Probability in the Engineering and Informational Sciences, **15**, 2001, 57–68. Printed in the U.S.A.

THE EXPECTED NUMBERS OF UNRESOLVED POSITIVE CLONES FOR VARIOUS RANDOM POOL DESIGNS

F. K. Hwang and Y. C. Liu
Department of Applied Mathematics
National Chiao Tung University
Hsinchu 300, Taiwan, Republic of China
E-mail: {fhwang; u8722518}@math.nctu.edu.tw

A pool design is random if it varies according to a probability distribution. There are four types of random design proposed in the literature: random incidence design, random *k*-set design, random distinct *k*-set design, and random *k*-size design. Recently Hwang gave an approximation to estimate the number of unresolved positives for random distinct *k*-set design. In this article, we give exact formulas for all four types of random designs for estimating the number of unresolved positives. We also do some numerical comparisons of the four designs.

1. INTRODUCTION

A *pool design* is used in identifying clones which contain a specific DNA fragment (referred to as a *probe*). A pool design can be represented by a $t \times n$ 0–1 matrix where each column is labeled by a clone and each row by a pool. A 1-entry in cell (i, j) implies that clone j is contained in pool i. For a given probe F, a clone is called *positive* if it contains F, and *negative* if not. The outcome of probing a pool is also binary: A *negative pool* means that the pool contains no positive clone; a *positive pool* means it does, but not knowing how many or which ones.

Ideally, by knowing the outcomes of the t pools in a pool design, we can identify all positive clones from the negatives. However, this may be achieved only with a very large t. To achieve economy, there is a trade-off between the size of t and the

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

possible unidentification. Thus, clones are classified into four categories: *resolved positives*, *unresolved positives*, *resolved negatives*, and *unresolved negatives*; "resolved" means identified. We will let P, $\bar{P}$, N, and $\bar{N}$ denote their respective numbers. Barillot et al. [2] first proposed to minimize $E(\bar{N})$.

A pool design is *deterministic* if the design matrix is fixed, and *random* if it varies according to a probability distribution. Although the best deterministic designs may minimize t, they are hard to find. On the other hand, random designs are usually surprisingly efficient. They also have the advantage of being applicable to all t and n, whereas good deterministic designs, relying on the availability of combinatorial designs, exist only for certain pairs of (t,n). There are four types of random designs proposed in the literature [1–4]:

1. Random incidence design. Each cell in the design matrix has a predetermined probability p of containing a 1-entry.
2. Random k-set design. Each column is a random k-subset of the set $\{1,\dots,t\}$ (the subset yields the set of rows containing 1-entries for that column).
3. Random distinct k-set design. Same as design 1 except the columns are distinct k-subsets.
4. Random k-size design. Each row is a random k-subset of the set $\{1,\dots,n\}$.

$E(\bar{N})$ has been computed for each of the above models. However, $E(\bar{P})$ is a much harder object to compute. An approximation of $E(\bar{P})$ was given in [4] for the random distinct k-set design, and the method was suggested to be extendable to the other models. In this article, we give an exact formula to compute $E(\bar{P})$ for all four models. We also do some computations to show how $E(\bar{N})$ and $E(\bar{P})$ depend on the choices of k or p and to compare the four designs.

2. EXACT FORMULAS FOR $E(\bar{P})$

For a given pool design, the set of negative pools is referred to as the *negative set.* Let d denote the number of positive clones. We first consider the random distinct k-set design.

Let $K_d(i)$ denote the probability that a given set of i pools is the negative set, and let $V_{d,i}(j)$ denote the probability of $\bar{N}=j$ given that the negative set is of size i. We quote two lemmas from [4]. Note that throughout this article, we follow the customary definition $x^0=1$, including $x=0$.

Lemma 1:

$$K_d(i)=\sum_{h=i}^{t}(-1)^{h-i}\binom{t-i}{h-i}\frac{\binom{\binom{t-h}{k}}{d}}{\binom{\binom{t}{k}}{d}}\qquad \textit{for } d>0.$$

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

LEMMA 2:

$$V_{d,i}(j) = \frac{\dbinom{\binom{t-i}{k}-d}{j}\dbinom{\binom{t}{k}-\binom{t-i}{k}}{n-d-j}}{\dbinom{\binom{t}{k}-d}{n-d}}.$$

Let $U_{d-1,i}(y)$ denote the probability that a positive clone (a k-subset) C has y indices not covered by the union of the other $d-1$ positive clones which yield a negative set of size i. We will refer to the y indices as *uncovered indices.*

LEMMA 3:

$$U_{d-1,i}(y) = \begin{cases} \dfrac{\binom{t-i}{k}-(d-1)}{\binom{t}{k}-(d-1)} & \textit{for } y=0 \\[2ex] \dfrac{\binom{i}{y}\binom{t-i}{k-y}}{\binom{t}{k}-(d-1)} & \textit{for } y>0. \end{cases}$$

PROOF: $y = 0$ implies that the k indices of C can be chosen randomly from the $t-i$ indices not in the negative set, except that C cannot be identical to any of the other $d-1$ positive clones. $y > 0$ implies that y indices of C must be chosen from the i indices of the negative set, and $k-y$ indices out of it. By noting that there are

$$\binom{t}{k}-(d-1)$$

ways of choosing a k-subset not identical to the other $d-1$ positive clones, Lemma 3 follows. ■

Therefore, if $d-1$ positive clones generate a negative set of size i, then adding another positive clone will generate a negative set of size $i-y$ with probability $U_{d-1,i}(y)$.

Finally, let $f(i,y,j)$ denote the probability that a positive clone C has at least one index not covered by the union of the other $d-1$ positive clones and the j unresolved negatives given that the negative set generated by the $d-1$ positive clones is of size i, and the number of uncovered indices of C is y.

Lemma 4:

$$f(i,y,j) = \begin{cases} \displaystyle\sum_{m=1}^{y} (-1)^{m-1} \binom{y}{m} \frac{\dbinom{\binom{t-i+y-m}{k} - (d-1)}{j}}{\dbinom{\binom{t-i+y}{k} - d}{j}} & \text{if } \dbinom{t-i+y}{k} \geq d+j \\ 0 & \text{otherwise.} \end{cases}$$

Proof: Consider a given set of m indices from the y uncovered indices of C. We compute the probability that these m indices are also not covered by the union of the j unresolved negatives. Note that the negative set generated by the d positive clones is of size $i-y$. An unresolved negative cannot contain any index of the negative set. It must also avoid the m indices out of the rest of the $t-i+y$ indices. There are

$$\binom{t-i+y-m}{k}$$

such choices, but $d-1$ of them are already taken by the $d-1$ positive clones. Therefore, there are

$$\binom{\binom{t-i+y-m}{k} - (d-1)}{j} \tag{1}$$

ways of choosing the j unresolved negatives. Adding this over all $m = 1,2,\ldots,y$, the total number of choices is

$$\binom{\binom{t-i+y}{k} - d}{j}. \tag{2}$$

So, the ratio of (1) and (2) is the probability we are computing. By noting that there are

$$\binom{y}{m}$$

choices of an m-subset and using the inclusion–exclusion principle, Lemma 4 follows. ■

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

Summarizing, we have Theorem 5.

THEOREM 5:

$$E(P) = d\sum_i \sum_y \sum_j \binom{t}{i} K_{d-1}(i) U_{d-1,i}(y) V_{d,i-y}(j) f(i,y,j).$$

Note that $E(\bar{P}) = d - E(P)$.

Let K^*, U^*, V^*, and f^* denote the terms corresponding to K, U, V, and f, respectively, except for the random k-set design (i.e., the columns do not have to be distinct). We gave K^*, U^*, V^*, and f^* without proofs since they are analogous to the distinct k-set case.

THEOREM 6: *For random k-set design,*

$$E^*(P) = d\sum_i \sum_y \sum_j \binom{t}{i} K^*_{d-1}(i) U^*_{d-1,i}(y) V^*_{d,i-y}(j) f^*(i,y,j),$$

where

$$K^*_{d-1}(i) = \sum_{h=i}^{t} (-1)^{h-i} \binom{t-i}{h-i} \left[\frac{\binom{t-h}{k}}{\binom{t}{k}} \right]^{d-1},$$

$$U^*_{d-1,i}(y) = \frac{\binom{i}{y}\binom{t-i}{k-y}}{\binom{t}{k}},$$

$$V^*_{d,i-y}(j) = \binom{n-d}{j} \left[1 - \frac{\binom{t-i+y}{k}}{\binom{t}{k}} \right]^{n-d-j} \left[\frac{\binom{t-i+y}{k}}{\binom{t}{k}} \right]^{j},$$

$$f^*(i,y,j) = \sum_{m=1}^{y} (-1)^{m-1} \binom{y}{m} \left[\frac{\binom{t-i+y-m}{k}}{\binom{t-i+y}{k}} \right]^{j} \quad \textit{for } k \geq t-i+y.$$

Let K', U', V', and f' denote the terms corresponding to K, U, V, and f, respectively, except for the random incidence design. Again, we give K', U', V', and f' without proofs.

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

THEOREM 7: *For the random incidence design*,

$$E'(P) = d\sum_i \sum_y \sum_j \binom{t}{i} K'_{d-1}(i)U'_{d-1,i}(y)V'_{d,i-y}(j)f'(y,j),$$

where

$$K'_{d-1}(i) = (1-p)^{(d-1)i}[1-(1-p)^{d-1}]^{t-i},$$

$$U'_{d-1,i}(y) = \binom{i}{y} p^y(1-p)^{i-y,}$$

$$V'_{d,i-y}(j) = \binom{n-d}{j}(1-p)^{(i-y)j}[1-(1-p)^{i-y}]^{n-d-j},$$

$$f'(y,j) = \sum_{m=1}^{y} (-1)^{m-1}\binom{y}{m}(1-p)^{jm}.$$

Let K'', U'', V'', and f'' denote the terms corresponding to K, U, V, and f, respectively, except for the random k-size design.

THEOREM 8: *For the random k-size design*,

$$E''(P) = d\sum_i \sum_y \sum_j \binom{t}{i} K''_{d-1}(i)U''_{d-1,i}(y)V''_{d,i-y}(j)f''(y,j),$$

where

$$K''_{d-1}(i) = \left(\frac{\binom{n-d+1}{k}}{\binom{n}{k}}\right)^{i}\left(1-\frac{\binom{n-d+1}{k}}{\binom{n}{k}}\right)^{t-i},$$

$$U''_{d-1,i}(y) = \begin{cases} \binom{i}{y}\left(\frac{\binom{n-d}{k-1}}{\binom{n-d+1}{k}}\right)^{y}\left(1-\frac{\binom{n-d}{k-1}}{\binom{n-d+1}{k}}\right)^{i-y} & \text{if } n-d+1 \geq k \\ \begin{cases} 1 & \text{if } y=0 \\ 0 & \text{otherwise} \end{cases} & \text{otherwise,} \end{cases}$$

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

$$V''_{d,i-y}(j) = \begin{cases} \binom{n-d}{j} \sum_{l=j}^{n-d} (-1)^{l-j} \binom{n-d-j}{l-j} \left[\frac{\binom{n-d-l}{k}}{\binom{n-d}{k}} \right]^{i-y} & \textit{if } n-d \geq k \\ \begin{cases} 1 & \textit{if } j = n-d \\ 0 & \textit{otherwise} \end{cases} & \textit{otherwise}, \end{cases}$$

$$f''(y,j) = \begin{cases} \sum_{m=1}^{y} (-1)^{m-1} \binom{y}{m} \left[\frac{\binom{n-d-j}{k-1}}{\binom{n-d}{k-1}} \right]^{m} & \textit{if } n-d \geq k-1 \\ 0 & \textit{otherwise}. \end{cases}$$

PROOF: $K''_{d-1}(i)$ was given in [4]. Let Y denote the set of uncovered indices, then a row r is in Y if and only if one of its kth 1-entries is at column C, and the other $k-1$ 1-entries are in the columns representing the negative clones. Hence, the probability of a row in Y is

$$\frac{\binom{n-d}{k-1}}{\binom{n-d+1}{k}}.$$

Because the rows are independent, the probability of a given set of y rows from the negative set equals to Y is given by $U''_{d-1,i}(y)$. Furthermore, the probability that j given negative clones (which do not appear in the $i-y$ pools of the negative set from the d positive clones) can be computed using the inclusion–exclusion formula to be

$$\sum_{l=j}^{n-d} (-1)^{l-j} \binom{n-d-j}{l-j} \left[\frac{\binom{n-d-l}{k}}{\binom{n-d}{k}} \right]^{i-y},$$

where

$$\frac{\binom{n-d-l}{k}}{\binom{n-d}{k}}$$

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

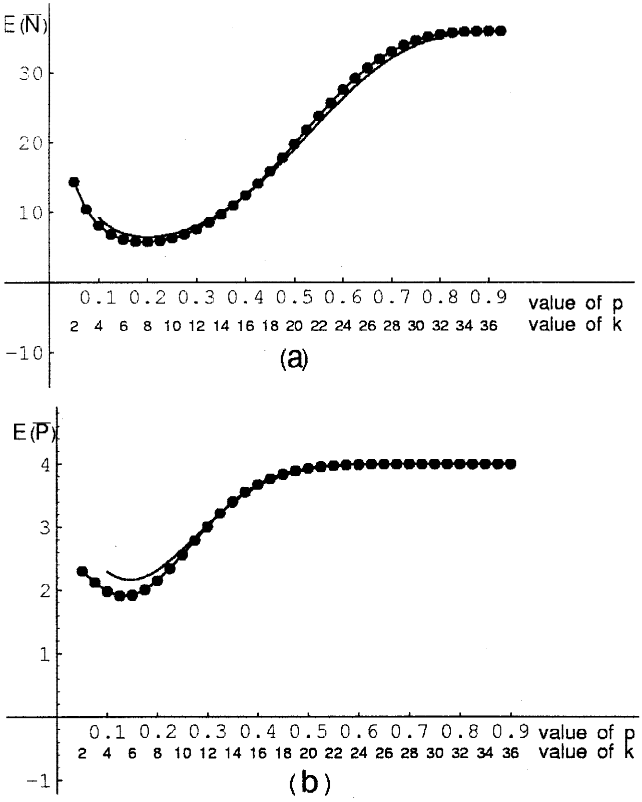

FIGURE 1. Random incidence and random k-size (in solid circles) design.

is the probability that a given pool in the $i - y$ pools does not contain C. $V''_{d,i-y}(j)$ is obtained by multiplying the above probability by the number of ways of choosing j clones from the $n - d$ negative clones. Finally, out of Y, at least one row not covered by the union of the j unresolved negatives is given by $f''(y, j)$. ■

$E(\bar{N})$ was given correctly in [4]. However, the formula for $V_{d,i}(j)$ used there to compute $P(\bar{N} = j)$ was slightly different from the one used in Theorem 8; hence, $P(\bar{N} = j)$ computed there was actually an approximation. Here, we give an exact formula.

THEOREM 9: *For the random k-size design,*

$$P(\bar{N} = j) = \sum_i \binom{t}{i} K_d(i) V_{d,i}(j).$$

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

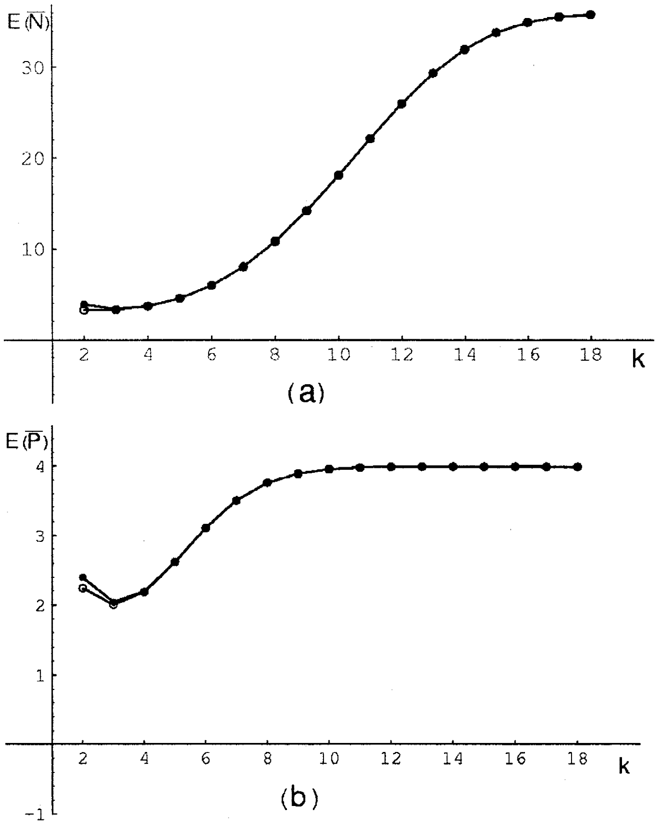

FIGURE 2. Random k-set and random distinct k-set (in open circles) design.

3. NUMERICAL RESULTS

We use the formulas given in [1,4] and in the last section to compute $E(\overline{N})$ and $E(\overline{P})$ for some values of n, t, d, k, and p for all four designs mentioned in Section 1. We draw the random k-set design and the random distinct k-set design together for easier eye comparison. We also draw the random incidence design and the random k-size design together, although their semblance is not a priori evident. For $n = 40$, $t = 20$, and $d = 4$, see Figure 1. For $n = 100$, $t = 30$, and $d = 5$, see Figure 2.

It was pointed out in [1] that $p = 1/(d + 1)$ minimizes $E(\overline{N})$ for the random incidence design. This is verified by Figures 1a and 3a. The p-value minimizing $E(\overline{P})$ for the first case is about 0.146, and for the second case, it is about 0.123, closer to $p = 1/(d + 3)$ than $1/(d + 1)$. For the random k-size design, it was pointed

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

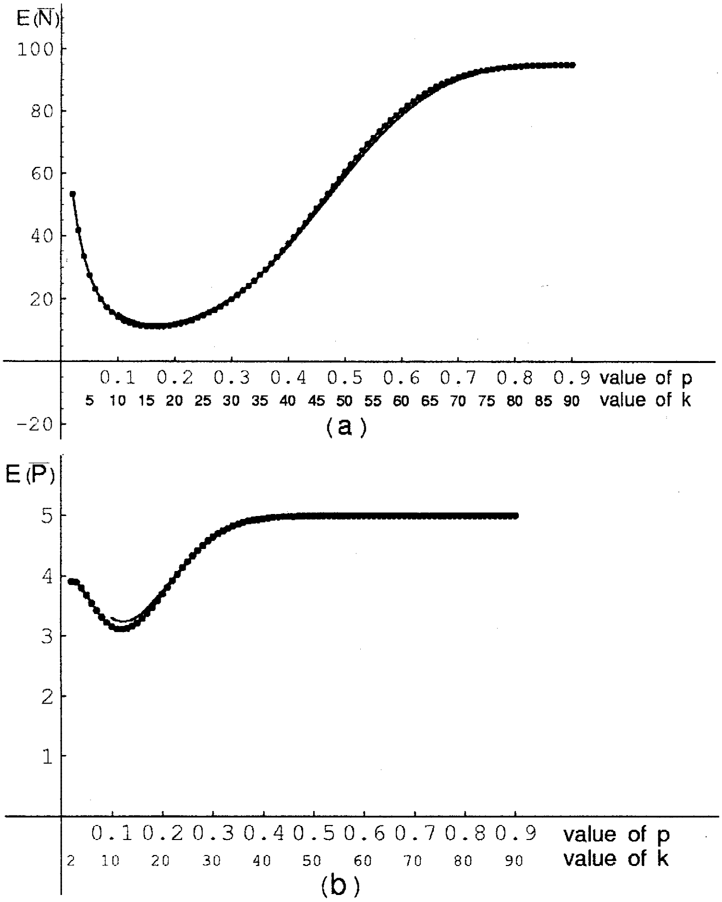

FIGURE 3. Random incidence and random k-size (in solid circles) design.

out in [4] that if we set $k = n/(d+1)$, then k equals the expected size of a random (incidence) pool design with optimal choice of p for $E(\bar{N})$. Therefore, we paired (p, k) with $k = np$ in the horizontal axis. Indeed, the optimal choices of k correspond to that of p in both cases and for both $E(\bar{N})$ and $E(\bar{P})$. The random k-size design is better than the random incidence design for both $E(\bar{N})$ and $E(\bar{P})$ around the optimal choices.

The random distinct k-set design is slightly better than the random k-set design, but the difference is observable only for small k (where the minimum $E(\bar{P})$ or $E(\bar{N})$

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press

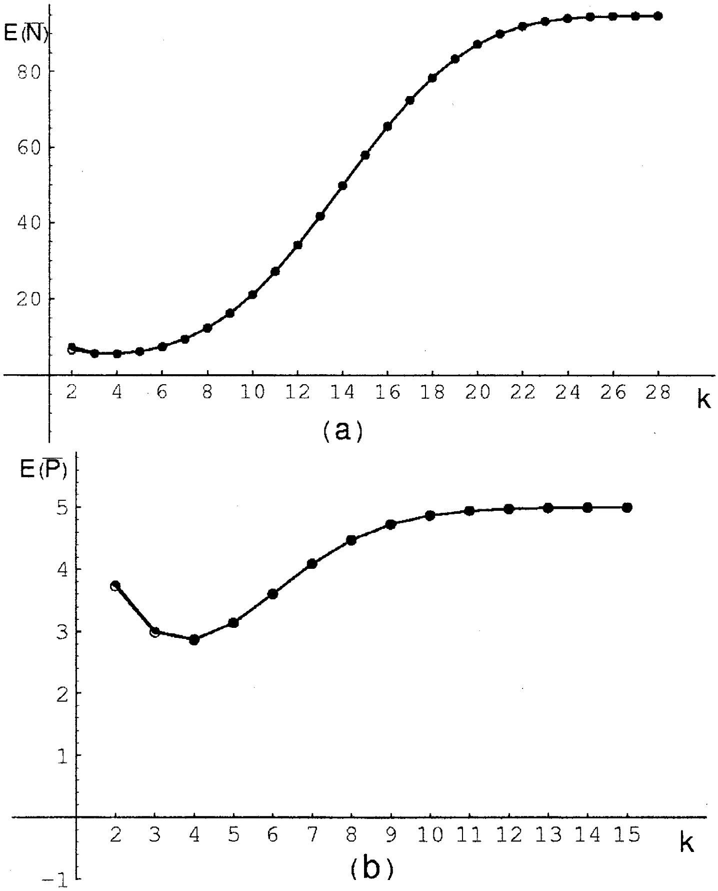

FIGURE 4. Random *k*-set and random distinct *k*-set (in open circles) design.

occurs). The optimal k-values are surprisingly small, ranging from 2 to 4 for both cases and for both $E(\bar{N})$ and $E(\bar{P})$ (see Fig. 4).

Across the board, we note that the curves shown in the figures are of the S type. In particular, these curves have the nice properties of having a unique minimum and a graceful degradation [i.e., a small slip in the k-value causing a small slip in $E(\bar{N})$ or $E(\bar{P})$]. Furthermore, the $E(\bar{N})$ curve and the $E(\bar{P})$ curve are in general agreement with respect to optimal choices of design parameters.

Comparisons of the different designs with their optimal choices of p or k are given in Table 1. It seems that the random incidence design is uniformly worst and the random k-size design is also not good for $E(\bar{N})$.

TABLE 1. Comparison of Designs with Optimal Choices of p or k

	$E(\bar{N})$		$E(\bar{P})$	
Value of n	40	100	40	100
Random incidence design	6.51	11.87	2.16	3.24
(value of p)	(0.2)	(0.167)	(0.146)	(0.123)
Random k-size design	5.77	11.1	1.91	3.11
(value of k)	(8)	(16)	(5)	(12)
Random k-set design	3.42	5.55	2.04	2.87
(value of k)	(3)	(4)	(3)	(4)
Random distinct k-set design	3.26	5.53	2.0	2.87
(value of k)	(2)	(4)	(3)	(4)

Acknowledgment

This work was supported by NCS grant 88-2115-M009-016.

References

1. Balding, D.J., Bruno, W.J., Knill, E., & Torney, D.C. (1996). A comparative survey of non-adaptive pooling designs. In T.P. Speed & M.S. Waterman (eds.), *Genetic mapping and DNA sequencing*, IBM Volumes in Mathematics and Its Applications. New York: Spring-Verlag, pp. 133–154.
2. Barillot, E., Laroix, B., & Cohen, D. (1991). Theoretical analysis of library screening using a n-dimensional pooling strategy. *Nucleic Acids Research* 19: 6241–6247.
3. Bruno, W.J., Knill, E., Balding, D.J., Bruce, D.C., Doggett, N.A., Sawhill, W.W., Stalling, R.L., Whittaker, C.C., & Torney, D.C. (1995). Efficient pooling designs for library screening. *Genomics* 26: 21–30.
4. Hwang, F.K. (to appear). Random k-set pool designs with distinct columns. *Probability in the Engineering and Information Science.*

https://doi.org/10.1017/S0269964801151041 Published online by Cambridge University Press