

# Pooling Strategies for Establishing Physical Genome Maps Using FISH

Fengzhu Sun<sup>1\*</sup>  
Gary Benson<sup>2</sup>  
Norm Arnheim<sup>3</sup>  
Michael Waterman<sup>3,4</sup>

<sup>1</sup>Department of Genetics  
Emory University School of Medicine  
Atlanta, GA 30322  
U. S. A.

<sup>2</sup>Department of Biomathematical Sciences  
The Mount Sinai School of Medicine  
New York, NY 10029-6574  
U. S. A.

Department of Mathematics<sup>4</sup> and <sup>3</sup>Molecular Biology  
University of Southern California  
Los Angeles, CA 90089-1113  
U. S. A.

\* To whom correspondence should be made. Phone (404)-727-1702, Fax (404)-727-3949.

Running head: GENOME MAP USING FISH

Keywords: FISH; Chromosome characterization; Experimental design; Coupon collector problem; Mathematical modeling

## Abstract

Often, in biological studies, it is necessary to identify an organism's chromosomes. In some organisms the individual chromosomes can be identified by staining procedures while many other species have a very large number of chromosomes, often of similar size, which defy identification by traditional staining methods. We have devised strategies, based on fluorescent *in situ* hybridization (FISH), which allow the assignment of a preset number of probes to each chromosome without prior chromosome identification. By hybridizing mixtures of probes labeled with different colored fluorescent molecules the chromosomal origin of each probe can be determined.

# 1 Introduction

The genome project is an effort to decipher the genetic code of humans and a small number of experimental organisms including the mouse, fruit fly, yeast and the common bacteria *E. coli*. A great many other species hold fascination for life scientists but are not likely to be analyzed in as much detail since their impact on health related research is not thought to be very high. Yet, genome analysis of many of these “orphan” species could provide fundamental information on important biological and evolutionary questions. Thus fast and relatively simple methods to investigate genome structure from virtually any organism would be of great value.

Fluorescent *in situ* hybridization (FISH) is a method of genome analysis that can localize a specific DNA sequence to a chromosome (Johnson et al. 1991, Le Beau 1996, Lichter et al. 1990, Speicher 1996). If the genome of an organism is fragmented and individual pieces isolated by recombinant DNA cloning, then the original chromosomal location of any fragment can be determined. First, the fragment (probe) is fluorescently labeled and then added to a preparation of cells in which the chromosomes remain intact. The fragment will anneal (hybridize) to its chromosome of origin and a microscope can be used to detect the fluorescent label on a single chromosome. By carrying out FISH with many probes, a large number of them can be associated with each individual chromosome.

In some organisms the individual chromosomes can be identified by staining procedures and thus probes can be localized by FISH to their particular chromosomes with little difficulty. On the other hand, many species have a very large number of chromosomes, often of similar size, and may defy identification by traditional staining methods. We have devised strategies which allow the assignment of a preset number of probes to each chromosome without prior chromosome identification. By hybridizing mixtures of probes labeled with fluorescent molecules of different colors, the chromosomal origin of each probe can be determined.

In this study, we formulate a problem involving the assignment of probes to chromosomes using FISH. We first formalize the problem and then discuss and analyze three separate solutions. The results of assigning probes to chromosomes might be the raw data for the problem of comparing sets of synthetic genes. See Ferretti et al. (1996).

## 2 Problem Description

We are given a set of  $N$  chromosomes. We want to assign at least  $m$  probes to each chromosome. Assignments are done by *experiments*. In each experiment, we use some number of probes. Each probe is colored with one of the available colors. Generally, the number of available colors is less than the number of probes in an experiment. The colored probes are presented to the  $N$  chromosomes. The origin of each probe is determined by observation of the occurrences of the colors on the chromosomes. Since there are usually more probes than colors, observing a color on a chromosome does not give precise information about which probe is on that chromosome. Several experiments may be required to determine which probe is on the chromosome. We emphasize that *the chromosomes are in no way distinguishable*. Observation that a probe  $x$  is on a chromosome does not tell us which chromosome it is on. All we know is that it is on the chromosome of origin for probe  $x$ . We call the method of combining probes and colors in an experiment the *protocol*. The method of collecting information from the experiments we denote the *strategy*. In this study, we present several solutions for the following problem.

**Chromosome Characterization Problem (CCP):** Devise a protocol for the experiments and a strategy for collecting information from the experiments such that the number of experiments required to assign at least  $m$  probes to each of the  $N$  chromosomes is *minimized*.

Since we have no control over the origin of each probe, this problem is naturally a statistical one. Our solutions therefore are designed to minimize the *expected* number of experiments. In what follows, we propose and analyze several experimental strategies. Below we assume that each probe in the library belongs to a unique chromosome. (If this is false, we discard the probe.)

We use the following parameters:

- $N$  = Number of chromosomes;
- $m$  = Number of probes to be assigned to each chromosome;
- $C$  = Number of available colors.

In this paper, we present three strategies to solve the chromosome characterization problem in three separate sections. In each section we first give the experimental protocol and simulations. Then we present the mathematical analysis for the experimental strategies. Biologists who are interested in the experiments can look at the experimental protocol without going into detail about the mathematical analysis.

## 3 The First Strategy

### 3.1 The Experiment Protocol

We divide the experiments into two jobs. First we assign one probe to each chromosome. Then for each fixed chromosome, we assign  $m - 1$  extra probes to the chromosome. In this way, we can assign  $m$  probes to all the chromosomes. We explain the two jobs in detail next. The strategy is closely related to the coupon collectors problem (Feller, 1968).

#### Job 1: One probe per chromosome

First, the experiments are done by analyzing one probe at a time until we assign at least one probe to each chromosome. Our protocol uses one color, without loss of generality, red. The number of probes per experiment increases as described below. Let  $k$  be the number of chromosomes that have been assigned one probe. Our strategy is as follows:

1. Randomly choose a probe and denote it as  $c_1$ . We call  $c_1$  a “success probe”. Set  $k = 1$ .
2. Randomly choose another probe which we call a “sample probe”. Color the  $k$  “success probes” and the “sample probe”. Put the success probes and the sample probe together into one experiment.
3. If the probes including the  $k$  “success probes” and the “sample probe” belong to  $k + 1$  different chromosomes, i.e., the “sample probe” does not belong to the same chromosome as any of the “success probes”, we add the “sample probe” to the group of “success probes”. Set  $k = k + 1$ .  
Otherwise, discard the sample probe.
4. Continue from step 2 until  $k = N$ , i.e., all the chromosomes are assigned one probe.

Notice that the last few chromosomes take more experimental steps to obtain one probe than the first few chromosomes do because a success probe is more difficult to find. We analyze the number of steps below.

Using this protocol, it will be shown that the expected number of experimental steps to finish job 1 for  $N = 25$  chromosomes is 95.

After Job 1 is completed, we can distinguish each individual chromosome by its single probe. We take advantage of this in the next job.

## Job 2: $m$ probes per chromosome

In our experiments, we can continue to analyze one probe a time, or alternatively, many probes can be analyzed simultaneously. The economics of scale suggest that adding multiple probes simultaneously is much more practical. Given a batch of  $l$  sample probes, it makes sense to analyze the result of all  $l$  sample probes at once and to either reach a conclusion or decide to continue sampling only after all  $l$  sample probes have been assigned. In this job, our objective is to assign additional  $m - 1$  probes to every chromosome.

We focus our attention on an individual chromosome. All chromosomes will be similarly analyzed. Suppose that we consider chromosome  $i$ , distinguished by its success probe  $c_i$ . Color  $c_i$  with one color, e.g., red. We use the following experimental strategy.

1. Randomly choose  $l$  new “sample probes” and color them with another color, e.g., green. Put the sample probes together with success probe  $c_i$  in one experiment. If one or more of the sample probes occur on the same chromosome as  $c_i$ , then one chromosome in the experiment will have one red probe and one or more green probes hybridized to it.
2. If we detect *exactly one* sample probe on chromosome  $i$ , we can identify it in  $\lceil \log_C(l) \rceil$  additional experiments as explained below. (Notice that we can distinguish between one and several green probes on a chromosome because we assume that there will be a detectable gap between the positions of hybridized probes.)  
Otherwise, discard these sampled probes.
3. Continue from step 1 until  $m - 1$  extra probes are assigned to chromosome  $i$ .

In step 2, we need to identify the sample probe  $c_x$  on chromosome  $i$ . We evenly divide the  $l$  probes into  $C$  groups, and give each of the first  $C - 1$  groups a different color. The probes in the last group are not colored. We color probe  $c_i$  with the remaining color. One of the  $C$  groups must contain  $c_x$ . We put all the colored probes in one experiment to see which group contains  $c_x$ . If none of the colored sample probes belongs to the same chromosome as  $c_i$ , then  $c_x$  must belong to the uncolored group. We then discard all the probes in the other groups. We repeat this procedure by dividing the chosen group again into  $C$  smaller groups until we identify  $c_x$ . In this way, we assign one additional probe to chromosome  $i$  in  $\lceil \log_C(l) \rceil$  steps.

We repeat this process until we assign  $m$  probes to each chromosome.

It is easy to see that using one color and  $l = 1$  we can also finish job 2. In general, if  $l$  is too small or too large, we will need more experimental steps to obtain exactly one new probe on chromosome  $i$ . Also if  $l$  is large, we need more experimental steps to identify the sample probe on chromosome  $i$ . Therefore the number of probes  $l$  is the most important parameter in this process. In the *Mathematical analysis* section, we will prove that the total expected

number of experimental steps, including job 1 and job 2, is

$$N \sum_{k=1}^N 1/k + N(m-1)(1/p + \lceil \log_C(l) \rceil)$$

where  $p = \frac{l}{N}(1 - \frac{1}{N})^{l-1}$ .

## Simulations

Using the above formula we can obtain the expected number of experimental steps using (a).  $C = 2$  and (b).  $C = 3$  different colors to assign at least  $m = 5$  probes to  $N = 25$  chromosomes for different values of sample probe pooling size  $l$  (Table 1).

To see the distribution for the number of experimental steps, we simulate the first strategy. For illustration, we take  $N = 25$ ,  $m = 5$  and  $C = 3$ . We choose  $l = 8, 9, 10$  in our three simulations. Figure 1(abc) show the histograms for the number of experimental steps. Through simulation results, we find that for  $l = 8$ , the mean for the number of experimental steps is 711 and standard deviation (stdv) is 47. For  $l = 9$ , mean = 680 and stdv = 45. For  $l = 10$ , mean = 756 and stdv = 43. We see that  $l$  determines the efficiency of our experiment, with  $l = 9$  being optimal.

/\* Note to the editor: Insert Table 1 here. \*/

/\* Note to the editor: Insert Figure 1 here. \*/

## 3.2 Mathematical Analysis

In this section, we study the number of experimental steps to assign at least  $m$  probes to each chromosome using the above protocol. We examine the two jobs separately. Let  $T_1$  and  $T_2$  be the number of experimental steps in job 1 and job 2 respectively.

**Job 1:** The number of experimental steps  $T_1$  in job 1 corresponds to the waiting time before a coupon collector obtains a whole collection of coupons in the *coupon collector's problem* (Feller 1968). Here “coupons” correspond to probes and “figures” correspond to chromosomes. In our first theorem, we give results on the expectation, variance, and limit distribution for  $T_1$  [Feller, 1968, Baum and Billingsley, 1969, Klaassen, 1994].

**Theorem 3.1** *Let  $T_1$  be the number of experimental steps in job 1. Then*

1. *The expectation of  $T_1$  is*

$$E(T_1) = N \sum_{k=1}^N 1/k \approx N \log(N)$$

2. The variance of  $T_1$  is

$$\text{Var}(T_1) = N \sum_{k=1}^N (N-k)/k^2 \approx N(\pi^2 N/6 - \log(N))$$

3. As  $N \rightarrow \infty$ ,  $Z = \exp(-(T_1/N) + \log(2N))$  converges to an exponential distribution with mean 2.

**Note 1:** Even when  $N$  is relatively small, such as  $N \geq 10$ , the approximations in the above theorem are still very good.

**Note 2:** From Theorem 3.1 (3), we can approximate the distribution of  $T_1$  by

$$F_1(t) = P\{T_1 < t\} \approx \exp(-N \exp(-t/N))$$

The derivation of this approximation is straightforward.

**Job 2:** For a fixed chromosome, let  $X$  be the number of experimental steps to assign one probe to it using the above protocol. Each probe belongs to the chromosome with probability  $1/N$ . Because we put  $l$  sample probes in one experiment, the number of sample probes on the chromosome is binomially distributed with success probability  $1/N$ , denoted by  $B(l, 1/N)$ . Therefore exactly one of the  $l$  probes belongs to the chromosome with probability  $l/N(1 - 1/N)^{l-1}$ . The number of experimental steps  $X^*$  to obtain exactly one of the  $l$  probes belonging to the chromosome is geometrically distributed with success probability  $p = l/N(1 - 1/N)^{l-1}$ , denoted by  $G(p)$ . It takes  $\lceil \log_C(l) \rceil$  experimental steps to identify the probe belonging to the chromosome. Therefore

$$X = X^* + \lceil \log_C(l) \rceil$$

where  $X^*$  is geometrically distributed with success probability  $p$ . Therefore

$$E(X) = E(X^*) + \lceil \log_C(l) \rceil = 1/p + \lceil \log_C(l) \rceil$$

and

$$\text{Var}(X) = \text{Var}(X^*) = (1-p)/p^2$$

Let  $X_{ij}$ ,  $1 \leq i \leq N$ ,  $1 \leq j \leq m-1$  be the number of steps to assign the  $j$ -th probe to chromosome  $i$  in job 2. Then according to our strategy,  $X_{ij}$  are independent identically distributed (iid) random variables and have the same distribution as  $X$  and

$$T_2 = \sum_{i=1}^N \sum_{j=1}^{m-1} X_{ij}$$

From the above discussions, we have the following result.

**Theorem 3.2** *Let  $T_2$  be the number of experimental steps in job 2 and  $p = \frac{l}{N}(1 - \frac{1}{N})^{l-1}$ . Then*

1. *The expectation of  $T_2$  is  $N(m - 1)(1/p + \lceil \log_C(l) \rceil)$ .*
2. *The variance of  $T_2$  is  $N(m - 1)(1 - p)/p^2$ .*
3. *As  $N \rightarrow \infty$ ,  $(T_2 - E(T_2))/\sqrt{\text{Var}(T_2)}$  is approximately  $N(0, 1)$ .*

Let  $T$  be the total number of experimental steps to finish our job. Then  $T = T_1 + T_2$  and the expectation and variance of  $T$  are:

$$E(T) = E(T_1) + E(T_2) = N \sum_{k=1}^N 1/k + N(m - 1)(1/p + \lceil \log_C(l) \rceil)$$

and

$$\text{Var}(T) = \text{Var}(T_1) + \text{Var}(T_2) = N \sum_{k=1}^N (N - k)/k^2 + N(m - 1)(1 - p)/p^2$$

When  $N$  is relatively large (such as  $N \geq 10$ ),  $p$  can be reasonably approximated by  $\frac{l}{N} \exp(-\frac{l}{N})$ .

## 4 The Second Strategy

### 4.1 The Experimental Protocol

As in the first strategy, we divide the experiment into two jobs. First we assign one probe to each chromosome. Then we assign  $m - 1$  extra probes to each chromosome. We try here to improve both jobs.

#### Job 1: One probe per chromosome

First let us improve job 1. In the first strategy, we put one sample probe along with the “success probes” into each experiment. This is reasonable for the first few steps because each time, we have high probability of obtaining a new “success probe”. But at the late stage of this job, much time is wasted waiting for a new success probe. We can design our experiment in such a way so that the number of “sample probes” depends on the number of “success probes” we have obtained. Based on a mathematical analysis that we describe below, our improved strategy is as follows: Let  $k$  be the number of “success probes”.

1. Randomly choose a probe and denote it as  $c_1$ . We call  $c_1$  a “success probe”. Set  $k = 1$ .
2. Randomly sample  $L = \lceil \log_{k/N}(A(C)) \rceil$  “sample probes”, where  $A(C)$  is the solution for the equation

$$x \log(x) \log(C + 1) + (1 - x)^2 = 0$$

( $A(1) \approx 0.513$ ,  $A(2) \approx 0.364$ ,  $A(3) \approx 0.294$ ,  $A(4) \approx 0.252$ .) Color all the  $k$  “success probes” and the “sample probes” with one color. Put them into one experiment to see overlapping patterns.

3. If the probes including the “success probes” and the “sample probes” belong to more than  $k$  chromosomes, then at least one “sample probe” does not belong to the same chromosome as any of the “success probes”. Identify one such probe in  $\lceil \log_{C+1}(L) \rceil$  experimental steps (see below). Then add it to the group of “success probes”. Set  $k = k + 1$ .

Otherwise discard all these sample probes.

4. Continue from step 2 until  $k = N$ , i.e., all the chromosomes are assigned one probe.

In step 3, to identify one probe that does not belong to the same chromosome as any of the success probes, we can evenly divide the “sample probes” into  $C + 1$  groups. Color each of the first  $C$  groups with a different color and do not color the probes in the last group. Color the  $k$  success probes with one of the  $C$  colors, e.g., red. Put the “sample probes” along

with the “success probes” into one experiment. If the red probes, including the  $k$  “success probes” and the red “sample probes”, belong to more than  $k$  chromosomes, then one of the red “sample probes” must not be on the same chromosome as any of the  $k$  “success probes”. Else if all the red probes belong to  $k$  chromosomes and at least one colored sample probe with color other than red is on a different chromosome from any of the red probes, then the group of probes containing that sample probe can be identified. If the above two situations do not happen, the group of un-colored sample probes must contain at least one probe not belonging to the same chromosome as any of the  $k$  success probes.

In this way we can identify one of the  $C + 1$  groups containing a sample probe not belonging to the same chromosome as any of the “success probes”. Then we evenly divide this group into  $C + 1$  groups again. Continue this process until we identify one sample probe not belonging to the same chromosome as any of the  $k$  success probes. This requires at most  $\lceil \log_{C+1}(L) \rceil$  experimental steps.

Using this strategy, the expected number of experimental steps in job 1 are 63, 54, and 48 using  $C = 1, 2,$  and  $3$  colors respectively for  $N = 25$  chromosomes. Note that in the first strategy, job 1 requires 95 steps compared to 63 steps in the second strategy using one color, a save of over 30%.

## Job 2: $m$ probes per chromosome

In the first strategy, we usually need to wait a long time to see exactly one probe on a fixed chromosome. During the waiting time, the experiments are wasted. To overcome this problem, we pool both the “success probes” and “sample probes” simultaneously.

As in the first strategy, suppose we have already finished job 1 and labeled each chromosome by a “success probe”. Then we assign  $m - 1$  extra probes to each chromosome. Our second strategy is as follows:

1. Randomly choose  $S < C$  “success probes” from the set of all the success probes and color each with a different color.
2. Randomly choose  $L$  “sample probes” and color them with one color which is different from the colors used in step 1.
3. Put the  $S$  “success probes” and  $L$  “sample probes” into one experiment and observe the number of “sample probes” on each of the  $S$  chromosomes labeled by the  $S$  “success probes”.
4. Identify all the “sample probes” on each of the  $S$  labeled chromosomes. This takes at most  $r \lceil \log_C(L/C) \rceil + 1$  experimental steps, where  $r$  is the number of probes on the chromosome (see Lemma 4.1).

5. Continue from step 2 until  $k \geq 1$  of the chosen chromosomes has been assigned at least  $m$  probes. Remove the success probes for these chromosomes from the set of  $S$  success probes. Then randomly choose  $k$  new “success probes” to replace them. If less than  $k$  success probes remain, choose all the remaining success probes.
6. Continue from step 2 until all the chromosomes have been assigned at least  $m$  probes.

## Simulations

We used simulations to test the second strategy. Table 3(ab) give upper bounds for the expected total number of experimental steps, including job 1 and job 2, to assign  $m=5$  probes to  $N=25$  chromosomes using (a)  $C = 2$  and  $C = 3$  colors and different success probe pooling sizes  $S$  and sample probe pooling sizes  $L$  according to the above protocol. From Table 3b we see that for  $C = 3$ , using sample probe pooling size  $L = 9$  gives the smallest upper bound for any values of success probe pooling size  $L$ . For  $C = 3$ ,  $S = 2$  and  $L = 9$ , we only need on average at most 395 steps to finish the job, a reduction of about 40% compared to the first strategy which used 680 steps using  $C = 3$  colors. To see the distribution for the number of experimental steps, we also give the histograms for the upper bounds for  $C = 3$ ,  $L = 9$  and (a).  $S = 1$  and (b).  $S = 2$  (Figure 2ab).

/\* Note to the editor: Insert Table 2 here. \*/

/\* Note to the editor: Insert Figure 2 here. \*/

## 4.2 Mathematical Analysis

We study the number of experimental steps in the two jobs separately.

**Job 1:** First we develop the formula we used for the sample probe pooling size given that there are  $k$  success probes.

Suppose we have already obtained  $k$  “success probes”. Then we sample a pool of  $L$  probes randomly. Each sample probe does not belong to the same chromosome as any of the  $k$  “success probes” with probability  $(N - k)/N = 1 - k/N$ . Because the  $L$  sample probes are sampled independently, the number of sample probes that do not belong to the same chromosome as any of the “success probes” is binomially distributed with success probability  $1 - k/N$ , denoted by  $B(L, 1 - k/N)$ . Therefore the probability that there is *at least one* such probe in the set of  $L$  sample probes is  $1 - (k/N)^L$ . The number of experimental steps  $X_k$  until we see at least one such probe is geometrically distributed with success probability  $1 - (k/N)^L$ , denoted by  $G(1 - (k/N)^L)$ . Once we see at least one such probe, we need  $\lceil \log_{C+1}(L) \rceil$  experimental steps to identify it, thus gaining a new success probe. Therefore,

the number of experimental steps  $T_k$  to identify a new “success probe” is

$$T_k = X_k + \lceil \log_{C+1}(L) \rceil$$

and the expectation of  $T_k$  is

$$E(T_k) = E(X_k) + \lceil \log_{C+1}(L) \rceil = (1 - (k/N)^L)^{-1} + \lceil \log_{C+1}(L) \rceil$$

We want to minimize the expected number of experimental steps in going from  $k$  “success probes” to  $k + 1$  “success probes”. Differentiating  $E(T_k)$  with respect to  $L$  (thinking of  $L$  as a continuous variable) we have

$$(E(T_k))'_L = \frac{A \log(A) \log(C + 1) + (1 - A)^2}{L \log(C + 1)(1 - A)^2}$$

where  $A = (k/N)^L$ . Letting  $(E(T_k))'_L = 0$ , we have

$$A \log(A) \log(C + 1) + (1 - A)^2 = 0 \tag{1}$$

and  $L = \log_{k/N}(A)$ . Because  $L$  needs to be an integer, we take  $L = \lceil \log_{k/N}(A) \rceil$  which gives the formula we used in the experimental protocol. To emphasize the dependency of  $L$  on  $k$ , we denote it as  $L_k$  below.

The total expected number of experimental steps in job 1 is

$$\sum_{k=1}^{N-1} E(T_k) = \sum_{k=1}^{N-1} (1 - (k/N)^{L_k})^{-1} + \sum_{k=1}^{N-1} \lceil \log_{C+1}(L_k) \rceil$$

**Job 2:** Next, we study the upper bound given in step 4 of job 2. Suppose that we have  $L$  probes and  $r$  of them belong to a specific chromosome. The next lemma gives an upper bound for the number of experimental steps to identify these  $r$  probes.

**Lemma 4.1** *Assume that in a pool of  $L$  probes,  $r$  of them belong to a specific chromosome. Using  $C$  colors, we can identify the  $r$  probes in at most  $r \lceil \log_C(L/C) \rceil + 1$  steps.*

**Proof:** We prove this lemma by induction on  $L$ .

(i). The lemma is obviously true for  $L \leq C$ .

(ii). Suppose the lemma is true for  $L \leq l - 1 \geq C$ . Next we prove that the lemma is true for  $L = l$ . First we evenly divide the  $l$  probes into  $C$  groups and color each of the first  $C - 1$  groups with a different color. The probes in the last group are not colored. Then color the success probe corresponding to the chromosome with the other color. Put these colored probes along with the success probe into one experiment to determine the groups containing

at least one of the  $c$  probes. For each group  $i$  containing at least one of the  $r$  probes, using induction, we can identify the probes in at most  $r_i \lceil \log_C(\lceil l/C \rceil / C) \rceil + 1$  steps, where  $r_i > 0$  is the number of probes in group  $i$  belonging to the specific chromosome. Therefore we can identify all of them in at most

$$\sum_{i, r_i \neq 0} (r_i \lceil \log_C(\lceil l/C \rceil / C) \rceil + 1) + 1 \leq r \lceil \log_C \lceil l/C \rceil / C \rceil + r + 1 \leq r \lceil \log_C(l/C) \rceil + 1$$

for  $l > C$ .

■

Next we heuristically analyze the number of experimental steps needed to assign  $m$  probes to each chromosome for  $S = 1$ . The analysis for  $S > 1$  is complicated. In each experiment, we sample  $L$  probes. The number of probes  $R$  on the chromosome is binomially distributed with success probability  $1/N$ , denoted by  $B(L, 1/N)$ . Therefore the probability that there are no probes on the chromosome is  $(1 - 1/N)^L$ . The waiting time to obtain at least one probe on the chromosome is geometrically distributed with success probability  $p_1 = 1 - (1 - 1/N)^L$ , denoted by  $G(p_1)$ . The expected number of steps until we obtain at least one probe on the chromosome is  $1/p_1$ . Given that there is at least one probe on the chromosome, the distribution of the number of chromosomes,  $R$ , is

$$P\{R = k | R > 0\} = \binom{L}{k} (1/N)^k (1 - 1/N)^{L-k} / p_1, \quad k = 1, 2, \dots, L$$

Therefore

$$E(R | R > 0) = \frac{L}{N p_1}$$

Because we need at most  $R \lceil \log_C(L/C) \rceil + 1$  experimental steps to identify the  $R$  sample probes, the expected number of steps to identify these probes is at most

$$\frac{L}{N p_1} \lceil \log_C(L/C) \rceil + 1$$

Therefore the total number of experimental steps to assign an average of  $\frac{L}{N p_1}$  probes to the chromosome is at most

$$\frac{1}{p_1} + \frac{L}{N p_1} \lceil \log_C(L/C) \rceil + 1$$

We refer to the set of experiments to find and identify at least one sample probe belonging to the chromosome using the above protocol as *a round* of experiments. Because we want to assign  $m - 1$  probes to the chromosome, we need  $(m - 1) / E(R | R > 0) = (m - 1) N p_1 / L$  rounds of experiments. Therefore we need at most

$$\left( \frac{1}{p_1} + \frac{L}{N p_1} \lceil \log_C(L/C) \rceil + 1 \right) \frac{(m - 1) N p_1}{L} = (m - 1) \left( \frac{N}{L} + \lceil \log_C(L/C) \rceil + \frac{N p_1}{L} \right)$$

experimental steps. The expected number of experimental steps to assign at least  $m - 1$  probes to all the chromosomes is

$$N(m - 1) \left( \frac{N}{L} + \lceil \log_C(L/C) \rceil + \frac{Np_1}{L} \right)$$

To optimize this strategy, we only need to choose the minimum point  $L$  of this function.

## 5 The Third Strategy

### 5.1 The Experimental Protocol

This strategy depends on a new set of ideas. It also attempts to conform to more realistic assumptions about the kind of experiments that can actually be performed. In a real experiment, both the number of colors and the number of probes should be small.

#### Dealing with overlaps

We begin with a few definitions:

**Definition 5.1** *When two colored probes appear on the same chromosome, we have an **overlap**. When the identity of both probes is unknown, we have an **unresolved overlap**. When we use further experiments to identify the unknown probes in an unresolved overlap we are **resolving the overlap**.*

**Definition 5.2** *A set of probes is **independent** if each probe in the set occurs on a separate chromosome. Independent probes can not overlap.*

In the rest of this section, we will use the following protocols for combining colored probes:

**Protocol 1:** An experiment contains  $2C$  probes and  $C$  colors, with two probes per color.  $C$  of the probes are independent, that is the  $C$  probes belong to  $C$  different chromosomes and can not overlap. Let each color contain exactly one of these  $C$  probes and one other probe.

**Protocol 2:** An experiment contains  $3C$  probes and  $C$  colors, with three probes per color.  $2C$  of the probes are independent. Each color contains exactly two of these  $2C$  probes and one other probe.

#### The sieve

As in the other strategies, we have two jobs. The first is to assign one probe to each chromosome. The next is to assign at least  $m - 1$  additional probes to each chromosome. Assume that we have completed job 1 and let  $M$  be the set of probes, one per chromosome that we obtained in job 1.

Our goal is to decrease the number of experiments by:

1. decreasing the number of probes that we examine, and/or

2. increasing the number of overlaps detected with a single experiment.

A key idea for the third strategy is that we start with a fixed pool  $P$  of probes and use the probes in  $M$  to sieve through this pool. Following the protocols, we choose either  $C$  or  $2C$  “success probes” from  $M$  and find all the overlaps between the success probes and the pool probes. Then, we resolve the overlaps simultaneously. Then, we **remove from the pool the pool probes that participated in an overlap**. Thus we remove all those probes from the pool that fall on the chromosomes labeled by the subset of success probes. This constitutes one pass of the pool through the sieve. We perform multiple passes, each time with a new set of success probes. Eventually, each of the success probes “catches” all of its overlapping pool probes.

The sieve achieves both points mentioned above. As the pool shrinks, the number of probes we examine in each pass decreases. At the same time, the number of overlaps that are detected by an experiment increases because as the number of chromosomes represented by probes in the pool decreases, the number of expected overlaps per experiment *of fixed size* increases.

Let  $M$  be the set of probes, one per chromosome obtained in job 1. We will use either protocol 1 or 2. Within one experiment, let

$$\begin{aligned} k &= \text{the number of probes that can } \textit{not} \text{ overlap given the protocol;} \\ r &= \text{the number of probes randomly chosen from the pool.} \end{aligned}$$

In protocol 1,  $k = C$ , and in protocol 2,  $k = 2C$ . In both protocols,  $r = C$ .

### The Sieve Procedure

1. Randomly select  $T$  pool probes to form a pool  $P$ ;
2. Choose  $k$  “success probes” from  $M$ . In protocol 1, each color gets one success probe. In protocol 2, each color gets two success probes. Randomly choose  $r$  probes from  $P$ . Each color gets one of the pool probes. Run an experiment and look for overlaps. Any overlaps that are detected are noted and the probes set aside to resolve later.
3. Repeat step 2 using the same  $k$  “success probes” and a new set of  $r$  probes from  $P$  until the probes in the pool are exhausted.
4. Resolve all the overlaps obtained in the repetition of step 2. With protocol 1, we can resolve one overlap using at most  $5(C - 1)/(C * (3C - 1))$  steps on average. With protocol 2, we can resolve one overlap using at most  $4(3C - 2)/(C * (5C - 1))$  steps on average as shown in the *Mathematical Analysis* section. When an overlap occurs between a pool probe and a “success probe,” **remove the pool probe from the pool**. Do nothing if two pool probes overlap.

5. Repeat steps 2, 3, and 4 using a new set of  $k$  “success probes” until all the success probes are exhausted.

Before we present the third strategy, we explain our intuition. Suppose we start with a pool large enough to contain  $m-1$  probes per chromosome. We could then use the sieve procedure once (with  $T$  equal to the entire pool size) and achieve job 2. There are two problems with this approach:

1. In order to have high probability that the pool has  $m-1$  probes per chromosome would require a pool larger than necessary most of the time.
2. In a large pool, a majority of the chromosomes will have many more than the required  $m-1$  probes.

We can solve the first problem by using the sieve procedure several times, each time with a relatively small number  $T$  of new pool probes. When all of the chromosomes finds  $m-1$  probes we can stop. In this way, our procedure (and the analysis) is geared to the expected case rather than the worst case.

This does not solve the second problem, because we will still have too many probes for many of the chromosomes. So let us make a second modification. Each time we run the sieve procedure, we check if any of the chromosomes have gathered the required  $m-1$  probes. For any that have, **we remove their success probes from the set  $M$** . The next run of the sieve therefore uses a smaller set  $M$ . There is a tradeoff here. If we do not use all the chromosomes to screen the  $T$  pool probes, then we will have some probes in  $T$  that never see their success probe. Call these the “lost probes”. The lost probes will persist until the end of the procedure and that means that all of the remaining probes in  $M$  will have to screen them. On the other hand, we are reducing the number of overlaps that we cannot really use (*i.e.*, the overlaps involving the lost probes) and resolving the overlaps is expensive. We will show in the analysis that the tradeoff works in our favor, making this a good strategy.

## **Job 2: $m$ probes per chromosome**

We assume that Job 1 has been completed. Later, we will show how this can be accomplished in the context of performing job 2.

1. Run the sieve procedure with  $T$  randomly sampled probes.
2. Remove from  $M$ , the success probes corresponding to those chromosomes that have already obtained  $m-1$  probes. Repeat from step 1 until all the chromosomes have been assigned  $m-1$  probes.

We can improve step 4. If two probes from the pool overlap, we can remove one of them and remember its representative so that when the representative is assigned to a chromosome, all the other probes it represents are also correctly assigned. Such overlaps occur and are resolved and there is no reason to throw away this information since by using it we can further reduce the pool size and thus further reduce the number of experiments. Because of its complications, though, we will not analyze this modification in detail.

## Simulations

We ran simulations to test the third strategy. For job 1, we used the same method as in *The Second Strategy*. In job 2, we used both Protocol 1 ( $k = r = 3$ ) and Protocol 2 ( $k = 6, r = 3$ ). Table 3 a (Protocol 1) and b (Protocol 2) give the simulated average number of experimental steps, including job 1 and job 2, to assign  $m = 5$  probes to all the  $N = 25$  chromosomes by assigning  $T$  probes a time. From this table we see that using three colors and the second protocol, we can finish the job in about 308 steps by assigning 50 probes a time, a saving about 25% compared to 395 steps using the second strategy.

Figure 3ab give the histogram for the number of experiments by assigning  $T = 50$  probes a time using the third strategy, with (a) Protocol 1 and (b) protocol 2 and  $C = 3$ .

/\* Note to the editor: Insert Table 3 here. \*/

/\* Note to the editor: Insert Figure 3 here. \*/

## Job 1 again

We return once more to job 1. Here, we show that **we do not have to do a preliminary search to find one probe per chromosome**. The sieve does this for us automatically.

Again, let  $M$  be the set of success probes found in job 1. Since we do not perform job 1, we start with  $M$  empty. We are going to add probes to  $M$  until every chromosome has one probe in  $M$ . Then as in the third strategy, we will remove a probe from  $M$  if its chromosome has found an additional  $m - 1$  probes. Note though, that **no probe is removed from  $M$  until every chromosome has one probe in  $M$** .

The idea is the following. We use only pool probes, but some of these become success probes. Suppose wolog that we are using protocol 1. We would like to start with a set of  $k$  success probes and  $r$  other probes. But, we do not yet have  $k$  success probes. So, we do the following.

1. Choose  $k$  unused success probes from  $M$ . If there are not enough unused success probes in  $M$ , choose as many as are present. Choose  $r$  pool probes. Color as in protocol 1, except some colors may not get their independent probe. Run the experiment and look for overlaps.

2. There are two possible outcomes:

- The experiment contained  $\geq k$  independent probes. Add the new independent probes to  $M$ , mark them as unused and complete the round using the first  $k$  unused independent probes in  $M$  as we would a round in the sieve procedure.
- We find  $n < k$  independent probes. Now, here is the beauty of this technique. It was okay to find less than  $k$  probes because **all the probes in the experiment fell on the  $n$  chromosomes represented by the  $n$  independent probes**. In other words, we didn't let any pool probes slip by! Add the  $n$  independent probes to  $M$ , mark them unused and repeat from 1.

3. After the round is over, that is, after all the pool probes have been screened by this set of independent probes, mark them used. If any probes remain in the pool, repeat from 1.

4. If each chromosome has one probe in  $M$  then stop, otherwise choose a new pool of probes, mark every probe in  $M$  unused and repeat from 1.

Once  $M$  is full, we continue with the unmodified third strategy, now free to remove probes from  $M$  if a chromosome finds  $m - 1$  additional probes.

It is difficult to analyze this modification in detail because we start from different structures for each probe pool. Note here that we only need at most  $k + r$  probes ( $= 2k$  probes for protocol 1) in each experiment, while in job 1 of the first two strategies, we need many more probes at the late stage. The overlap information we obtain while completing job 1 can also be used in job 2 to further reduce the number of experiments there.

## 5.2 Mathematical Analysis

### Resolving overlaps

First, we present some theorems about resolving overlaps under protocol 1 and protocol 2. First note that overlaps occur that do not need to be resolved. In protocol 1, overlaps between probes with the same color do not need to be resolved.

#### Protocol 1

**Lemma 5.3** *Let an experiment be conducted under protocol 1. Let an unresolved overlap occur between two probes of different colors. Then we can resolve the overlap with 2 probes and 1 color in two experiments.*

**Proof:** Since one of the four possible combinations is prohibited, we need to resolve only three combinations. Using 2 probes and 1 color, we test each of these combinations. If after

the second experiment we have not found the overlap, then it is the third combination by default. ■

**Lemma 5.4** *Let an experiment be conducted under protocol 1. Let the distribution of overlaps between probes (of those that can overlap) be uniform. Let an unresolved overlap occur between two probes. Then the expected number of experiments to resolve the overlap using 2 probes and 1 color according to the strategy in Lemma 5.3 is  $5(C - 1)/(3C - 1)$ .*

**Proof:** Using  $2C$  probes, there are  $\binom{2C}{2}$  possible combinations.  $\binom{C}{2}$  of these occur between independent probes and are therefore prohibited, leaving  $C(3C - 1)/2$  ( $= \binom{2C}{2} - \binom{C}{2}$ ). Of these,  $C$  occur between probes of the same color and require no additional experiments to resolve. The remaining  $3C(C - 1)/2$  ( $= C(3C - 1)/2 - C$ ) occur between probes of different colors and require additional steps to resolve them. With probability  $1/3$ , we can resolve the overlap in one step and with probability  $2/3$ , we resolve the overlap in two steps. Therefore the expected number of experiments is

$$\frac{2C}{C(3C - 1)} * 0 + \frac{3C(C - 1)}{C(3C - 1)}((1/3) * 1 + (2/3) * 2) = \frac{5(C - 1)}{3C - 1}.$$

■

**Theorem 5.5** *Using  $C$  colors and  $2C$  probes, we can resolve  $C$  overlaps under protocol 1 in two experiments. The expected cost to resolve an overlap is  $(1/C) \cdot 5(C - 1)/(3C - 1)$ .*

**Proof** According to the strategy in Lemma 5.3, we can resolve one overlap in at most two experiments using 2 probes and 1 color. Using  $2C$  probes and  $C$  colors, we can resolve  $C$  overlaps in at most two experiments. The expected number of experiments running  $c$  overlaps per experiment is  $(1/C) \cdot 5(C - 1)/(3C - 1)$  from Lemma 5.4.

## Protocol 2

**Lemma 5.6** *Let an experiment be conducted under protocol 2. Let an unresolved overlap occur between two probes of the same color. Then we can resolve the overlap with 2 probes and 1 color in one experiment.*

**Proof:** There are two possible combinations. Resolving them takes one experiment. ■

**Lemma 5.7** *Let an experiment be conducted under protocol 2. Let an unresolved overlap occur between two probes of different colors. Then we can resolve the overlap with 3 probes and 1 color in three experiments.*

**Proof:** Let the unresolved overlap involve the colors RED (R) and GREEN (G). Each color has three probes,  $R_A, R_B, R_C$ , and  $G_A, G_B, G_C$ . Let  $R_B, R_C, G_B$  and  $G_C$  be the independent probes that can not overlap. Of the nine possible combinations with 3 probes of one color and 3 probes of another, four are prohibited because they are overlaps between independent probes. Thus we have five possible combinations. Proceed as follows:

Experiment 1: Color  $G_A, R_B$  and  $R_C$  the same color. If no overlap occurs, go to experiment 2. Otherwise, we have two possible combinations, either  $G_A$  and  $R_B$  or  $G_A$  and  $R_C$ . These can be resolved with 2 probes and one color in one additional experiment for a total of two experiments for these combinations.

Experiment 2: Color  $R_A, G_B$  and  $G_C$  the same color. If no overlap occurs, then the overlap is between  $G_A$  and  $R_A$  by default. (Giving a total of two experiments for this combination.) Otherwise, as above, we have two possible combinations, either  $R_A$  and  $G_B$  or  $R_A$  and  $G_C$ . These can be resolved with 2 probes and one color in one additional experiment, for a total of three experiments for these combinations.

■

**Lemma 5.8** *Let an experiment be conducted under protocol 2. Let the distribution of overlaps between probes (of those that can overlap) be uniform. Then the expected number of experiments to resolve an overlap using 3 probes and 1 color is  $\frac{4(3C-2)}{5C-1}$ .*

**Proof:** With  $3C$  probes, there are  $\binom{3C}{2}$  possible combinations. Of these,  $\binom{2C}{2}$  occur between independent probes and are therefore prohibited, leaving  $C(5C-1)/2$ .  $2C$  occur between probes of the same color and require one experiment to resolve. The remainder,  $5C(C-1)/2$ , occur between probes of different colors. With probability  $3/5$ , we resolve the overlap in two experiments and with probability  $2/5$  we resolve the overlap in three experiments. Therefore the expected number of experiments is

$$\frac{4C}{C(5C-1)} * 1 + \frac{5C(C-1)}{C(5C-1)} * \left( \frac{3}{5} * 2 + \frac{2}{5} * 3 \right) = \frac{4(3C-2)}{5C-1}$$

■

**Theorem 5.9** *Using  $C$  colors and  $3C$  probes we can resolve  $C$  overlaps under protocol 2 in at most three steps. The expected cost to resolve an overlap is  $\frac{4(3C-2)}{C(5C-1)}$ .*

**Proof:** The proof is the same as for theorem 5.5. ■

## The expected number of experimental steps

Suppose we have  $T_1 + T_2 + \cdots + T_N = T$  pool probes, with  $T_i$  the number of probes on chromosome  $i$ . Given  $T$ , then  $(T_1, T_2, \cdots, T_N)$  has multinomial distribution with probability distribution

$$P\{T_1 = k_1, T_2 = k_2, \cdots, T_N = k_N\} = \frac{T!}{k_1!k_2!\cdots k_N!} \left(\frac{1}{N}\right)^T$$

where  $k_1 + k_2 + \cdots + k_N = T$ . The expectation and variance of  $T_i$  are given by

$$E(T_i) = T/N, \quad Var(T_i) = (N-1)T/N^2$$

Before we analyze the third strategy, let us first give a lemma about the multinomial distribution.

**Lemma 5.10** *Let  $(X_1, X_2, \cdots, X_N)$  have multinomial distribution,  $Multinomial(T; p_1, p_2, \cdots, p_N)$ , i.e.*

$$P\{X_1 = k_1, X_2 = k_2, \cdots, X_N = k_N\} = \frac{T!}{k_1!k_2!\cdots k_N!} p_1^{k_1} p_2^{k_2} \cdots p_N^{k_N}$$

Then

$$E \frac{X_i(X_i - 1)}{X_{j+1} + X_{j+2} + \cdots + X_N - 1} = \frac{Tp_i^2}{\sum_{\mu>j} p_\mu} \quad i > j$$

In particular, if  $p_1 = p_2 = \cdots = p_N = 1/N$ , then

$$E \frac{X_i(X_i - 1)}{X_{j+1} + X_{j+2} + \cdots + X_N - 1} = \frac{T}{N(N-j)} \quad i > j$$

**Proof:** First note that if the numerator is not zero, the denominator can not be zero either. Therefore the above equation is well-defined. Let  $S_j = \sum_{\mu>j} X_\mu$ . Then the joint probability distribution of  $(X_i, S_j)$  is given by

$$P\{X_i = l_1, S_j = l_1 + l_2\} = \frac{T!}{l_1!l_2!(T - l_1 - l_2)!} \left(\sum_{\mu \leq j} p_\mu\right)^{T-l_1-l_2} (p_i)^{l_1} \left(\sum_{\mu>j, \mu \neq i} p_\mu\right)^{l_2}$$

The joint probability generating function of  $(X_i, S_j)$  is

$$\begin{aligned} f(x, s) &= Ex^{X_i} s^{S_j} \\ &= \sum_{l_1+l_2 \leq T} P\{X_i = l_1, S_j = l_1 + l_2\} x^{l_1} s^{l_1+l_2} \\ &= \left( p_i x s + \left( \sum_{\mu>j, \mu \neq i} p_\mu \right) s + \left( \sum_{\mu \leq j} p_\mu \right) \right)^T \end{aligned}$$

Therefore

$$\begin{aligned}
E \frac{X_i(X_i - 1)}{S_j - 1} &= \int_0^1 \frac{d^2 f(1, s)}{dx^2} \frac{1}{s^2} ds \\
&= T(T - 1)p_i^2 \int_0^1 \left( \sum_{\mu > j} p_\mu \right) s + \sum_{\mu \leq j} p_\mu \Big)^{T-2} ds \\
&= \frac{T p_i^2}{\sum_{\mu > j} p_\mu}
\end{aligned}$$

■

Next we give a lemma on the expected number of experimental steps to assign  $T$  randomly chosen probes to  $l$  chromosomes using the sieve procedure. For simplicity, we use the real number  $x$  when the ceiling of  $x$  should be used. This can cause some problems especially when  $x$  is small. But when  $x$  is large, this does not cause any major problem for our approximation.

**Lemma 5.11** *Suppose we first randomly choose  $T$  probes from the  $N$  chromosomes. Then we assign these  $T$  probes to  $l$  randomly chosen chromosomes according to the sieve procedure (i.e., using protocol 1 or 2 with  $k$  success probes and  $r$  probes at a time chosen from pool  $P$ ). Let  $A$  be the average number of experimental steps needed to resolve one overlap. Then the expected number of experimental steps is*

$$\frac{T}{N} \left( A l \left( 1 + \frac{r-1}{2k} \right) + \frac{1}{r} \left( \frac{Nl}{k} - \frac{l}{2} \left( \frac{l}{k} - 1 \right) \right) \right)$$

**Proof:** First we fix  $T_1, T_2, \dots, T_N$ , the number of probes on each chromosome. Randomly choose  $r$  pool probes from the  $T$  pool probes. Then the number of pool probes  $X$  out of the  $r$  randomly chosen probes on a fixed chromosome  $i$  is hyper-geometric with probability distribution

$$P\{X = x\} = \frac{\binom{T_i}{x} \binom{T-T_i}{r-x}}{\binom{T}{r}}$$

The expectation and variance of  $X$  are

$$E(X) = \frac{rT_i}{T}, \quad Var(X) = \frac{rT_i}{T} \left( 1 - \frac{T_i}{T} \right) \left( 1 - \frac{r-1}{T-1} \right)$$

Therefore the number of pairwise overlaps between these pool probes on chromosome  $i$  is

$$E\binom{X}{2} = 1/2(EX^2 - EX) = 1/2(Var(X) + (EX)^2 - EX) = \frac{r(r-1)T_i(T_i-1)}{2T(T-1)}$$

The expected number of pairwise overlaps among pool probes on all the chromosomes is  $\sum_{i=1}^N \frac{r(r-1)T_i(T_i-1)}{2T(T-1)}$ .

On chromosome  $i$ , the expected number of overlaps between the success probe and the pool probes is  $rT_i/T$ . Summing over all the selected success probes  $1, 2, \dots, k$ , we have the number of overlaps between success probes and pool probes  $\sum_{i=1}^k rT_i/T$ .

There are totally  $T$  probes and each time we choose  $r$  pool probes. Therefore we need  $T/r$  times to exhaust all the pool probes. Notice for each randomly chosen  $r$  pool probes, the distribution of overlaps is the same. Therefore, for the first *round*, where a round means the process of assigning the probes to the selected  $k$  chromosomes, the total number of overlaps is

$$\left( \sum_{i=1}^k \frac{rT_i}{T} + \sum_{i=1}^N \frac{r(r-1)T_i(T_i-1)}{2T(T-1)} \right) * \frac{T}{r} = \sum_{i=1}^k T_i + \sum_{i=1}^N \frac{(r-1)T_i(T_i-1)}{2(T-1)}$$

In the first round, we also need  $T/r$  experiments to detect the overlaps. Let  $A$  be the average number of experimental steps to resolve one overlap, then the expected total number of experiments is

$$A \left( \sum_{i=1}^k T_i + \sum_{i=1}^N \frac{(r-1)T_i(T_i-1)}{2(T-1)} \right) + \frac{T}{r}$$

Taking expectation with respect to  $(T_1, T_2, \dots, T_N)$  and using Lemma 5.10, we have the expected total number of experimental steps in the first round

$$A \left( \sum_{i=1}^k ET_i + \sum_{i=1}^N E \frac{(r-1)T_i(T_i-1)}{2(T-1)} \right) + \frac{T}{r} = \frac{T}{N} \left( kA + \frac{(r-1)A}{2} + \frac{N}{r} \right)$$

After the first round, the pool probes on chromosomes  $1, 2, \dots, k$  are assigned to their corresponding chromosomes. We have  $S_k = T_{k+1} + T_{k+2} + \dots + T_N$  pool probes left. We screen these  $S_k$  pool probes with another set of randomly chosen  $k$  chromosomes, without loss of generality, chromosomes  $k+1, k+2, \dots, 2k$ . With the same argument as above, for fixed  $(T_1, T_2, \dots, T_N)$ , the expected number of experimental steps to detect and to resolve the overlaps is

$$A \left( \sum_{i=k+1}^{2k} T_i + \sum_{i=k+1}^N \frac{(r-1)T_i(T_i-1)}{2(S_k-1)} \right) + \frac{S_k}{r}$$

Taking expectation with respect to  $(T_1, T_2, \dots, T_N)$  and using Lemma 5.10, we obtain the expected number of steps in the second round

$$A \left( \sum_{i=k+1}^{2k} ET_i + \sum_{i=k+1}^N E \frac{(r-1)T_i(T_i-1)}{2(S_k-1)} \right) + \frac{ES_k}{r} = \frac{T}{N} \left( kA + \frac{(r-1)A}{2} + \frac{N-k}{r} \right)$$

Continuing this process, we find that the expected number of experimental steps in the  $i+1$ -st round is  $\frac{T}{N} \left( kA + \frac{(r-1)A}{2} + \frac{N-ik}{r} \right)$ . Because we only want to assign probes on chromosomes

$1, 2, \dots, l$ , we need  $l/k$  rounds and the total expected number of experimental steps to assign probes to chromosomes  $1, 2, \dots, l$  would be

$$\frac{T}{N} \left( \left(1 + \frac{r-1}{2k}\right) Al + \frac{1}{r} \sum_{i=0}^{l/k-1} (N - i * k) \right) = \frac{T}{N} \left( Al \left(1 + \frac{r-1}{2k}\right) + \frac{1}{r} \left( \frac{Nl}{k} - \frac{l}{2} \left( \frac{l}{k} - 1 \right) \right) \right)$$

■

Given the above two lemmas, we are now ready to analyze the total number of experimental steps to assign  $\beta T$  probes to the chromosomes ( $\beta$  an integer  $\geq 1$ ) according to the third strategy.

**Lemma 5.12** *The expected number of experimental steps to assign  $\beta T$  probes to the chromosomes according to the third strategy is*

$$T \left( A \left(1 + \frac{r-1}{2k}\right) \sum_{i=0}^{\beta-1} B(iT) + \frac{1}{r} \left( \frac{N}{k} \sum_{i=0}^{\beta-1} B(iT) - \frac{1}{2} \sum_{i=0}^{\beta-1} B(iT) \left( \frac{NB(iT)}{k} - 1 \right) \right) \right)$$

where  $B(iT) = \sum_{j=0}^{m-2} \binom{iT}{j} (1/N)^j (1 - 1/N)^{iT-j}$  and  $B(0) = 1$ .

**Proof:** We first randomly choose  $T$  probes and assign them to the  $N$  chromosomes. The expected number of experimental steps is given by Lemma 5.11 with  $l = N$ . After that, a chromosome  $i$  has not been assigned  $m - 1$  probes with probability  $B(T) = \sum_{j=0}^{m-2} \binom{T}{j} (1/N)^j (1 - 1/N)^{T-j}$ . The expected number of chromosomes that have not obtained  $m - 1$  probes is  $NB(T)$ .

Next we randomly choose another  $T$  probes and assign them to the  $NB(T)$  chromosomes. The expected number of experimental steps is given by Lemma 5.11 with  $l = NB(T)$ . After that, the expected number of chromosomes that have not obtained  $m - 1$  probes is  $NB(2T)$ .

Continuing this process until we assign  $\beta T$  probes to the chromosomes. The expected total number of experimental steps is

$$T \left( A \left(1 + \frac{r-1}{2k}\right) \sum_{i=0}^{\beta-1} B(iT) + \frac{1}{r} \left( \frac{N}{k} \sum_{i=0}^{\beta-1} B(iT) - \frac{1}{2} \sum_{i=0}^{\beta-1} B(iT) \left( \frac{NB(iT)}{k} - 1 \right) \right) \right)$$

■

In order to study the number of experimental steps to assign at least  $m - 1$  probes to all the chromosomes, we need to study the distribution of the stopping time  $\mathcal{I}$  for the third strategy. Clearly,  $\{\mathcal{I} \leq i\}$  if and only if, using the first  $iT$  probes, all the chromosomes have been assigned at least  $m - 1$  probes. Because the number of probes on the chromosomes is multinomial( $iT; 1/N, \dots, 1/N$ ), we have

$$P\{\mathcal{I} \leq i\} = \sum_{\substack{l_1 \geq m-1, \dots, l_N \geq m-1, \\ l_1 + \dots + l_N = iT}} \frac{(iT)!}{l_1! \dots l_N!} \left( \frac{1}{N} \right)^{iT}$$

Therefore the expected total number of experimental steps is

$$\begin{aligned} T & \left( A \left( 1 + \frac{r-1}{2k} \right) E \left( \sum_{i=0}^{\mathcal{I}-1} B(iT) \right) + \frac{1}{r} \left( \frac{N}{k} E \left( \sum_{i=0}^{\mathcal{I}-1} B(iT) \right) - \frac{1}{2} E \left( \sum_{i=0}^{\mathcal{I}-1} B(iT) \left( \frac{NB(iT)}{k} - 1 \right) \right) \right) \right) \\ & = T \left( \left( A + \frac{(r-1)A}{2k} + \frac{N}{rk} + \frac{1}{2r} \right) \sum_{i=0}^{\infty} B(iT) P\{\mathcal{I} > i\} - \frac{N}{2rk} \sum_{i=0}^{\infty} B^2(iT) P\{\mathcal{I} > i\} \right) \end{aligned}$$

## Divide and conquer

Here we confirm our intuition of the performance of the third strategy. From Lemma 5.11 we see that the expected number of experimental steps to assign  $T$  probes to all the  $l = N$  chromosomes is a linear function of  $T$ . Therefore if we want to assign all  $\beta T$  probes to all  $N$  chromosomes, separating the probes to several groups does not give any improvement on the expected number of experimental steps. But, if our goal (as in the third strategy) is to assign  $m - 1$  probes to each of the  $N$  chromosomes, we can assign  $T$  probes first and then assign another  $T$  probes *only to the chromosomes that have not been assigned  $m - 1$  probes*, continuing this process until we screen all the  $\beta T$  probes. The expected total number of overlaps using this strategy is

$$T/r(k + (r-1)/2) \sum_{i=0}^{\beta-1} B(iT)$$

which is less than the total number of overlaps

$$\beta T/r(k + (r-1)/2)$$

if we assign the  $\beta T$  probes at the same time from Lemmas 5.11 and 5.12. The expected total number of experiments to detect these overlaps using the separating strategy is

$$\frac{T}{r} \sum_{i=0}^{\beta-1} \left( \frac{N}{k} B(iT) - \frac{1}{2} B(iT) \left( \frac{NB(iT)}{k} - 1 \right) \right)$$

from Lemma 5.12. Because  $B(iT) \leq 1$  and  $Nx/k - x/2(Nx/k - 1)$  is increasing in  $x \leq 1$ , we have

$$NB(iT)/k - B(iT)/2(NB(iT)/k - 1) \leq N/k - 1/2(N/k - 1) = (N/k + 1)/2$$

Thus the total number of experiments to detect the overlaps would be less than

$$\beta T(N/k + 1)/(2r)$$

which is the total number of experiments to detect the overlaps if we assign the  $\beta T$  probes in a single application of the sieve procedure (Lemma 5.11 again). From the above analysis we see the separating strategy can decrease both the number of overlaps and the number of experiments.

## 6 Conclusion

We have defined the Chromosome Characterization Problem and presented three different strategies for its solution. The three strategies have different flavors. The first strategy is the most straight forward and requires the most work. This is the most intuitive one and gives one way to solve this problem. The second strategy modifies the first strategy by pooling “success probes” and “sample probes” at the same time. By suitably choosing the “success probe” pooling size and “sample probe” pooling size, we can drastically reduce the number of experimental steps. The design of both strategies is quite easy, but they may be difficult to use in practice due to the larger number of probes required in some experiments. The third strategy requires the least work and has two main advantages. First, it uses a constant, small number of probes in each experiment. Second, many of the experiments can be run at the same time. Its disadvantage is that the arrangement and coloring of probes in the experiments are tricky and must be carefully arranged.

## Acknowledgments

We thank Vladimir Grebinski and Gregory Kucherov for their comments on a preliminary version of the paper. Their comments made us aware that we used more colors than necessary. This paper is supported by grants from the National Science Foundation (CCR-9623532 to GB) and the National Institutes of Health (GM 362320 to MSW). F. Sun was supported in part by the University Research Committee of Emory University.

## References

1. Baum, L. E., and Billingsley, P. 1969. Asymptotic distributions for the coupon collector's problem. *Ann. of Math. Statist.* 36, 1835-1839.
2. Feller, W. 1968. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, New York.
3. Ferretti, F., Nadeau, J.H., and Sankoff, D. 1996. Original synteny. *Lecture Notes in Computer Science. Combinatorial Pattern Matching, 7th Annual Symposium*, 1075, 159-167.
4. Johnson, C.V., Singer, R.H., and Lawrence, J.B. 1991. Fluorescent detection of nuclear RNA and DNA: implications for genome organization *Meth. Cell Biol.* 35, 73-98.
5. Klaassen, A. J. 1994. Dixie Cups: sampling with replacement from a finite population. *J. Appl. Prob.* 31, 940-948.
6. Le Beau, M. 1996. One FISH, two FISH, red FISH, blue FISH *Nature Genetics* 12, 341-344.
7. Lichter, P., Tang, C. J. C, Call, K., Hermanson, G., Evans, G. A., Housman, D., and Ward, D. C. 1990. High-resolution mapping of human chromosome 11 by *in situ* hybridization with cosmid clones. *Science* 247, 64-69.
8. Speicher, M.R., Ballard, S.G., and Ward, D.C. 1996, Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12, 368-375.

(a).  $C = 2$

$l$	6	7	8	9	10
# steps	906	852	811	880	856

(b).  $C = 3$

$l$	6	7	8	9	10
# steps	806	752	711	680	756

Table 1: The expected number of experimental steps using the first strategy to assign  $m = 5$  markers to  $N = 25$  chromosomes using (a).  $C = 2$  and (b).  $C = 3$  different colors for different values of sample probe pooling size  $l$ .

(a).  $C = 2$

$S, L$	6	7	8	9	10
1	782	721	676	744	716

(b).  $C = 3$

$S, L$	6	7	8	9	10
1	677	616	572	537	612
2	468	437	414	395	485

Table 2: Upper bounds for the expected number of experimental steps using the second strategy with success probe pooling size  $S$ , sample probe pooling size  $L$  and number of available colors (a).  $C = 2$  and (b).  $C = 3$ .

(a). Protocol 1:  $k = r = 3$

$T$	50	100	150	200	250	300
# experiments	381	431	466	534	609	700

(b). Protocol 2:  $k = 6, r = 3$

$T$	50	100	150	200	250	300
# experiments	308	348	378	443	494	561

Table 3: The simulated average numbers of experimental steps to assign  $m = 5$  probes to each of the  $N = 25$  chromosomes using the third strategy with  $T$  probes a time and (a). Protocol 1 with  $k = r = 3$  and (b). Protocol 2 with  $k = 6, r = 3$ .

## Figure Legends

- Figure 1. Histograms for the number of experimental steps using the first strategy after 5000 replications to assign  $m = 5$  probes to each of the  $N = 25$  chromosomes using  $C = 3$  colors for sample probe size (a).  $l = 8$ , (b).  $l = 9$ , and (c).  $l = 10$ .
- Figure 2. Histograms for the number of experimental steps using the second strategy after 5000 replications to assign  $m = 5$  probes to each of the  $N = 25$  chromosomes using  $C = 3$  colors for sample probe size  $L = 9$  and success probe size (a).  $S = 1$  and (b).  $S = 2$ .
- Figure 3 Histograms for the number of experimental steps using the third strategy after 5000 replications to assign  $m = 5$  probes to each of the  $N = 25$  chromosomes using  $C = 3$  colors using (a). Protocol 1 with  $k = r = 3$  and (b). Protocol 2 with  $k = 6$ ,  $r = 3$ .