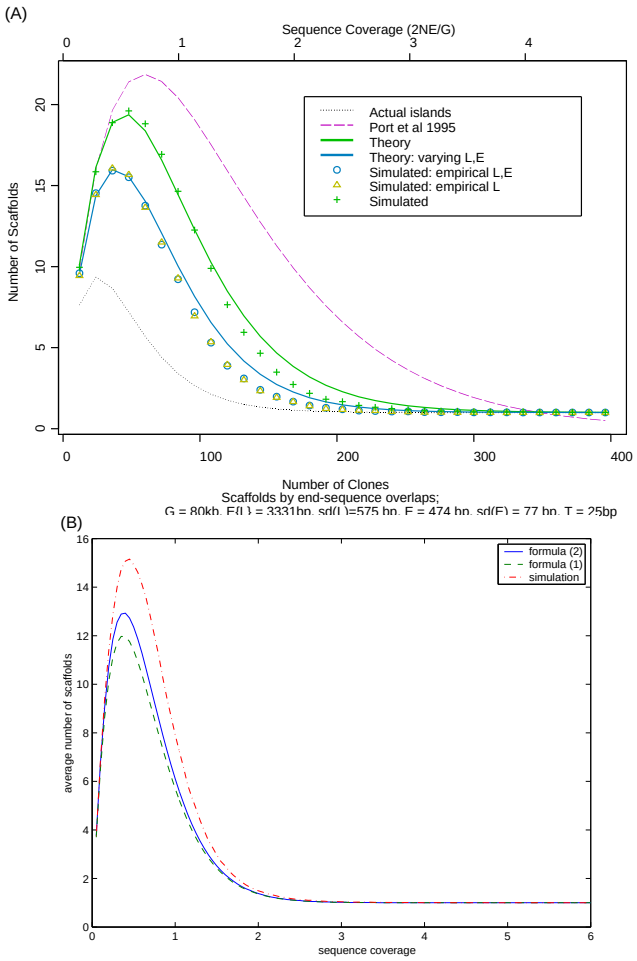

Figure 3: Comparison of theoretical predictions of the expected number of scaffolds to simulation results. (A) Each simulated point is the average of 500 independent simulations of the P1 clone sequencing project at the Berkeley Drosophila Genome Project: $G = 80$ kb, $T = 25$ bp, and clone library of empirical $F_L(\cdot)$ and $F_E(\cdot)$ with $L = 3331$ bp, $\sigma_L = 575$ bp, $E = 474$ bp, $\sigma_E = 77$ bp. Simulations were carried out with constant clone and end-sequence read lengths (|), empirical clone lengths and constant read lengths ($\triangle$), and empirical clone and read lengths ($\circ$). The greedy island formula with constant clone and read lengths from [6] and actual island formula from [5] are plotted for reference. (B) Comparison of simulations to the theoretical formula using different approximation for $p(t)$. Each simulated point represent the average of 3000 simulations of a project with $G = 80$ kb, $T = 25$ bp, constant read length $E = 500$ bp, and an equal mixture of two constant clone species of 5 kb and 2 kb.

Figure 3 and 4 compare the theoretical prediction in Theorem 1 to simulated results for the expected number of scaffolds and average scaffold size in bp using two different sets of parameters. In each simulated paired-end sequencing project, the locations of clones were represented as *i.i.d.* draws from a uniform distribution on the integers $\in [0, G-L]$; sizes of clones and end-sequence read lengths were either constant or *i.i.d.* samples of the corresponding empirical distribution as noted.

Counting scaffold contigs In practice, measures of *scaffold contigs* are usually observed instead and they are better indicators for the project progress. A *scaffold contig* excludes singletons and refers to a scaffold containing multiple clones. An easy modification to Theorem 1 can be derived for scaffold contigs.

Theorem 2 (Formula for scaffold contigs)
With the same notation and assumption as in Theorem 1,

(i) The probability that a clone starting at t does not overlap remaining clones at the ends,

$$\tau(t) = \tau = e^{-(G-2L)\lambda E}. \tag{7}$$

(ii) The expected number of scaffold contigs $= (1-e^{-\lambda L})(1-\tau) + N(p-\tau)$.

(iii) The expected length of a scaffold contig $= \frac{G(1-e^{-\lambda L})-Le^{-2\lambda L}}{(1-e^{-\lambda L})(1-r)+N(p-r)}$.

(iv) $\mathrm{E}\{$*No. of apparent SMGs per scaffold contig*$\} = \frac{1+(2N-1)e^{-2\lambda E}-r(1-e^{-\lambda L}+2N)}{(1-e^{-\lambda L})(1-r)+N(p-r)} - 1$.

(v) $\mathrm{E}\{$*size of an apparent SMG in a scaffold contig*$\} =$

$$\frac{\frac{G(1-e^{-\lambda L})-Le^{-2\lambda L}}{N(p_a-e^{-2\lambda L})} - \frac{G(1-e^{-2\lambda E})-2Er(1-e^{-\lambda L}+N)}{(1-e^{-\lambda L})(1-r)+N(p-r)}}{\frac{1+(2N-1)e^{-2\lambda E}-2r(1-e^{-\lambda L}+N)}{(1-e^{-\lambda L})(1-r)+N(p-r)} - 1}.$$

Proof.

(i) See Methods (Section 4) for the derivation of $\tau(t)$.

(ii) A clone starting at t begins a scaffold contig with probability $p(t)-\tau(t)$ by definition of scaffolds and contigs. So the number of scaffold contigs starting in $(0, G]$, N_c, is a Poisson random

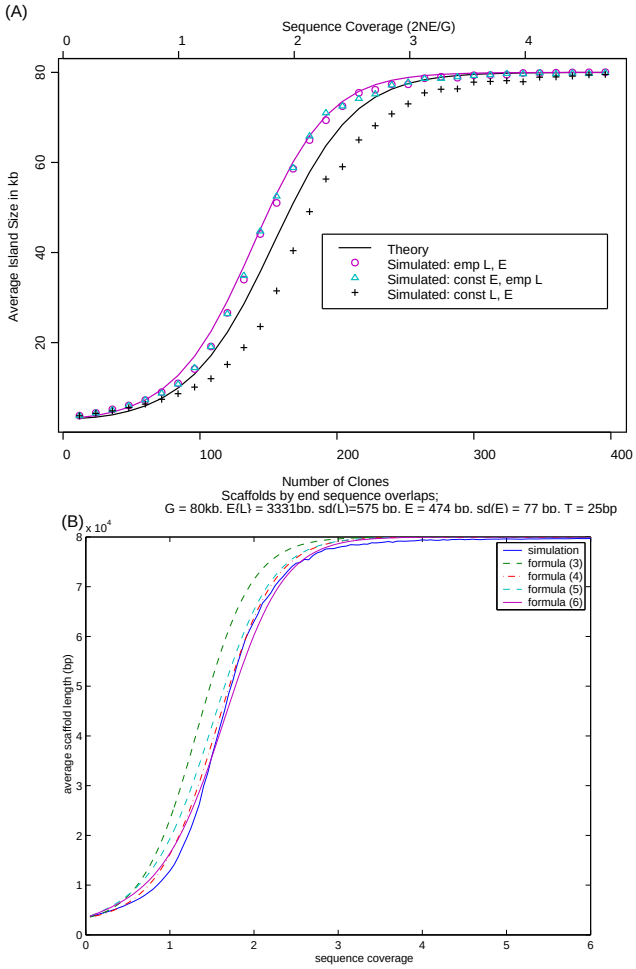

Figure 4: Average length of a scaffold. (A) Theoretical curve based on formula (3) for p_e. The parameters are the same as Figure 3(A). (B) Comparison of formula (3)-(6) for p_e using parameters as in Figure 3(B).

variable with mean $N(p(t) - \tau(t))$. With a similar edge-effect correction to counting scaffolds, the number of scaffold contigs $= \mathbf{1}(0 \text{ is covered by a non-singleton clone}) + N_0$, and

$$\mathbf{E}\{\text{No. of scaffold contigs}\} = (1 - e^{-\lambda L})(1 - \tau) + N(p - \tau).$$

(iii) Following the derivation of the expected length of a scaffold, and arguing that scaffold contigs are effective scaffolds, we have

$$\begin{aligned} & \mathbf{E}\{\text{length of scaffold contigs}\} \\ \approx & \frac{\mathbf{E}(\text{total scaffold coverage} - \text{total coverage by actual singletons})}{\mathbf{E}\{\text{No. of scaffold contigs}\}} \\ = & \frac{G(1 - e^{-\lambda L}) - Le^{-2\lambda L}}{(1 - e^{-\lambda L})(1 - \tau) + N(p - \tau)}. \end{aligned}$$

(iv) By the definition of apparent SMGs and scaffold contigs,

$$\begin{aligned} & \mathbf{E}\{\text{No. of apparent SMGs per scaffold contig}\} \\ = & \frac{\mathbf{E}\{\text{No. of sequence islands} - \text{No. of singleton ends in } (0, G)\}}{\mathbf{E}\{\text{No. of scaffold contigs}\}} - 1 \\ = & \frac{1 + (2N - 1)e^{-2\lambda L} - \tau(1 - e^{-\lambda L} + 2N)}{(1 - e^{-\lambda L})(1 - \tau) + N(p - \tau)} - 1 \end{aligned}$$

(v) Following the argument for the expected size of apparent SMG in a scaffold,

$$\begin{aligned} & \mathbf{E}\{\text{apparent SMG size in a scaffold contig}\} \\ = & \frac{\mathbf{E}\{\text{scaffold contig length}\} - \frac{\Sigma\{\text{sequence island coverage in scaffold contigs}\}}{\Sigma\{\text{No. of scaffold contigs}\}}}{\mathbf{E}\{\text{No. of apparent SMGs in a scaffold contig}\}} \\ = & \frac{\frac{G(1 - e^{-\lambda L}) - Le^{-2\lambda L}}{N(p_+ - e^{-2\lambda L})} - \frac{G(1 - e^{-2\lambda L}) - 2Er(1 - e^{-\lambda L} \mid N)}{(1 - e^{-\lambda L})(1 - r) \mid N(p - r)}}{\frac{1 \mid (2N - 1)e^{-2\lambda L} - 2r(1 - e^{-\lambda L} \mid N)}{(1 - e^{-\lambda L})(1 - r) \mid N(p - r)} - 1}. \end{aligned}$$

3 Discussion

In this paper, we presented a Lander-Waterman type analysis for various measures of sequencing/mapping project progress utilizing a

paired-end sequencing scheme. The theoretical predictions yield satisfactory results when compared to simulations using parameters similar to that were used in the Berkeley Drosophila Genome Project and Celera's Human Genome Project.

The analysis and simulations reveal an interesting feature of the paired-end sequencing scheme: in terms of mean values of various project progress measures, varying end-sequence length ($\sigma_E > 0$) has little effect on the progress of scaffolds (compared to constant length case $\sigma_E = 0$ with same E), whereas variable clone length significantly speeds up project progress ($\sigma_L > 0$ compared to $\sigma_L = 0$ with same L) even though it makes no difference for forming actual islands. This is readily explained, since end-sequence length contributes to scaffolds in the form of sequence islands, whose mean number does not change by the variation of end sequence length. In contrast, clone length is the source of paired-end dependency, and varying clone length indeed increases the possibility of clone overlaps at their ends to some degree. It can be seen from the closed-form approximation of α and β that the contribution of σ_L^2 is bounded by the constant case, and hence $\sigma_L^2 > 0$ accelerates the growth of scaffolds.

How about increasing L or E? What is the best choice of L and E for a project of size G? Longer E is obviously beneficial since it results in more sequence coverage and larger chance for end overlaps, but E is usually limited by the sequencing technology. Currently $E \approx 500 - 600$ bp for most large-scale sequencing operations using automated sequencers. Thus it reduces to the selection of L/E. Given fixed G, E, N, σ_L^2, longer L yields higher clone redundancy and less actual islands, but the chance of clone overlap at the ends remains unchanged if $G \gg L > 3E$, implying no effect on scaffolds. This phenomenon can be easily seen from the absence of L in $\tau(t)$. However, all project progress measures are dependent on L, when effective genome size $G_e = G - L$ is used in place of G as finite correction for moderate L/G. So larger L accelerates project progress as a result of increasing clone intensity and a smaller region for scaffold inclusion.

In practice, the starting material for sequencing is usually a circular clone formed by vector of size V and insert DNA of size G (the real sequencing target). For instance, $G \approx 80$ kb with $V = 16$ kb for a P1 clone, and $G \approx 150$ kb with $V \approx 8$ kb for a BAC clone. So the subclone process intensity is actually $\frac{N}{G+V}$ and the linearized clone has target insert $(0, G)$ and vector at both sides $(-\frac{V}{2}, 0) \cup (G, G + \frac{V}{2})$. Typically subclones containing only vector sequence are screened

out and excluded from contig assembly. Those vector-insert hybrid subclones provide *anchoring* information for mapping clones. The *real* project progress should involve subclones containing at least some insert DNA (clones starting in $(-L, G)$), and scaffolds start in $(0, G - L)$ for recovery of the sequencing target. With this modification, larger L indeed yields faster project progress as a result of more subclones containing insert DNA, and a shorter region for scaffold inclusion. This modification including the vector size also gives a more realistic prediction when applying the theoretical analysis to the real case. In addition, a longer subclone library usually has a larger σ_L^2 which also helps the growth of scaffolds somewhat.

4 Methods

Proposition 1 (ii): Derivation of $p(t)$ for $L > 3E$ with negligible clone subsumption. The event that L_i starts a scaffold can be decomposed into the following two disjoint events:

$[A_1]$: t is uncovered by any clones and L_i begins an actual island. This event has probability $e^{-\lambda L}$.

$[A_2]$: t is internal to some clone L_p, and L_p is unlinked to L_i via their ends.

In this case, t is covered only by clones which *interlace* or *encompass* L_i without end overlaps. In other words, clone L_p interlaces L_i when $s < s + E_{pL} < t < t + E_{iL} < s + L_p - E_{pR} < s + L_p < t + L_i - E_{iR}$ so they overlap but their ends do not (only happens with $L_i > E_{iL} + E_{iR} + E_{pR}$ and $L_p > E_{pL} + E_{pR} + E_{iL}$). Clone L_p encompasses L_i when $s + E_{pL} < t < t + L_i < s + L_p - E_{pR}$ so L_i is subsumed in the unsequenced portion of L_p.

Event A_2 summarizes the difference between actual islands and scaffolds by paired-end sequencing, and can be further broken down into the following events:

$$A_2 = B \cap ((C_1 \cap D_1 \cap D_2) \cup (C_2 \cap D_1 \cap D_2))$$

$[B]$: no clone starting before t overlaps L_i at the ends; and

$[C_1]$: L_i interlaces some clone L_p, $s < t$; or

$[C_2]$: L_i is subsumed under some clone L_p, $s < t$.

$[D_1]$: given B and C_1, no end-sequences link E_{pR} to E_{iL} and/or E_{iR} for an interlacing clone L_p;

$[D_2]$: given B, C_1, D_1, no other complex links exist between E_{iL}, E_{iR} and E_{sL}, E_{sR}.

$[E_1]$: given B and C_2, no end-sequences link E_{sR} to E_{iR} and/or E_{sL} to E_{iL}.

$[E_2]$: given B, C_2, E_1, no other complex links exist between E_{sL}, E_{sR} and E_{iL}, E_{iR}.

$\Pr\{D_2\}$ and $\Pr\{E_2\}$ both ≈ 1, since given no direct bridge end-sequences for two interlacing or subsuming clones L_s and L_i, having more complex links would require many clones of very restrictive range of length and location with small probability. So

$$\begin{aligned}\Pr\{A_2\} &= \Pr\{B, C_1\}\Pr\{D_1\}\Pr\{D_2\} + \Pr\{B, C_2\}\Pr\{E_1\}\Pr\{E_2\} \\ &\approx \Pr\{B, C_1\}\Pr\{D_1\} + \Pr\{B, C_2\}\Pr\{E_1\}\end{aligned}$$

In addition, for length distribution $F_L(\cdot)$ with small variance σ_L^2, clone subsumption is an unlikely event, and it suffices to calculate $\Pr\{B\}$ and $\Pr\{D_1\}$ for $\Pr\{A_2\} \approx (\Pr\{B\} - \Pr\{A_1\})\Pr\{D_1\}$.

A clone L_s starting at s overlaps L_i at the ends whenever

$[B_{RL}]$: E_{sR} overlaps E_{iL}: $t < s + L_s$ and $s + L_s - E_{sR} < t + E_{iL}$;

$[B_{LL}]$: E_{sL} overlaps E_{iL}: $t < s + E_{sL}$ and $s < t + E_{iL}$;

$[B_{RR}]$: E_{sR} overlaps E_{iR}: $t + L_i - E_{iR} < s + L_s$ and $s + L_s - E_{sR} < t + L_i$;

$[B_{LR}]$: E_{sL} overlaps E_{iR}: $s < t + L_i$ and $t + L_i - E_{iR} < s + E_{sL}$.

Note that B_{RL}, B_{LR}, and $B_{LL} \cup B_{RR}$ are disjoint events for $L > 3E$. So the above has probability

$$\begin{aligned}q(s) &= \Pr\{B_{RL} \cup B_{LL} \cup B_{RR} \cup B_{LR}\} \\ &= \Pr\{B_{RL}\} + \Pr\{B_{LR}\} + \Pr\{B_{LL}\} + \Pr\{B_{RR} \backslash B_{LL}\} \\ &= \Pr\{t - s < L_s < t - s + E_{iL} + E_{sR}\} + \Pr\{s - t < L_i < s - t + E_{sL} + E_{iR}\} \\ &\quad + \Pr\{t - s < E_{sL}, s - t < E_{iL}\} \\ &\quad + \Pr\{t - s > E_{sL} \text{ or } s - t > E_{iL}; t - s - E_{iR} < L_s - L_i < t - s + E_{sR}\} \\ &= \mathbf{E}[F_L(t - s + E_{iL} + E_{sR}) - F_L(t - s)] + \mathbf{E}[F_L(s - t + E_{iR} + E_{sL}) - F_L(s - t)] \\ &\quad + 1 - F_E(|t - s|) + \int F_E(|t - s|)(1 - F_E(|t - s - d|))\, dH(d)\end{aligned}$$

Now we can apply thinning arguments to derive $\Pr\{B\}$: thin the original process by removing clones that do not cover t and those

do not overlap L_i at the ends. So the thinned process has intensity $\mu(s) = \lambda q(s)\ \mathbf{1}(s < t)$, and

$$\begin{aligned}
\Pr\{B\} &= \Pr\{\text{no events in } (-\infty, t) \text{ in the thinned process}\} \\
&= \exp\{-\int_{-\infty}^{t} \mu(s)ds\} \\
&= \exp\{-\lambda(\mathbf{E}[\int_0^{\infty} (1 - F_L(v)) - (1 - F_L(v + E_{iL} + E_{iR}))\, dv] \\
&\quad + \int_0^{\infty} (1 - F_E(v))\, dv + \int\int_0^{\infty} F_E(v)(1 - F_E(|v - d|))\, dv\, dH(d)\}\} \\
&= \exp\{-\lambda(\mathbf{E}[\int_0^{E_{iL} + E_{iR}} (1 - F_L(v))\, dv] + E \\
&\qquad + \int\int_0^{\infty} F_E(v)(1 - F_E(|v - d|))\, dv\, dH(d)\}\} \\
&= \exp\{-\lambda(3E + \int\int_0^{\infty} F_E(v)(1 - F_E(|v - d|))\, dv\, dH(d)\}\} \\
&= \exp\{-(3 + \alpha)\lambda E\} \\
\text{where } \alpha &\equiv \frac{1}{E}\int\int_0^{\infty} F_E(v)(1 - F_E(|v - d|))\, dv\, dH(d).
\end{aligned}$$

The remaining component is $\Pr\{D_1\}$. $L_p (s < t)$ interlaces L_i if

$$s + E_{pL} < t < t + E_{iL} < s + L_p - E_{pR} < s + L_p < t + L_i - E_{iR}.$$

D_1 is a complicated event that no remaining end-sequences link E_{pR} to E_{iL} and/or E_{iR} for an interlacing clone L_p. Instead of trying to analyze the clone layouts that satisfy D_1, we approximate D_1 by a simpler necessary event $D_1' \cup D_2'$:

$[D_1']$: no remaining *left* end-sequences overlap E_{pR} - no clones start in $(s + L_p - E_{pR} - E_j, s + L_p)$.

$[D_2']$: no *left* end-sequences starting in $(t, t + L_i)$ overlap E_{iL} nor E_{iR} - no clones start in $(t, t + E_{iL}) \cup (t + L_i - E_{iR} - E_j, t + L_i - E_{iR})$.

The condition is restricted to the left end-sequences of remaining clones since B guarantees no right end-sequences can possibly bridge the ends of L_p and L_i. Let $Y(\cdot|t)$ denote the conditional distribution of $v = t - s$ given that L_p interlaces L_i. $Y(\cdot|t) \sim \text{Uniform}(max(E_{pL}, L_p - L_i + E_{iR}), L_p - E_{pR} - E_{iL})$. With similar thinning arguments,

$$\Pr\{D_1'\} = e^{-\lambda\mathbf{E}[\int_{-\infty}^{t+L_p} (1 - F_E(t + L_p - \xi - E_{iR}))d\xi]} = e^{-\lambda\mathbf{E}[\int_{-E_{iR}}^{\infty} (1 - F_E(u))\, du]} = e^{-2\lambda E},$$

$$\begin{aligned}
\Pr\{D_2'\} &= e^{-\lambda(\int_t^{\infty}(1-F_E(\xi-t))\,d\xi+\int_{-\infty}^{t+\Delta_L-\Delta_E}(1-F_E(t+L_t-E_{tR}-\xi))\,d\xi)} = e^{-2\lambda E},\\
\Pr\{D_1' \cap D_2'\} &= \exp\{-\lambda(4E-\mathbf{E}[\int(\int_{-\infty}^{s+L_s}(1-F_E(s+L_s-\xi-E_{sR}))\,d\xi \\
&\qquad + \int_{-\infty}^{s+L_s}(1-F_E(t+L_t-E_{tR}-\xi))\,d\xi)\,dY(t-s)])\} \\
&= \exp\{-4\lambda E+\lambda\mathbf{E}[\int(\int_{u+L_s-E_{sR}-E_{tL}}^{\infty}(1-F_E(v))\,dv \\
&\qquad + \int_{u-\Delta_L-E_{tR}}^{\infty}(1-F_E(v))\,dv)\,dY(u)]\} \\
&= \exp\{-4\lambda E+\lambda\mathbf{E}[\frac{dY(u)}{du}](2\int v(1-F_E(v))\,dv)\} \\
&= \exp\{-4\lambda E+\lambda(E^2+\sigma_E^2)\mathbf{E}[\frac{dY(u)}{du}]\}.
\end{aligned}$$

Use first moment approximation.

$$\begin{aligned}
\mathbf{E}[\frac{dY(u)}{du}] &= \mathbf{E}[\{L_s-E_{sR}-E_{sL}-E_{tL}+(\Delta_L-E_{sR}+E_{tL})\mathbf{1}_{\Delta_L>E_{sR}-E_{tL}}\}^{-1}] \\
&\approx \{L-3E-\mathbf{E}[(\Delta_L-\Delta_E)\mathbf{1}_{\Delta_L>\Delta_E}]\}^{-1} \\
&= \{L-3E-\frac{1}{2}\mathbf{E}|\Delta_L-\Delta_E|\}^{-1} \text{ since } \Delta_L-\Delta_E \text{ is symmetric.}
\end{aligned}$$

$$\begin{aligned}
\text{So } \Pr\{D_1\} &\approx e^{-4\lambda E}(2e^{2\lambda E}-e^{\frac{\lambda(E^2+\sigma_E^2)}{L-3E-\frac{1}{2}\mathbf{E}|\Delta_L-\Delta_E|}}) = e^{-2\lambda E}(2-e^{\beta\lambda E}) \\
\text{where } \beta &\equiv 2-(E+\frac{\sigma_E^2}{E})/(L-3E-\frac{1}{2}\mathbf{E}|\Delta_L-\Delta_E|).
\end{aligned}$$

Putting the components together, we have an approximation (formula 1) for $p(t)$, the thinning probability of scaffolds by end overlaps:

$$p(t) \approx e^{-\lambda L}+(e^{(2+\alpha)\lambda E}-e^{-\lambda L})\cdot e^{-2\lambda E}(2-e^{\beta\lambda E})$$

Closed form approximation for α and β. Rewrite αE as follows:

$$\begin{aligned}
\alpha E &= \int\int_0^{\infty} F_E(u)(1-F_E(u-d))\,du\,dH(d) \\
&= \mathbf{E}\int_0^{\infty}\Pr\{E_i<u, E_j>|u-\Delta_L|\}\,du \qquad \text{for some } E_i, E_j \text{ i.i.d.} \sim F_E
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\int_0^\infty \Pr\{max(E_i, \Delta_L - E_j) < u < \Delta_L + E_j\}\, du \\
&= \mathbf{E}\{\mathbf{1}_{\Delta_L \mid E_j,\ E_i>0}(\Delta_L + E_j - E_i) - \mathbf{1}_{\Delta_L\ E_j,\ E_i>0}(\Delta_L - E_j - E_i)\} \\
&= \frac{1}{2}\mathbf{E}|\Delta_L - \Delta_E| - \mathbf{E}[(\Delta_L - E_j - E_i)\mathbf{1}_{\Delta_L\ E_j,\ E_i>0}].
\end{aligned}
$$

Notice that

$$
(\mathbf{E}|\Delta_L - \Delta_E|)^2 + \mathrm{Var}(|\Delta_L - \Delta_E|) = \mathbf{E}(|\Delta_L - \Delta_E|^2) = \mathrm{Var}(\Delta_L - \Delta_E) = 2\sigma_L^2 + 2\sigma_E^2.
$$

So $\sqrt{2(\sigma_L^2 + \sigma_E^2)}$ is an upper bound for both $\mathbf{E}|\Delta_L - \Delta_E|$ and $2\alpha E$.
In addition, $0 \le \alpha \le 2$ and $0 \le \beta \le 2$ since

$$
0 \le \alpha E = \int_0^\infty \Pr\{B_{RR} \backslash B_{LL}\}\, du \le \int_0^\infty \Pr\{B_{RR}\}\, du = 2E;
$$

$$
2E = \int_0^\infty \Pr\{B_1\} du \le (2+\beta)E = \int_0^\infty \Pr\{B_1 \cup B_2\} du \le \int_0^\infty (\Pr\{B_1\} + \Pr\{B_2\}) du = 4E.
$$

Therefore

$$
\begin{aligned}
\alpha &\approx min(2, \frac{\sqrt{(\sigma_L^2 + \sigma_E^2)/2}}{E}); \\
\beta &\approx min(2, max(0, 2 - \frac{E + \sigma_E^2/E}{L - 3E - \sqrt{(\sigma_L^2 + \sigma_E^2)/2}})).
\end{aligned}
$$

Proposition 1 (iii): Derivation of $p(t)$ for $L > 3E$ with clone subsumptions. With the contribution of clone subsumption,

$$
\begin{aligned}
\Pr\{A_2\} &= \Pr\{B, C_1\}\Pr\{D_1\} + \Pr\{B, C_2\}\Pr\{E_1\} \\
&= p_1 \cdot (\Pr\{B\} - \Pr\{A_1\}) \cdot \Pr\{D_1\} + (1 - p_1) \cdot (\Pr\{B\} - \Pr\{A_1\}) \cdot \Pr\{E_1\}
\end{aligned}
$$

where p_1 is defined as the conditional probability that some clone interlaces L_i given t is covered by some clone not overlapping L_i at the ends (that either interlaces or encompasses L_i). Note that $L_s (s < t)$ interlaces L_i when $s \in I_1 = \{s : s + E_{sL} < t < t + E_{iL} < s + L_s - E_{sR} < s + L_s < t + L_i - E_{iR}\}$, and $L_s (s < t)$ encompasses L_i when $s \in I_2 = \{s + E_{sL} < t < t + L_i < s + L_s - E_{sR}\}$, so

$$
p_1 \equiv \Pr\{B \cap C_1 | B \cap (C_1 \cup C_2)\}
$$

$$= \frac{\int_{-\infty}^{t} \Pr\{s \in I_1\}\,ds}{\int_{-\infty}^{t} \Pr\{s \in I_1 \text{ or } I_2\}\,ds}$$

$$= \frac{\mathbf{E}\int_0^{\infty} F_E(v)[F_L(v+L_i-E_{ir}) - F_L(v+E_{pr}+E_{qj})]\,dv}{\mathbf{E}\int_0^{\infty} F_E(v)[F_L(v+L_i-E_{ir}) - F_L(v+E_{pr}+E_{qj}) + 1 - F_L(v+L_i+E_{pr})]\,dv}$$

In addition, following the argument of the calculation of $\Pr\{D_1\}$, approximate E_1 by $E_1' \cup E_2'$:

$[E_1']$: E_{iR} can't be extended by left end of other clones;

$[E_2']$: E_{pR} can't be extended by left end of other clones.

$$\begin{aligned}\Pr\{E_1\} &= \Pr\{E_1'\} + \Pr\{E_2'\} - \Pr\{E_1' \cap E_2'\} \\ &= e^{-\beta\lambda E} + e^{-2\lambda E} - e^{-2\lambda E \mid \int_{x+L_i}^{x+L_i} \int_{-\infty}^{t} F_E(s \mid L_i,\ E_{ir},\ x)\,dx\,dY(i,\ s)]}.\end{aligned}$$

Taking all the components together, we have Formula (2)

$$\begin{aligned}p(t) &= \text{Formula (1)} + (1-p_1)(\Pr\{B\} - \Pr\{A_1\})(\Pr\{E_1\} - \Pr\{D_1\}) \\ &= \text{Formula (1)} + (1-p_1)(e^{-(2\,|\,\alpha)\lambda E} - e^{-\lambda L}) \\ &\quad \cdot (e^{-\lambda E} - e^{-2\lambda E} + e^{-(2\,|\,\beta)\lambda E} - e^{-2\lambda E \mid \int_{x+L_i}^{x+L_i} \int_{-\infty}^{t} F_E(s \mid L_i,\ E_{ir},\ x)\,dx\,dY(i,\ s)]})\end{aligned}$$

Theorem 1 (iii): Derivation of $p_c(t)$. By the definition of an effective scaffold, only one extra condition is needed for specifying an effective scaffold S_i starting at t in addition to above events A_1 and A_2 for a scaffold: when another scaffold S_p starting at $s < t$ interlaces S_i, S_p must terminate before S_i. In other words, the ending clone of S_p has its right end-sequence E_{pR} falling in between the ends of some clone in S_i, and no remaining left end-sequences extends E_{pR}. This just adds a factor of $e^{-2\lambda E}$ to the probability of event A_2, hence we have Formula (3):

$$\begin{aligned}p_c(t) &= \Pr\{\text{an effective scaffold starts at } t\} \\ &\approx e^{-\lambda L} + e^{-\gamma\lambda E}(e^{-\alpha\lambda E} - e^{-\lambda(L-2E)})(2 - e^{-\beta\lambda E})\end{aligned}$$

Alternative approximations (Formula (4)-(6)) for p_c can also be used. Formula (4) is a simple modification to Formula (3), and just uses $\Pr\{D_1\} \approx e^{-2\lambda E}$ for the fact that L_i can not be extended to the right for effective scaffolds S_i. Formula (5) uses a modified $\Pr\{D_1\} \approx \Pr\{D_1'\} - \Pr\{D_1' \cap D_2'\}$ to correct for double counting those singletons L_i. Formula (6) uses the additional factor $2e^{-\lambda E} - e^{-(2\,|\,\beta)\lambda E} - e^{-(1\,|\,\alpha)\lambda E}$ instead to compensate for the fact that scaffold S_i can not be extended to the right but clone L_i can, ignoring those singletons L_i.

Theorem 2 (1): Derivation of $\tau(t)$. Consider the thinned process in the derivation of $\Pr\{B\}$. Then

$$\begin{aligned}\tau(t) &= \Pr\{\text{no events in } (-\infty,\infty) \text{ in the thinned process}\} \\ &= \exp\{-\int \mu(s)ds\} \\ &= \exp\{2\log(\Pr\{B\})\} = e^{(6\,|\,2\alpha)\lambda E}.\end{aligned}$$

5 Acknowledgment

RFY and TPS would like to thank the Berkeley Drosophila Genome Project for originating this work, especially B. Kimmel and S. Celniker. RFY was partially supported by Burrows Wellcome Foundation-Program in Mathematics and Molecular Biology Fellowship for this work. XL and MSW would like to thank Granger Sutton and Gene Myers for suggesting this work. XL is grateful to the University of Southern California for University Fellowship support, and both XL and MSW were partially supported by NIH grant 53-4955, and by Celera Genomics.

Ru-Fang Yeh
Center for Bioinformatics and Molecular Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560

Terence P Speed
Department of Statistics, University of California, Berkeley, CA 94720-3860
Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3050, Australia

Michael S Waterman
Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089
Celera Genomics, Rockville, MD 20878

Xiaoman Li
Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113
Celera Genomics, Rockville, MD 20878

References

[1] R. Arratia, E.S. Lander, S. Tavare, M.S. Waterman. Genomic mapping by anchoring random clones: a mathematical analysis. *Genomics*. **11(4)**: 806-27. 1991.

[2] A. Edwards and C. T. Caskey. Closure strategies for random DNA sequencing. *Methods: Companion Methods Enzymol.* 3(1): 41-47. 1991.

[3] Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, Erfle H, Caskey CT, Ansorge W. Automated DNA sequencing of the human HPRT locus. *Genomics*. **6(4)**:593-608. 1990.

[4] P. Hall. Introduction to the Theory of Coverage Processes. John Wiley and Sons. 1988.

[5] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2: 231-239. 1988.

[6] Port, E., Sun, F., Martin, D., and Waterman, M.S. (1994). Genomic mapping by end-characterized Random clones: A mathematical analysis. *Genomics* 26 : 84-100.

[7] J. Roach. Random Subcloning. *Genome Research* 5: 464-473. 1995.

[8] J. Roach, C. Boysen, K. Wang, and L. Hood. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics* 26: 345-353. 1995.

[9] J.C. Venter *et al.* The sequence of human genome. *Science* **291(16)**: 1304-1351. 2001.

[10] J.L. Weber and E.W. Myers. Human whole-genome shotgun sequencing. *Genome Res.* **7(5)**:401-9. 1997.