Pages 1–25 (Preprint posted electronically on January 6, 2003)

CLONE-ARRAY POOLED SHOTGUN MAPPING AND SEQUENCING: DESIGN AND ANALYSIS OF EXPERIMENTS

MIKLÓS CSŰRÖS AND ALEKSANDAR MILOSAVLJEVIC

Abstract. This paper studies methods for sequencing and mapping that rely solely on BAC pooling and shotgun sequencing. First, we scrutinize and improve the recently proposed Clone-Array Pooled Shotgun Sequencing (CAPSS) method, which delivers a clone-linked assembly of a whole genome sequence. Secondly, we introduce a novel physical mapping method, called *Clone-Array Pooled Shotgun Mapping* (CAPS-MAP), which computes the physical ordering of BACs in a random library. Both CAPSS and CAPS-MAP are based on constructing subclone libraries from pooled genomic clones.

After pointing out some shortcomings of the original CAPSS proposal, we propose algorithmic and experimental improvements that make CAPSS a viable option for sequencing a set of BACs. We provide the first probabilistic model of CAPSS sequencing progress. The model leads to theoretical results supporting previous, less formal arguments on the practicality of CAPSS. We demonstrate the usefulness of CAPS-MAP for clone overlap detection with a probabilistic analysis. Our analysis indicates that CAPS-MAP is well-suited for detecting BAC overlaps in a highly redundant library, relying on a low amount of shotgun sequence information. Consequently, it is a practical method for computing the physical ordering of clones in a random library, without requiring additional clone fingerprinting. Since CAPS-MAP requires only random shotgun sequence reads, it can be seamlessly incorporated into a sequencing project with almost no experimental overhead.

Keywords: sequencing, physical mapping, pooled shotgun sequencing.

1. Introduction

A new BAC-based sequencing strategy, called Clone-Array Pooled Shotgun Sequencing (CAPSS), was proposed recently (Cai et al. 2001). CAPSS assembles the complete sequences of individual BACs using a

small number of shotgun libraries compared to clone-by-clone sequencing strategies. In a clone-by-clone approach, clones are sequenced independently: DNA is extracted from each clone and used for subclone library preparations. In this way one subclone library is constructed for every clone. The clone's initial sequence is assembled after collecting a sufficient number of sequence reads from the subclone library. In a CAPSS approach, DNA from clones are pooled together, and subclone libraries are prepared from the pools. A CAPSS experiment is designed so that the number of subclone libraries is much smaller than the number of clones, yet the pooling design enables the assembly of individual clone sequences. In what follows, we refer to random shotgun sequence reads collected from a subclone library that was constructed using pooled BACs as *pooled shotgun fragments*. For the computational aspects of sequence assembly, pooled shotgun fragments are random subsequences originating from a set of clone sequences.

The CAPSS proposal (Cai et al. 2001) relies on a simple rectangular pooling design defined by an array layout of BACs (Figure 1). The pools correspond to the rows and columns. An array layout reduces the number of shotgun library preparations to the square root of the number of BACs when compared to clone-by-clone sequencing. This reduction can be important given the fact that for a mammalian genome, even a minimally overlapping tiling path contains between twenty and thirty thousand clones (IHGSC 2001).

This paper has two goals. First, after pointing out some shortcomings of the original CAPSS proposal, we propose algorithmic and experimental improvements that make CAPSS a viable option for sequencing a set of BACs. We provide the first probabilistic model of CAPSS sequencing progress. The model leads to theoretical results supporting previous, less formal arguments on the practicality of CAPSS. The paper's second goal is to introduce the *Clone-Array Pooled Shotgun Mapping (CAPS-MAP)* method to detect clone overlaps in a random BAC library. The information on clone overlaps is used to compute the physical ordering of clones in the library, without requiring additional clone fingerprinting. CAPS-MAP operates in the same experimental framework as CAPSS. It needs only shotgun sequence reads, which makes it a cost-effective method that can be seamlessly integrated into a sequencing project with very little experimental overhead. We demonstrate the usefulness of CAPS-MAP for clone overlap detection with a probabilistic analysis.

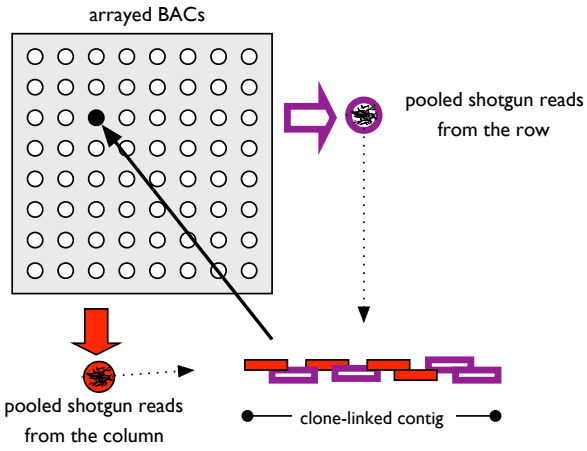


FIGURE 1. CAPSS strategy for arrayed BACs. DNA extracted from each clone is pooled together with other clones in the same row and column. Subclone libraries are prepared from the pools, and random reads are collected from the sublibraries. Reads are assembled into contigs. If a contig contains reads from a row and a column pool's sublibrary, the contig is assigned to the BAC at the intersection of the row and the column.

2. TRANSVERSAL DESIGNS

It was proposed by Cai et al. (2001) that CAPSS be used in hybrid projects, combining whole-genome shotgun (WGS) (Weber and Myers 1997) and pooled shotgun fragments. The idea is that the pooled shotgun fragments can provide the localization information for the whole-genome shotgun fragments so that the latter can be used for a clone-linked sequence assembly. After fragments are assembled into contigs, the contigs need to be mapped to individual BACs. There are at least two algorithmic approaches to assigning the contigs to BACs in a hybrid CAPSS project. The first one, which we call *clone-based contig mapping*, was suggested by Cai et al. (2001), and consists of assembling contigs for each BAC individually. For each BAC, WGS fragments are combined with the fragments from the pools that the BAC is included in, i.e., from the BAC's row and column pools in the original CAPSS design. Each assembled contig that contains fragments from two of the BAC's pools is assigned to the BAC. Another possibility, which we call *collective contig mapping*, is to combine all whole-genome and pooled shotgun fragments together, and to assign a BAC to each resulting contig according to the pools that the fragments originate from. In

collective contig mapping, contigs may be simultaneously mapped to more than one clone.

What are the difficulties in contig mapping? We mention here three main problems: false negatives, ambiguities, and false mapping. A false negative refers to a situation where a BAC is not sampled in a pool it is included in, due to the low number of pooled shotgun fragments. A false negative for a simple rectangular design means that no contigs can be mapped to the BAC. Ambiguities and false mappings are caused by overlapping clones, or more generally, by clones that share identical sequences. The mapping of a contig is ambiguous if it is not possible to decide which clones the contig should be assigned to, in cases where two or more clone sets are equally likely choices for the mapping. For instance, if a contig contains pooled fragments from two row pools and two column pools, then there are four BAC candidates for the mapping, at the four intersections of the rows and columns. If they are numbered by B_{11}, B_{12}, B_{22}, and B_{21} in a clockwise direction, then the clone pairs (B_{11}, B_{22}) and (B_{12}, B_{21}) are equally likely choices. False mapping occurs when an insufficient number of pooled shotgun sequence reads are collected, and a contig that covers overlapping BACs gets assigned to the wrong clone or clone set. With the example of the four BACs in two rows and two columns, a false negative would occur if a contig were to contain fragments from only one row pool and column pool, and if, accordingly, we were to conclude that the contig should be mapped to, say, clone B_{11}, when in reality it belonged to an overlap between B_{12} and B_{21}. False mapping is more detrimental than ambiguity since it is not detected during contig mapping. In a clone-based mapping approach, contigs that cover clone overlaps are shorter, and get falsely mapped more often. With the example of the four clones, contigs covering the overlap might be assigned to each of the four clones, whereas a collective mapping approach would classify them as ambiguous. In view of the more harmful effects of clone overlaps in the clone-based mapping approach, collective contig mapping is the preferable method.

One strategy used to overcome the problems of ambiguity, false negatives, and false mapping involves transversal pooling designs (Csűrös and Milosavljevic 2002; Du and Hwang 2000). Using a transversal double-array design, the same set of BACs is independently arrayed and pooled twice. Each of the two resulting arrays contains the same set of BACs, albeit in a different two-dimensional arrangement. Thus, each BAC ends up being sampled in four pools: two column-pools and two row-pools. One of the arrays contains an arbitrary arrangement of BACs, while the other is "reshuffled" relative to the first.

More generally, a transversal pooling design with $n = 2d$ pool sets can be used to arrange clones on d reshuffled arrays. For a transversal design with n pool sets, every clone is included in n pools, and any subset with at least 2 of those pools uniquely identifies the clone.

The number of arrays in a transversal design may be adjusted to allow unambiguous and correct contig mapping for any redundancy in a BAC library. Specifically, it can be shown (Csűrös and Milosavljevic 2002; Du and Hwang 2000) that a d-array transversal design can accurately resolve BACs at up to $(2d - 1)$X redundancy. We previously described and analyzed transversal designs in the context of pooled shotgun experiments (Csűrös and Milosavljevic 2002) and compared their performance to other designs. Even though our analysis was performed for the Pooled Genomic Indexing (PGI) method in the context of comparative physical mapping, the results are generally valid for CAPSS and CAPS-MAP as well. Specifically, our results indicate that transversal designs reduce the frequency of false negatives and false mappings when compared to a simple rectangular design. Furthermore, when compared to other more complicated designs, they achieve an optimal balance between the number of shotgun library preparations and the frequency of contig mapping problems. Transversal designs also enjoy a practical advantage over more complicated combinatorial designs, in that they are readily implemented using existing automated clone arraying technologies.

When a transversal design is used, collective contig mapping can be implemented very efficiently, based on an algorithm that runs in $O(N + M)$ time for mapping M contigs onto N BACs. Without going into details, the main idea of the algorithm is to first build in $O(N)$ time a hash table that maps pool pairs to BACs. Based on the property of transversal designs that two pools identify a clone, this table contains all pool pairs that identify a unique clone. For each contig, it takes $O(1)$ time using the hash table to either identify the most likely clone set to which the contig can be mapped, or to declare the contig ambiguous.

3. Pooled shotgun fragments for sequencing

In this section we analyze CAPSS sequencing progress in a hybrid project that uses whole-genome and pooled shotgun fragments. Pooled shotgun fragments are collected using a transversal design with n pool sets, i.e., $n/2$ arrays. In order to derive a probabilistic model for such experiments, we introduce the following notations along with some standard simplifying assumptions. Assume that every clone has the same

length L (100–200 thousand base pairs in practice), and that each shotgun fragment has the same length ℓ (500 bp in practice). Let a be the coverage by pooled shotgun fragments, i.e., if F_{p} pooled shotgun sequence reads are collected, then $a = \frac{F_{\mathrm{p}}\ell}{NL}$ where N is the total number of clones. Let w denote the coverage by whole genome shotgun fragments, i.e., if F_{w} WGS sequence reads are collected, then $w = \frac{F_{\mathrm{w}}\ell}{G}$ where G is the genome length. Notice that $w = 0$ is possible. The WGS and pooled fragments are combined and compared to each other to find overlaps between them. Overlapping fragments form *islands*. Islands with two or more fragments are called *contigs*. An overlap between two fragments is detected if it is at least of length $\vartheta\ell$ where $0 < \vartheta \leq 1$. Statistics for islands, contigs, and gaps between islands are well known (Lander and Waterman 1988; Wendl and Waterston 2002). We are interested in statistics for *clone-linked contigs*, those that are assigned to BACs using the pooling information.

Here we consider the simplest case of assembling the sequence of a single clone that does not overlap with any other clone. Such a clone is covered by a total coverage of $(a + w)$. Although we concentrate on sequencing a particular clone, the transversal design allows the sequencing of overlapping clone regions by combining fragments from many (or even all) pools in a collective contig mapping method. Regions of overlapping clones have higher coverage since they are covered by more pooled fragments than a single clone. The sequencing of overlapping regions progresses thus faster than what is suggested by the statistics for a single clone. We examine the case of assigning contigs to overlapping BACs in §4. Two fragments from different pools suffice to assign a contig to a single BAC. In a practical setting, it may be advantageous to require more stringent criteria in order to avoid false mappings. The following theorem can be readily adapted for such criteria, albeit resulting in bulkier formulas.

Theorem 1. *Let $\sigma = 1 - \vartheta$ where ϑ is the fraction of length two fragments must share in order for the overlap to be detected. Consider a BAC that does not overlap with other clones. Define $c = w + a$, the total coverage. Let $X_1 = \frac{w+\frac{a}{n}}{c}$, $X_2 = \frac{w}{c}$, and $Y_i = 1 - (1 - e^{-c\sigma})X_i$ for $i = 1, 2$.*

(i) *The expected number of clone-linked contigs covering the clone equals*

$$\frac{L}{\ell} c e^{-c\sigma} p_{\mathrm{link}}, \tag{1}$$

where

$$p_{\text{link}} = \begin{cases} 1 - e^{-c\sigma}\left(n\frac{X_1}{Y_1} - (n-1)\frac{X_2}{Y_2}\right) & \text{if } w > 0; \\ \frac{1-e^{-a\sigma}}{1+\frac{1}{n-1}e^{-a\sigma}} & \text{if } w = 0. \end{cases} \tag{2}$$

(ii) *The expected number of fragments in a clone-linked contig is*

$$F_{\text{link}} = \begin{cases} \frac{e^{c\sigma}}{p_{\text{link}}}\left(1 - e^{-2c\sigma}\left(n\frac{X_1}{Y_1^2} - (n-1)\frac{X_2}{Y_2^2}\right)\right) & \text{if } w > 0; \\ e^{a\sigma} + \frac{1-\frac{1}{n-1}}{1+\frac{e^{-a\sigma}}{n-1}} & \text{if } w = 0. \end{cases} \tag{3}$$

(iii) *Define*

$$F_{\text{nolink}} = \frac{n\frac{X_1}{Y_1^2} - (n-1)\frac{X_2}{Y_2^2}}{n\frac{X_1}{Y_1} - (n-1)\frac{X_2}{Y_2}}, \tag{4}$$

and

$$\lambda_{\text{CBC}} = \frac{e^{c\sigma} - 1}{c} + \vartheta. \tag{5}$$

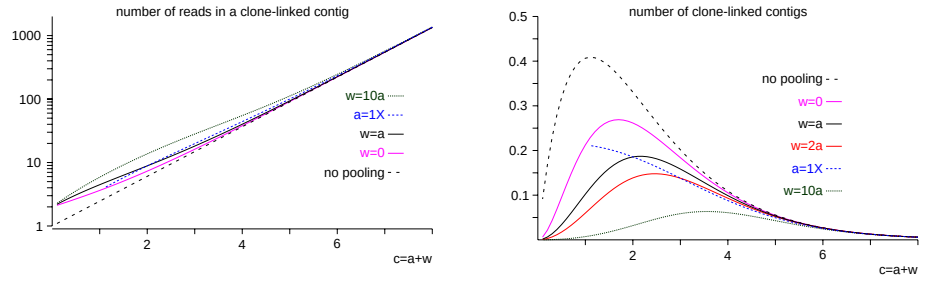
The expected length of a clone-linked contig can be written as $\ell\lambda_{\text{link}}$ *where* λ_{link} *is bounded as*

$$\frac{\lambda_{\text{CBC}} - \left(F_{\text{nolink}}\sigma + \vartheta\right)(1 - p_{\text{link}})}{p_{\text{link}}} \leq \lambda_{\text{link}} \leq \lambda_{\text{CBC}}. \tag{6}$$

Furthermore, when $\mu = a/c$ *is kept constant,* F_{nolink} *increases monotonically with* c *and*

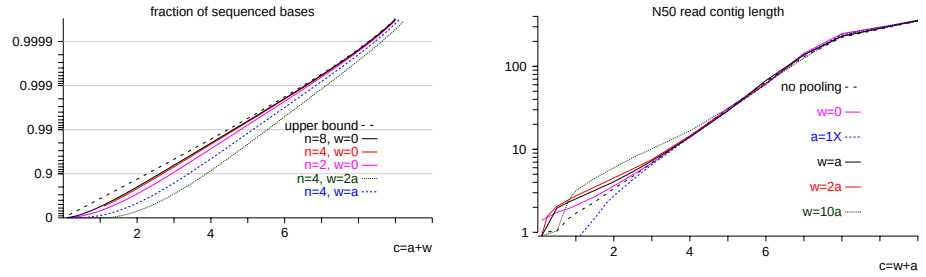
$$\lim_{c\to\infty} F_{\text{nolink}} = \begin{cases} \mu^{-1}\frac{(3n^2-3n+1)-\mu(2n^2-3n+1)}{(2n^2-3n+1)-\mu(n^2-2n+1)} & \text{if } w > 0; \\ \frac{n}{n-1} & \text{if } w = 0. \end{cases} \tag{7}$$

Proof. The proof relies on a Poisson process model, following the technique of Waterman (1995). We model the location of the shotgun fragments as a Poisson process with rate c. Define $\mu = a/c$, the fraction of pooled shotgun fragments. Every fragment is either a WGS fragment with probability $(1 - \mu)$, or comes from each one of the clone's pools with probability μ/n. First we state the well-known facts (Lander and Waterman 1988; Waterman 1995) about apparent islands, whether or not they are linked to a clone. The event E that a given fragment is the right-hand end of an apparent island has probability $J = \mathbb{P}E = e^{-c\sigma}$.



a b

FIGURE 2. CAPSS (Theorem 1): clone-linked contig statistics. The values are calculated from Theorem 1 for two-array transversal designs and different pooled coverage levels a. Overlaps between fragments are detected with $\vartheta = 0.1$. The number of contigs on the right-hand side is given in multiples of L/ℓ. The abscissa is the total coverage c.



a b

FIGURE 3. CAPSS (Theorem 1): sequencing progress. The left-hand side plots the fraction of bases covered by clone-linked contigs as a function of total coverage ($c = a + w$) for different designs. Notice that the improvement from two arrays to four arrays ($n = 4$ vs. $n = 8$) is marginal. The right-hand side plots the N50 values for different designs with two arrays, as multiples of ℓ. All values were calculated with fragment overlap detection $\vartheta = 0.1$. The N50 plot was obtained from simulation: each point is an average of 200 measurements.

For the k-th read, define M_k as the number of fragments from its right-hand end until the first gap towards the left. The probability that an island has j fragments in it equals

$$\mathbb{P}\Big\{ M_k = j \;\Big|\; E \Big\} = (1 - J)^{j-1} J.$$

An island can be mapped to a clone if it contains fragments from at least two pools. The probability of mapping the island ending at the k-th read (event D_k) depends on the number of fragments in the island. Using inclusion-exclusion:

$$\begin{aligned}(8)\quad &\mathbb{P}\Big\{D_k \;\Big|\; M_k = j\Big\} \\ &= 1 - \sum_{\text{pools}} \mathbb{P}\Big\{\text{fragments from only one pool+WGS} \;\Big|\; M_k = j\Big\} \\ &+ (n-1)\mathbb{P}\Big\{\text{only WGS reads} \;\Big|\; M_k = j\Big\} \\ &= 1 - n\Big(1 - \frac{n-1}{n}\mu\Big)^j + (n-1)(1-\mu)^j.\end{aligned}$$

By Equation (8), the number of fragments in a clone-linked island is distributed by the probabilities

$$\begin{aligned}(9)\quad \mathbb{P}\Big\{D_k, M_k = j \;\Big|\; E\Big\} &= \mathbb{P}\Big\{D_k \;\Big|\; M_k = j, E\Big\}\mathbb{P}\Big\{M_k = j \;\Big|\; E\Big\} \\ &= \left(1 - n\Big(1 - \frac{n-1}{n}\mu\Big)^j + (n-1)(1-\mu)^j\right)(1-J)^{j-1}J \\ &= P_0(j) - nP_\mu^{(n)}(j) + (n-1)P_\mu^{(\infty)}(j).\end{aligned}$$

with

$$P_0(j) = (1-J)^{j-1}J; \tag{10a}$$

$$P_\mu^{(n)}(j) = \left((1-J)\Big(1 - \frac{n-1}{n}\mu\Big)\right)^{j-1} J\Big(1 - \frac{n-1}{n}\mu\Big); \tag{10b}$$

$$P_\mu^{(\infty)}(j) = \Big((1-J)(1-\mu)\Big)^{j-1} J(1-\mu). \tag{10c}$$

Now, for all $0 < z \leq 1$,

$$\sum_{j=1}^{\infty}(1-z)^{j-1} = \frac{1}{z}; \quad \sum_{j=1}^{\infty} j(1-z)^{j-1} = \frac{1}{z^2}. \tag{11}$$

Using Equation (11),

$$\begin{aligned}\mathbb{P}\Big\{D_k \;\Big|\; E\Big\} &= \sum_{j=1}^{\infty}\mathbb{P}\Big\{D_k, M_k = j \;\Big|\; E\Big\} \\ &= 1 - \frac{nJ(1 - \frac{n-1}{n}\mu)}{1 - (1-J)\Big(1 - \frac{n-1}{n}\mu\Big)} + \frac{(n-1)J(1-\mu)}{1-(1-J)(1-\mu)}.\end{aligned}$$

In Equation (2), $p_{\text{link}} = \mathbb{P}\Big\{ D_k \;\Big|\; E \Big\}$. Equation (1) follows from the fact that the expected number of fragments covering the clone equals cL/ℓ.

By definition of the conditional probability,

$$\mathbb{P}\Big\{ M_k = j \;\Big|\; D_k, E \Big\} = \frac{\mathbb{P}\Big\{ D_k, M_k = j \;\Big|\; E \Big\}}{\mathbb{P}\Big\{ D_k \;\Big|\; E \Big\}} = \frac{P_0(j) - nP_\mu^{(n)}(j) + (n-1)P_\mu^{(\infty)}(j)}{p_{\text{link}}},$$

where the values can be plugged in from Equations (2) and (10). By Equation (11),

(12)

$$\mathbb{E}\Big[M_k \;\Big|\; D_k, E \Big] = \frac{p_{\text{link}}^{-1}}{J}\left(1 - \frac{nJ^2\Big(1 - \frac{n}{n-1}\mu\Big)}{\Big(1 - (1-J)(1 - \frac{n-1}{n}\mu)\Big)^2} + \frac{(n-1)J^2(1-\mu)}{\Big(1 - (1-J)(1-\mu)\Big)^2} \right),$$

which corresponds to (ii) with $F_{\text{link}} = \mathbb{E}\Big[M_k \;\Big|\; D_k, E \Big]$. It is interesting to notice that when $\mu = 1$, in Equation (12),

$$\frac{2}{J(\sigma)} \geq \mathbb{E}\Big[M_k \;\Big|\; D_k, E \Big] > \frac{1}{J(\sigma)},$$

and that $\mathbb{E}\Big[M_k \;\Big|\; D_k, E \Big] J^{-1}(\sigma)$ decreases when the coverage c increases.

By Equation (2),

(13)

$$\mathbb{P}\Big\{ \overline{D_k} \;\Big|\; E \Big\} = 1 - p_{\text{link}} = 1 - J\frac{1 - \Big(1-J\Big)\Big(1 - \frac{n-1}{n}\mu\Big)(1-\mu)}{\Big(1 - \Big(1-J\Big)\Big(1 - \frac{n-1}{n}\mu\Big)\Big)\Big(1 - \Big(1-J\Big)\Big(1-\mu\Big)\Big)}.$$

The expected number of fragments in an island that is not mapped to a clone equals

$$\mathbb{E}\Big[M_k \;\Big|\; \overline{D_k}, E \Big] = \frac{\mathbb{E}\Big[M_k \;\Big|\; E \Big] - \mathbb{E}\Big[M_k \;\Big|\; D_k, E \Big]\mathbb{P}\Big\{ D_k \;\Big|\; E \Big\}}{\mathbb{P}\Big\{ \overline{D_k} \;\Big|\; E \Big\}}.$$

Using $\mathbb{E}\Big[M_k \;\Big|\; E \Big] = J^{-1}$ and Equations (12), (2), and (13), we get Equation (4) with the notation $F_{\text{nolink}} = \mathbb{E}\Big[M_k \;\Big|\; \overline{D_k}, E \Big]$.

Let $\ell\lambda_k$ be the length of the island ending with the k-th fragment. The length of a non-linked island can be bounded as $\ell\mathbb{E}\Big[\lambda_{\text{nolink}} \;\Big|\; \overline{D_k}, E \Big]$ with

$$1 \leq \mathbb{E}\Big[\lambda_{\text{nolink}} \;\Big|\; \overline{D_k}, E \Big] \leq \mathbb{E}\Big[M_k \;\Big|\; \overline{D_k}, E \Big]\sigma + \vartheta.$$

The bounds of Equation (6) follow from

$$\mathbb{E}\Big[\lambda_k \;\Big|\; D_k, E\Big] = \frac{\mathbb{E}\Big[\lambda_k \;\Big|\; E\Big] - \mathbb{E}\Big[\lambda_k \;\Big|\; \overline{D_k}, E\Big]\mathbb{P}\Big\{\overline{D_k} \;\Big|\; E\Big\}}{\mathbb{P}\Big\{\overline{D_k} \;\Big|\; E\Big\}},$$

where $\mathbb{E}\Big[\lambda_k \;\Big|\; E\Big] = \lambda_{\text{CBC}} = \frac{J^{-1}-1}{c} + \vartheta$ (Waterman 1995). □

Figures 2 and 3 compare different experimental designs based on Theorem 1 and simulations. Figure 2 plots the island statistics from the theorem. It illustrates that for lower coverages (about $c < 4$), the ratio of pooled shotgun fragments makes a large difference in the sequencing. This difference is mainly shown in the number of clone-linked contigs, as the contig sizes do not differ much. At large coverage levels, when sequencing is nearly completed, the impact of pooled fragments is less, i.e., WGS sequence reads can make up for a lower pooled coverage.

Figure 3a shows that while more arrays increase the sequencing success, the improvements are very small after the second array. Notice that if the clones are selected from a minimally overlapping tiling path, then no part of the genome is covered by more than two BACs, and thus two arrays suffice for the unambiguous mapping of all contigs that cover clone overlaps. Figure 3b plots the N50 values. The N50 contig length is the value l such that half of the sequenced nucleotides belong to contigs of length at least l. The statistics for all designs converge to those of a non-pooled sequencing project as the coverage increases. In other words, the negative effects of pooling diminish and the project progresses just as without pooling: for example, at total coverage 4–5X, 99% of the clone is sequenced.

The value λ_{CBC} in the theorem is the expected island length in a non-pooled sequencing project. By Equation (7), and the fact that $\lim_{c\to\infty} p_{\text{link}} = 1$, we have $\lim_{c\to\infty} \lambda_{\text{link}} = \lambda_{\text{CBC}}$ when the ratio of pooled shotgun fragments is kept constant. This limit result is not surprising given that every island can be assigned to a clone with near certainty when the sequence read coverage is large.

4. Pooled shotgun fragments for overlap detection

The key observation for this section is that a transversal design makes it possible to map a contig unambiguously to more than one BAC at once. Now, a contig that is mapped to two clones simultaneously can be viewed as evidence that the two clones overlap. Taking the idea further, an entire set of BACs can be tested for overlaps in this manner,

which leads us to the Clone-Array Pooled Shotgun Mapping (CAPS-MAP) method that is described as follows. A redundant collection of random BACs covering a large genome is grouped into subsets of size q^2. Pooled shotgun sequence reads are collected from each clone group using a transversal design with d arrays of size $q \times q$. Partitioning into subsets may be dictated by the practical concerns of chemistry, biology and robotic automation. For array sizes that are multiples of 8 or 12 or both (yielding standard dimensions of a 96-well microtiter plate), such as $q = 24$, or $q = 48$, there exist known (Colbourn and Dinitz 1996) transversal designs. A pooling design with a few ($d = 2, 3, 4$) arrays suffice to compute the physical ordering of BACs in the library, depending on the library's redundancy and the array sizes. In addition to the pooled shotgun fragments, WGS fragments are used to increase read contig lengths. The fragments are compared to each other to find the overlaps between them, and overlapping fragments are assembled into contigs. Contigs that map unambiguously to more than one clone are taken as evidence that the clones overlap. Figure 4 shows how overlaps between clones in different clones can be detected. Figure 5 shows how overlaps between clones in the same clone group can be detected even in the presence of false negative errors. The clone overlap information can then be used to compute the physical ordering of the BACs in the library, and to select a minimal tiling path for complete sequencing, just as if the overlaps were detected using a fingerprinting scheme (Marra et al. 1997). The following theorem considers the case of detecting an overlap between two clones in different clone groups. Similar analyses can be carried out for more general cases with more overlapping clones, or clones in the same clone group, resulting in more cumbersome formulas.

Theorem 2. *Let two clones from different clone groups share an overlap. Define $c_2 = 2a+w$, the total sequence read coverage for the overlap. Define*

$$\beta_1 = \frac{w + (1 + \frac{1}{n})a}{c_2} \quad \beta_2 = \frac{w + a}{c_2} \quad \beta_3 = \frac{w + \frac{2a}{n}}{c_2} \quad \beta_4 = \frac{w + \frac{a}{n}}{c_2} \quad \beta_5 = \frac{w}{c_2};$$
$$\gamma_i = 1 - (1 - e^{-c_2\sigma})\beta_i \quad \textit{for } i = 1, \ldots, 5.$$

(i) *An apparent island in the overlap consisting of $j > 0$ fragments is mapped to the two clones simultaneously with probability $1 - q(j)$ where*

$$q(j) = 2n\beta_1^j - 2(n-1)\beta_2^j - n^2\beta_3^j + 2n(n-1)\beta_4^j - (n-1)^2\beta_5^j < 2n\beta_1^j. \tag{14}$$

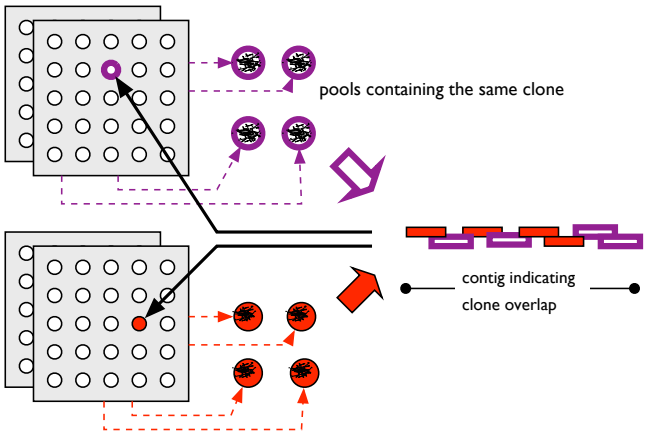


FIGURE 4. CAPS-MAP detects overlaps between clones by identifying situations where a read contig maps simultaneously to two clones. This figure illustrates a transversal pooling design for array pairs containing identical sets of BACs. The transversal design guarantees that the intersection of any two pools out of four possible for each BAC (two row and two column pools) uniquely identifies the BAC.

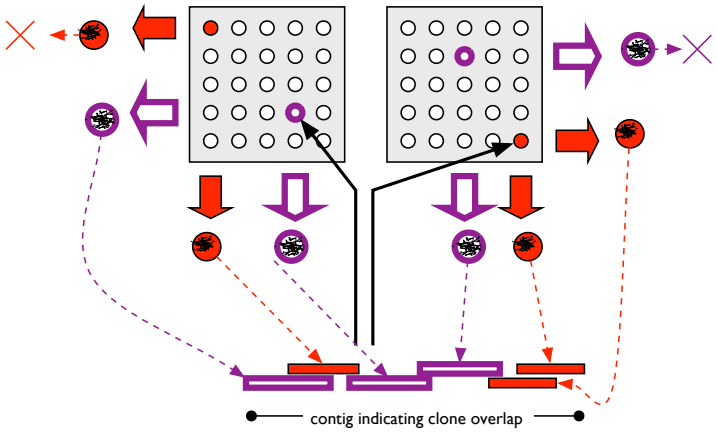


FIGURE 5. Overlaps between clones on the same array can also be detected by a transversal design, even in the presence of false negatives, i.e., situations where a particular BAC is not represented in a particular pool. Specifically, overlap between the two BACs illustrated in the figure is detected despite the fact that each BAC is sampled in only three pools.

(ii) *An apparent island covering the overlap is mapped to the two clones simultaneously with probability*

(15)
$$p_2 = 1 - e^{-c_2\sigma}\left(2n\frac{\beta_1}{\gamma_1} - 2(n-1)\frac{\beta_2}{\gamma_2} - n^2\frac{\beta_3}{\gamma_3} + 2n(n-1)\frac{\beta_4}{\gamma_4} - (n-1)^2\frac{\beta_5}{\gamma_5}\right).$$

Proof. The overlap is detected if it is covered by an island that can be simultaneously mapped to the two clones. We model the location of the shotgun fragments as a Poisson process with rate c_2. Define $\mu_2 = \frac{2a}{c_2}$, the fraction of pooled fragments covering the overlap. Every fragment is either a WGS fragment with probability $(1-\mu_2)$, or comes from each one of the two clones' pools with probability $\mu_2/(2n)$. The event E_2 that a given fragment is the right-hand end of an apparent island has probability $J_2 = \mathbb{P}E_2 = e^{-c_2\sigma}$. For the k-th fragment, define M_k as the number of fragments from its right-hand end until the first gap towards the left. The probability that an island has j fragments in it equals

$$\mathbb{P}\Big\{M_k = j \;\Big|\; E_2\Big\} = \Big(1 - J_2\Big)^{j-1} J_2.$$

The probability of mapping the island that ends at the k-th fragment (event D_k) depends on the number of fragments in the island. We calculate the probability of event $\overline{D_k}$ in separate cases. Let $p_{0,0}(j)$ denote the event that the island consists of WGS fragments only given that it has j reads. Then

$$p_{0,0}(j) = (1-\mu_2)^j. \tag{16a}$$

Let $p_{0,*}(j)$ denote the event that the island consists of pooled fragments for one clone only and WGS fragments, given that it has j fragments:

$$p_{0,*}(j) = \Big(1 - \frac{\mu_2}{2}\Big)^j. \tag{16b}$$

Let $p_{1,0}(j)$ denote the event that the island consists of pooled fragments from a fixed pool and WGS fragments, given that it has j fragments:

$$p_{1,0}(j) = \Big(1 - \frac{n-\frac{1}{2}}{n}\mu_2\Big)^j - p_{0,0}(j). \tag{16c}$$

Let $p_{1,1}(j)$ denote the event that the island consists of pooled fragments from a fixed pool for one clone, from another fixed pool for the other clone, and WGS fragments, given that it has j fragments:

$$p_{1,1}(j) = \Big(1 - \frac{n-1}{n}\mu_2\Big)^j - 2p_{1,0}(j) - p_{0,0}(j). \tag{16d}$$

Let $p_{1,+}(j)$ denote the event that the island consists of pooled fragments from a fixed pool for one clone, at least one pooled fragment for the

other clone, and WGS fragments:

$$p_{1,+}(j) = \Big(1 - \frac{n-1}{2n}\mu_2\Big)^j - p_{0,*}(j) - p_{1,0}(j). \tag{16e}$$

Using inclusion-exclusion,

$$\mathbb{P}\Big\{\overline{D_k} \;\Big|\; E_2, M_k = j\Big\} = \Big(2p_{0,*}(j) - p_{0,0}(j)\Big) + 2np_{1,+}(j) - n^2 p_{1,1}(j).$$

By Equations (16a–16e),

$$\mathbb{P}\Big\{\overline{D_k} \;\Big|\; E_2, M_k = j\Big\} = 2n\Big(1 - \frac{n-1}{2n}\mu_2\Big)^j - 2(n-1)\Big(1 - \frac{\mu_2}{2}\Big)^j \tag{17}$$
$$- n^2\Big(1 - \frac{n-1}{n}\mu_2\Big)^j + 2n(n-1)\Big(1 - \frac{n-\frac{1}{2}}{n}\mu_2\Big)^j - (n-1)^2(1-\mu_2)^j,$$

which corresponds to Equation (14) with $q(j) = \mathbb{P}\Big\{\overline{D_k} \;\Big|\; E_2, M_k = j\Big\}$. Using the same technique as before

$$1 - p_2 = \mathbb{P}\Big\{\overline{D_k} \;\Big|\; E_2\Big\} = \sum_{j=1}^{\infty} \mathbb{P}\Big\{\overline{D_k} \;\Big|\; E_2, M_k = j\Big\}\mathbb{P}\Big\{M_k = j \;\Big|\; E_2\Big\},$$

leading to Equation (15).

Recall that $q(j)$ is the probability of failing to map a contig of j reads to the two clones simultaneously. In order to show that the inequality in Equation (14) holds, we prove that

$$q(j) < 2n\beta^j - (2n-1)\beta_3^j < 2n\beta_1^j. \tag{18}$$

Notice that $\beta_5 < \beta_4 < \beta_3 < \beta_2 < \beta_1$ and thus $q(j) \rightsquigarrow 2n\beta_1^j$. Since $\beta_4 = (\beta_3 + \beta_5)/2$, it follows from the convexity of x^j that

$$2\beta_4^j \le \beta_3^j + \beta_5^j. \tag{19}$$

(Alternatively, notice that the same inequality follows from $p_{1,1}(j) \ge 0$ in Equation (16d).) We proceed by rearranging the equality of Equation (14):

$$\begin{aligned} 2n\beta_1^j - (2n-1)\beta_3^j - q(j) &= 2(n-1)\beta_2^j + (n-1)^2\beta_3^j - 2n(n-1)\beta_4^j + (n-1)^2\beta_5^j \\ &= (n-1)^2 \underbrace{\Big(\beta_3^j + \beta_5^j - 2\beta_4^j\Big)}_{>0 \text{ by Eq. (19)}} + 2(n-1) \underbrace{\Big(\beta_2^j - \beta_4^j\Big)}_{>0 \text{ since } \beta_2 > \beta_4}, \end{aligned}$$

which proves Equation (18). □

It is difficult to derive useful closed formulas for the probability of overlap detection. For example, based on Equation (15), the number of contigs in the overlap that are simultaneously mapped to the clones can be modeled as a Poisson random variable with expected value $c_2 e^{-c_2\sigma} p_2$.

For practical values of c_2, this model seriously underestimates the probability of overlap detection. The problem is similar to the one of using Lander-Waterman statistics (Lander and Waterman 1988) at high coverages (see Wendl and Waterston (2002) for a discussion). For a more suitable model, let G be the number of gaps entirely contained in the overlap, and number the islands from 0 to G. Let $j_0, j_2, \ldots, j_G$ denote the number of fragments in the islands. The probability that none of the islands can be mapped simultaneously to the two clones can be calculated as $p_{\text{nomap}}(j_0, \ldots, j_G) = \prod_{i=0}^{G} q(j_i)$. Notice that G and the j_i are random variables. We are interested in the expected value

$$p_{\text{nomap}} = \mathbb{E}p_{\text{nomap}}(j_0, \ldots, j_G). \tag{20}$$

In order to get a good assessment of CAPS-MAP performance, we found that it is best to use a Monte-Carlo estimation of this expected value; see Figure 6. Alternatively, by the inequality of Equation (14), $p_{\text{nomap}} < \mathbb{E}\left[\beta_1^R (2n)^{G+1}\right]$ where R is the number of fragments in the overlap, and thus $R = \sum_{i=0}^{G} j_i$. Following an approach similar to that of Wendl and Waterston (2002), which models the underlying coverage process more carefully, we can derive bounds (see Appendix) that are useful for large values of c_2 (e.g., $c_2 = 7$), but at lower coverages, this approach also underestimates the overlap detection probabilities significantly.

Based on Figure 6, the probability of detecting an overlap increases exponentially toward 1 with the overlap length. The same exponential behavior is characteristic of clone anchoring methods for overlap detection (Arratia et al. 1991). Consequently, clone contig statistics for CAPS-MAP can be calculated using a clone anchoring model with an appropriate anchoring process intensity. Clone contig statistics can also be estimated using a fingerprinting model (Lander and Waterman 1988) by noticing that clone overlaps above a certain length are detected with near certainty. Figure 6 indicates that using 1X pooled shotgun coverage and 2–5X WGS coverage, BAC overlaps of more than 20000 bp are detected almost certainly. While CAPS-MAP uses only the fact that a contig is mapped to multiple BACs, and not the actual contig sequence, the sequence information is used in the ensuing sequencing phase, and thus CAPS-MAP represents very little overhead in a whole-genome sequencing project.

It is worth pointing out here that CAPS-MAP detects very short, or even *negative* clone overlaps with non-negligible probability. A short region of the genome that is not covered by BACs in the library can be bridged by WGS fragments. The bridging WGS fragments may form a

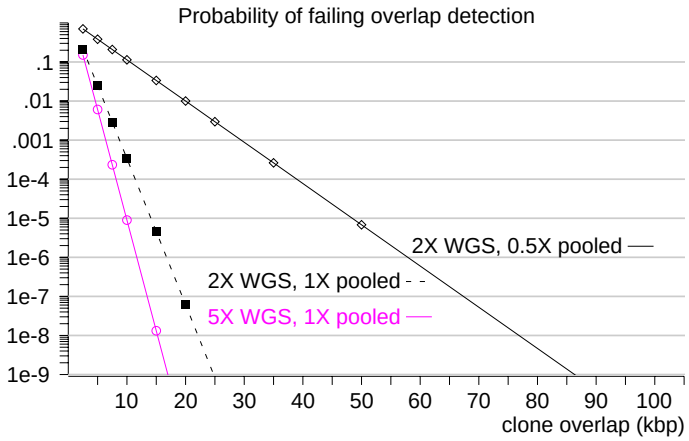


FIGURE 6. Clone overlap detection. The graph shows the probability of not detecting an overlap between two clones, as a function of the overlap size. The plots were calculated by a Monte-Carlo method using Theorem 2. All plots use $\vartheta = 0.1$ for shotgun fragment overlap detection, and $\ell = 500$ for shotgun fragment length.

contig with pooled fragments from the two BACs at the gap's ends that can be mapped to the two clones simultaneously. This unique feature of CAPS-MAP among clone overlap detection methods does not interfere with the calculation of the physical ordering of BACs. At the same time, it does decrease the necessary BAC library size for sequencing the genome completely. After the clones are selected for complete sequencing, the already collected WGS fragments are included in the genome sequence assembly. Consequently, negative overlaps detected by CAPS-MAP are already covered by shotgun fragments in the sequencing phase, and pose no additional requirements for shotgun sequence collection.

5. Discussion

Our analyses presented here indicate the theoretical feasibility of the CAPS-MAP method and provide guidance for the design of genome-scale CAPS-MAP experiments. In particular, our analysis indicates that transversal pooling designs can accommodate high levels of clone redundancy and perform well even at low levels of shotgun sequence coverage of clone pools. Moreover, transversal designs also perform well in cases where BACs are not equally represented in the pools, thus making certain BACs "invisible" in certain pools.

Practical biological and technical considerations may set a limit to the array size. In case of large genomes, the limitations may imply that the set of BACs is partitioned and that pooling is applied separately to individual subsets. This results lower clone redundancy within individual arrays and a larger number of pools. The analysis presented here allows for the partitioning of clones. It also allows for the possibility of including whole-genome shotgun sequence reads. Thus, our analysis covers a realistic and practical scenario of the CAPS-MAP method's application. Pooling methods are ideally suited for hybrid sequencing strategies that combine BAC-based and whole-genome sequencing methods. The whole-genome sequencing strategy (Weber and Myers 1997) was originally proposed to bypass the need for physical mapping. However, the absence of BAC localization information hampers assembly and finishing of the highly-repetitive mammalian genomes such as that of humans, naturally leading to a hybrid strategy involving a combination of BAC-based and whole-genome sequencing. In addition to having been used to assemble the initial draft of the human genome (Venter et al. 2001), such hybrid strategies are currently being pursued in the context of sequencing the mouse and rat genomes. The overhead cost of CAPS-MAP in a hybrid whole-genome project is very small, since the collected shotgun reads can be reused in the sequencing phase.

The product of applying CAPS-MAP is a set of BAC contigs. The contigs themselves are not anchored to particular chromosomes. Thus, a sparse set of sequence markers such as mapped STS sequences would be necessary to anchor the islands onto chromosomes and thereby complete the physical mapping.

Alternatively, individual BACs can be comparatively mapped onto sequenced genomes of related species using the recently described Pooled Genomic Indexing (PGI) method (Csűrös and Milosavljevic 2002). The advantage of PGI is that it uses the same pooling and shotgun sequencing information generated for the purpose of CAPS-MAP. Thus, PGI can be applied concurrently and without any additional experimental effort to obtain comparative physical maps.

One practically appealing aspect of these pooling methods is that they can be introduced with minimal alterations of existing clone-by-clone sequencing pipelines, such as the one at the Human Genome Sequencing Center at Baylor College of Medicine. Another appealing aspect is the ease of adjusting experimental parameters, such as the depth of shotgun sequencing of pools and the number of arrays in the course of a whole-genome experiment. The possibility of making

adjustments translates into a significantly better control over the final outcome of a whole-genome project.

CAPS-MAP, CAPSS, and PGI rely on essentially the same experimental data. One difference is that CAPSS in principle requires higher pool coverage than the one required for mapping by CAPS-MAP or comparative mapping by PGI. Another difference is that CAPSS may optionally be performed on a subset (tiling path) of BACs selected based on a map obtained by CAPS-MAP, by PGI, or by other methods, whereas CAPS-MAP arrays must contain highly redundant collections of BACs in order to guarantee a high number of clone overlaps. Despite these differences all three pooled shotgun methods reduce the experimental "instruction set" for both mapping and sequencing to just pooling and shotgun sequencing, thus allowing more streamlined, specialized, and controlled data production.

Acknowledgments

We are grateful to Richard Gibbs and George Weinstock for sharing pre-publication information on CAPSS and for useful comments. Our discussion of computing CAPS-MAP overlap detection probabilities has greatly benefited from conversations with Luc Devroye and Michael Waterman. This work was supported by grants 1 RO1 HG02583-01 from NHGRI at the National Institutes of Health, 250391-02 from the Natural Sciences and Engineering Research Council of Canada; Howard Hughes Medical Institute faculty startup funds, and Université de Montréal faculty startup funds.

References

Arratia, R., E. S. Lander, S. Tavaré, and M. S. Waterman (1991). Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics 11*, 806–827.

Cai, W.-W., R. Chen, R. A. Gibbs, and A. Bradley (2001). A clone-array pooled strategy for sequencing large genomes. *Genome Research 11*, 1619–1623.

Colbourn, C. J. and J. H. Dinitz (Eds.) (1996). *The CRC Handbook of Combinatorial Designs*. Boca Raton: CRC Press.

Csűrös, M. and A. Milosavljevic (2002). Pooled genomic indexing (PGI): mathematical analysis and experiment design. In *Algorithms in Bioinformatics: Second International Workshop*, Volume 2452 of *LNCS*, pp. 10–28. Berlin Heidelberg: Springer-Verlag.

Du, D.-Z. and F. K. Hwang (2000). *Combinatorial Group Testing and Its Applications* (2nd ed.). Singapore: World Scientific.

Ewens, W. J. and G. R. Grant (2001). *Statistical Methods in Bioinformatics: An Introduction*. New York: Springer-Verlag.

IHGSC (2001). Initial sequencing and analysis of the human genome. *Nature 609*(6822), 860–921.

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

Marra, M. A., T. A. Kucaba, N. L. Dietrich, E. D. Green, B. Brownstein, R. K. Wilson, K. M. McDonald, L. W. Hillier, J. D. McPherson, and R. H. Waterston (1997). High throughput fingerprint analysis of large-insert clones. *Genome Research 7*, 1072–1084.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. (2001). The sequence of the human genome. *Science 291*, 1304–1351.

Waterman, M. S. (1995). *Introduction to Computational Molecular Biology: Maps, Sequences and Genomes*. Boca Raton: Chapman & Hall.

Weber, J. L. and E. W. Myers (1997). Human whole-genome shotgun sequencing. *Genome Research 7*, 401–409.

Wendl, M. C. and R. H. Waterston (2002). Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. *Genome Research 12*, 1943–1949.

Appendix

Here we expand our discussion on the probability of overlap detection in CAPS-MAP. In particular, we derive formulas that show the exponential decay of the probability of not detecting an overlap when the coverage c_2 is not too small. We start with the bound

$$p_{\text{nomap}} < \mathbb{E}\Big[\beta_1^R (2n)^{G+1}\Big] \tag{21}$$

Define

$$\mathcal{G}_r(z) = \mathbb{E}\Big[z^G \;\Big|\; R\Big],$$

the probability generating function for the distribution of the number of gaps conditioned on the number of fragments. Define the events A_i for $i = 1, \ldots, r-1$: A_i denotes the event that the i-th fragment is followed by a gap, conditioned on the event $\{R = r\}$. For arbitrary g, and set of indexes $i_1 < i_2 < \cdots < i_g$,

$$\mathbb{P}\Big\{A_{i_1} A_{i_2} \cdots A_{i_g}\Big\} = (1 - g\delta)_+^r,$$

where $\delta = \frac{\sigma\ell}{\Theta L}$, and $(x)_+ = \max\{0, x\}$ (Ewens and Grant 2001; Wendl and Waterston 2002). Let

$$\begin{aligned} S_0 &= 1 \\ S_g &= \sum_{i_1 < \cdots < i_g} \mathbb{P}\Big\{A_{i_1} A_{i_2} \cdots A_{i_g}\Big\} = \binom{r-1}{g} (1 - g\delta)_+^r. \end{aligned}$$

Using inclusion-exclusion,

$$\mathbb{P}\Big\{G = g \;\Big|\; R = r\Big\} = \sum_{j=g}^{r-1} \binom{j}{g} (-1)^{j-g} S_j.$$

Hence,

$$\begin{aligned} \mathcal{G}_r(z) &= \sum_{g=0}^{r-1} z^g \sum_{j=g}^{r-1} \binom{j}{g} (-1)^{j-g} S_j \\ &= \sum_{j=0}^{r-1} S_j \sum_{g=0}^{j} (-1)^{j-g} \binom{j}{g} z^g \\ &= \sum_{j=0}^{r-1} S_j (z-1)^j. \end{aligned}$$

Substituting the S_j values:

$$\mathcal{G}_r(z) = \sum_{j=0}^{r-1} \binom{r-1}{j} (1-j\delta)_+^r (z-1)^j, \tag{22}$$

a result interesting on its own.

Returning to Equation (21), we have

$$p_{\text{nomap}} < \mathbb{E}\left[2n\beta_1^R \sum_{j=0}^{R-1} \binom{R-1}{j} (1-j\delta)_+^R (2n-1)^j\right], \tag{23}$$

where R is a Poisson random variable with mean

$$\lambda = \frac{c_2 \Theta L}{\ell}$$

For every $r \geq 0$, $(1-j\delta)_+^r \leq e^{-jr\delta}$, hence

$$\sum_{j=0}^{r-1} \binom{r-1}{j} (1-j\delta)_+^r (2n-1)^j \leq \left(1 + (2n-1)e^{-r\delta}\right)^{r-1}.$$

Consequently, by Equation (23),

$$p_{\text{nomap}} < \mathbb{E}\left[2n\beta_1^R \left(1 + (2n-1)e^{-R\delta}\right)^{R-1}\right].$$

Recall that the random value we take the expectation of is an upper bound on $p_{\text{nomap}}(j_0, \ldots, j_G)$, and thus if it is larger than one, it is useless. Let

$$f(r) = \min\left\{1, 2n\beta_1^r \left(1 + (2n-1)e^{-r\delta}\right)^{r-1}\right\}.$$

So we have in fact the bound

$$p_{\text{nomap}} < \mathbb{E} f(R). \tag{24}$$

In order to achieve exponential decay in the bound, we would like to have

$$\beta_1 \left(1 + (2n-1)e^{-r_0\delta}\right) < 1$$

for some $r_0 < \lambda$. Rearranging the inequality, we have

$$(2n-1)\frac{n(a+w)+a}{(n-1)a} < e^{(2a+w)\sigma}, \tag{25}$$

which is satisfied when a and w are not too small (see Figure 7).

There are several possible ways to exploit the fact that the exponential component of $f(r)$ becomes for r less than the expected value λ. The main idea is that when evaluating $\mathbb{E} f(R) = \sum f(r)\mathbb{P}\{R = r\}$ in Equation (24), either the probability of $R = r$ is small, or the value of $f(r)$ is small. Let $0 < k < \lambda$ be a threshold (that we specify later),

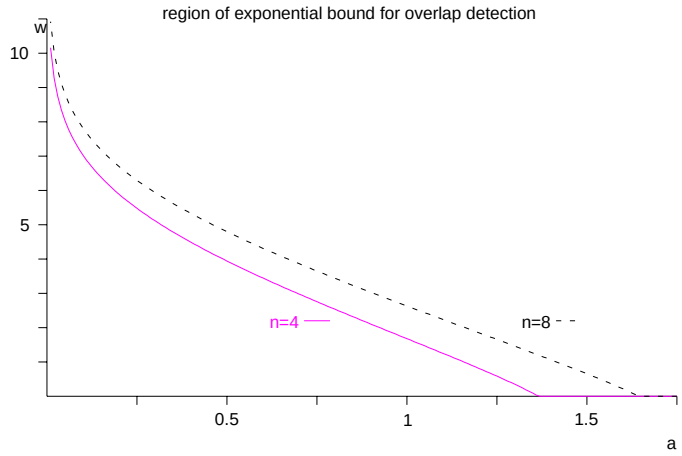


FIGURE 7. Values of the pooled shotgun coverage a and WGS coverage w, for which the clone overlap detection bound applies, are above the graphs (see Equation (25))

.

and let $\alpha = k/\lambda$. To proceed with Equation (24), we condition on the event $\{R \leq \alpha\lambda\}$. We use the bound

$$\mathbb{P}\{R \leq \alpha\lambda\} < \frac{e^{-\lambda(1-\alpha)^2/2}}{(1-\alpha)\sqrt{2\pi\alpha\lambda}}. \tag{26}$$

By definition,

$$\begin{aligned}\mathbb{P}\{R \leq \alpha\lambda\} &\leq \sum_{r=0}^{k} \frac{\lambda^r}{r!} e^{-\lambda} < e^{-\lambda}\frac{\lambda^k}{k!}\sum_{r=0}^{k}\left(\frac{k}{\lambda}\right)^r \\ &< e^{-\lambda}\frac{\lambda^k}{k!}(1-\alpha)^{-1} < e^{-\lambda(1-\alpha+\alpha\ln\alpha)}\frac{1}{(1-\alpha)\sqrt{2\pi\alpha\lambda}},\end{aligned}$$

where we used a Stirling approximation: $k! > (k/e)^k/\sqrt{2\pi k}$. Using a Taylor series expansion,

$$1-\alpha+\alpha\ln\alpha = \frac{1}{2}(1-\alpha)^2 + \frac{1}{6}(1-\alpha)^3 + \frac{1}{12}(1-\alpha)^4\ldots$$

and thus $1-\alpha+\alpha\ln\alpha > \frac{1}{2}(1-\alpha)^2$ for $0 < \alpha < 1$, and Equation (26) follows.

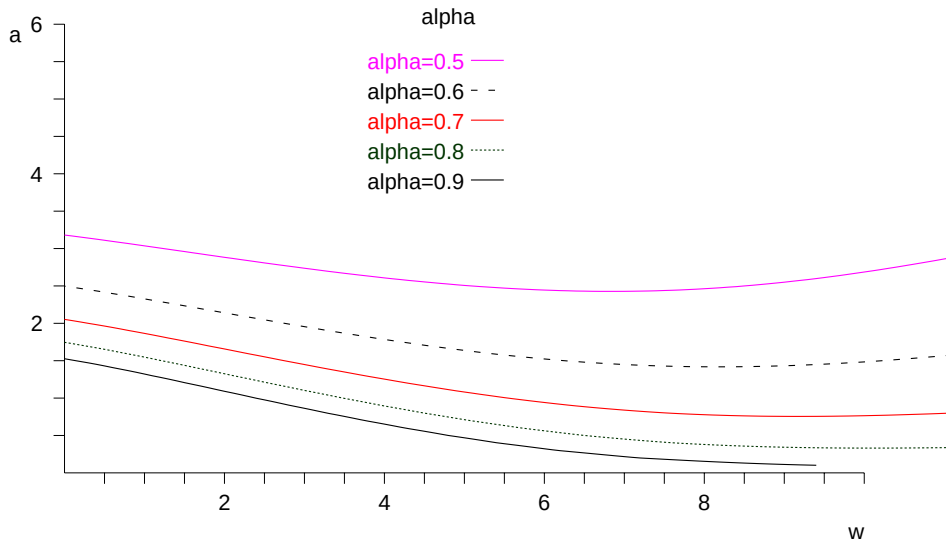


FIGURE 8. Balanced α values for our exponential bound.

Now,

$$\begin{aligned}
\mathbb{E}f(R) &= \mathbb{E}\Big[f(R) \;\Big|\; R \le \alpha\lambda\Big]\mathbb{P}\{R \le \alpha\lambda\} + \mathbb{E}\Big[f(R) \;\Big|\; R > \alpha\lambda\Big]\mathbb{P}\{R > \alpha\lambda\} \\
&\le \mathbb{P}\{R \le \alpha\lambda\} + \mathbb{E}\Big[f(R) \;\Big|\; R > \alpha\lambda\Big] \\
&< \frac{e^{-\lambda(1-\alpha)^2/2}}{(1-\alpha)\sqrt{2\pi\alpha\lambda}} + \frac{2ne^{-\lambda}\sum_{r=0}^{\infty}\frac{\left(\beta_1(1+(2n-1)e^{-\alpha\delta\lambda})\right)^r}{r!}}{1+(2n-1)e^{-\alpha\delta\lambda}} \\
&= \frac{\exp\Big(-\lambda(1-\alpha)^2/2\Big)}{(1-\alpha)\sqrt{2\pi\alpha\lambda}} + \frac{2n\exp\Big(-\lambda\Big(1-\beta_1(1+(2n-1)e^{-\alpha c_2\sigma})\Big)\Big)}{1+(2n-1)e^{-\alpha c_2\sigma}},
\end{aligned}$$

where we used $\delta\lambda = c_2\sigma$. Figure 8 shows values of α for different a, w pairs that balance the exponents in the two terms.

After choosing a balancing α value for a given (a, w) pair, we obtain

$$\mathbb{E}f(R) < X_1 \exp(-X_2\sigma L),$$

where X_1 and X_2 are constants that do not depend on σ. The bound becomes small ($< 10^{-8}$) for larger c_2 values (e.g., $c_2 = 7$), but even then, it is not very tight. Based on simulation results, the tightness is lost with the inequality of Equation (21), and not in the following steps. For

example, we evaluated the bounds of Equations (23) and (24) numerically. While they are fairly close to each other, and to the exponential bound using α, they already bound the expected value of Equation (20) rather loosely in many cases. Furthermore, even for (a, w) pairs for which we cannot establish exponential decay using the inequality of Equation (21), the overlap detection probability may get very close to one. For instance, a two-array design with $a = 0.5$ and $w = 2$ falls below the curve of Figure 7, yet can be employed efficiently in CAPS-MAP as shown in Figure 6. Therefore, we prefer using a Monte-Carlo evaluation of Equation (20) to predict the experimental performance of CAPS-MAP.

MC: Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada. Phone: +1 (514) 343-6111x1655, Fax: +1 (514) 343-5834.

E-mail address: csuros@iro.umontreal.ca

URL: http://www.iro.umontreal.ca/~csuros/

AM: Bioinformatics Research Laboratory and Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA.

E-mail address: amilosav@bcm.tmc.edu