

OPTIMIZING NONADAPTIVE GROUP TESTS FOR OBJECTS WITH HETEROGENEOUS PRIORS*

W. J. BRUNO[†], F. SUN[‡], AND D. C. TORNEY[†]

Abstract. We investigate nonadaptive group testing designs for heterogeneous mixtures of objects, independently *positive* with individual prior probabilities. In our model of the prior probabilities, the objects occur in one of several disjoint subsets and the number of positives in each subset is known. Furthermore, the positives are “uniformly distributed” within the subsets. The expected number of unresolved negative objects is minimized, and a unique global minimum is found for a family of stochastic, *random incidence* designs: all v group tests are constructed independently. The optimum incidence probabilities for the objects are well approximated by an asymptotic power series in v^{-1} . We find the three leading coefficients of this series. The dependence of the optimum incidence probability upon the prior probability is, to leading order, logarithmic. Objects with larger prior probability of being positive have smaller optimum incidence probability. Furthermore, this logarithmic dependence can be nonnegligible for screening collections of cloned DNA sequences.

Key words. experimental design, group testing, constrained optimization, asymptotic power series, clone library screening

AMS subject classifications. 26B99, 34A10, 41A60, 60C05, 62K05, 92A90, 94B99

PII. S0036139996305062

1. Introduction. A binary group test ordinarily indicates whether any of the objects in the group satisfy a criterion. Objects meeting the criterion are called *positives*, and our group test yields a positive result if the group contains any positives. If a group test is feasible, then a coordinated family of group tests efficiently identifies the positives. To illustrate the potential of group testing, consider S. Ulam’s example [18]: *I am thinking of one integer, which could be any integer from 1 to n . I will provide a correct, binary answer to any questions.*¹ *What is the minimum number of questions guaranteeing the identification of the integer?* It is not difficult to see that the number of questions must be at least as large as the logarithm base two of n and that a strategy involving bisection could achieve this bound. A fair proportion of present-day research has an analogous objective: *identify the needles in the haystack*. For example, consider the hunter of genes, or the troubleshooter—are all systems “go”?—or the prover of theorems. Because of this broad applicability, it is appropriate to employ general terms—of collections of objects and of positives. Whenever there are, at the outset, many possibilities, of which only a few are likely to be realized, group testing could logarithmically reduce the amount of work required for ascertainment. A reasonably reliable group test is, of course, the key for capitalizing upon this potential. Although many effective group tests exist, it behooves us to improve group testing techniques. In this manuscript, we take the existence of an effective group test for granted and

*Received by the editors June 10, 1996; accepted for publication (in revised form) February 25, 1997; published electronically May 11, 1998. This work was performed under the auspices of the U. S. Department of Energy and was supported by the Los Alamos Center for Human Genome Studies, the Los Alamos Center for Nonlinear Science, and a Los Alamos Laboratory Directed Research and Development grant.

<http://www.siam.org/journals/siap/58-4/30506.html>

[†]T-10, Theoretical Biology and Biophysics, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545 (billb@lanl.gov, dct@lanl.gov).

[‡]Department of Genetics and Molecular Medicine, 1462 Clifton Road, 429F Emory University School of Medicine, Atlanta, GA 30322 (fsun@gmm.gen.emory.edu).

¹Ulam also commended the situation in which incorrect answers might be provided.

we focus upon optimizing the group testing *design*: the incidence of objects in all group tests. Group testing is only one of the important applications of combinatorial designs [6, 14].

We began to study group testing designs because of the interest in nonadaptively screening large collections of clones of DNA sequences [8, 12, 16]. Nonadaptive designs are those in which the groups to be tested are constructed in advance and are all used in one pass, regardless of the test outcomes, facilitating the automation of screening of large collections. The experimental objective is, ordinarily [8], to identify all clones containing a given DNA sequence: the positives. The Polymerase Chain Reaction (PCR) is the basis of a premier group test, reduplicating a given sequence of interest—even from a single initial molecule [1, 17]. The presence of any positives out of thousands of clones is detectable using the PCR [13], for which we have employed Bayes statistics to reconcile experimental errors [15]. Frequently, there exist clone length data and estimates of the extent of overlap between pairs of clones [3], both of which should be incorporated into the prior joint probability distribution for the positives.

Whenever group testing is performed, similar information about the joint distribution is likely to be available. It is inconceivable that it would be, in general, a good practice either to ignore this information or to separate the objects into batches prior to group testing. When optimizing group testing designs, it is conventional to use a variant of the model in which there are j positive objects in a set of n objects, and each subset of size j is equally likely to be the set of positives. The objective of this manuscript is to take the first steps toward optimizing nonadaptive group testing designs—including all the objects—given an arbitrary joint prior probability distribution for the positives.

Our model of the prior probability distribution allows an arbitrary partition of the set of objects into disjoint, nonempty subsets, or *parts*, each with a fixed number of uniformly distributed positives. This model recommends itself because it allows quantitation of the improved performance of designs which make use of the prior probabilities. Our model might be credited with accommodating the fundamental novelty: individual priors. To be sure, parts with fixed numbers of positives will yield results which are only qualitatively related to the situation in which objects are independently positive, allowing a continuum of priors. Also, our model cannot address “higher-order” aspects of prior distributions: for example, priors on the outcomes for pairs of objects. As we explain in section 5, more general models should prove substantially more difficult to analyze. Rest assured that our model, despite its shortcomings, yields the first insights into nonadaptive group testing designs particularly appropriate for the screening of collections of objects, independently positive with individual priors. The applicability is illustrated in section 5, using a particular collection of clones.

An effective group testing design enables correct inference of the set of positives, so the status of each of the positives and the negatives should be unambiguous or *resolved*. The criterion for design performance we will employ is the expected number of *unresolved negative* objects: negative objects occurring only in positive group tests. When using group testing, it is usually assumed that the negative objects vastly outnumber the positive objects, which motivates our focus upon the negative objects. Because further experiments would be required to distinguish unresolved negative objects from positive objects, they manifest shortcomings of a group testing design. Furthermore, in our random incidence designs, the probability that a negative object

is unresolved is also an upper bound on the probability that a positive object is unresolved, when these objects are from the same part of the partition. We are not the first to recommend this performance criterion for screening clone libraries [5]. Other performance criteria are described elsewhere [8, 2, 15].

We employ a very simple stochastic design, a *random incidence design*, described more fully in section 2. The focus of this manuscript is to determine, for our model, the optimum incidence probabilities for this design. Most treatments of nonadaptive group testing designs are purely combinatoric [4]. It is, however, unrealistic to expect to be able to construct an appropriate combinatorial design, such as a Steiner system, for every new experiment. The additional desiderata arising from our model—objects occurring in different numbers of group tests—make matters worse. Alternatively, stochastic designs, such as the one we use, have the advantages that they can be easily generated and characterized [2, 8, 15]. Although the performance of stochastic designs is not anticipated to be the best possible, optimizing stochastic designs recommends itself for preliminary characterization of appropriate designs. The simplest and least constrained of all stochastic designs is the random incidence design. Although our decision to employ this design might be worrisome because the performance of a design should, to some extent, reflect the quality of its “organization,” in section 5 we demonstrate that random incidence designs asymptotically achieve respectable results, requiring a number of group tests less than twice the minimum number possible for our model. In practice, computational methods could yield better stochastic designs, which could be viewed as approximations to combinatorial designs [8].

2. Particulars of the model, design, and criterion. Our model comprises L subsets of the set of n_T objects which are the disjoint, nonempty subsets, or *parts*, of a partition of the objects. The number of parts L is therefore between one and the number of objects. Let l index these parts; part l contains n_l objects, of which j_l are positive, and every subset of size j_l of the n_l objects has the same prior probability of being the set of positives. It is expedient to restrict the parameters of the model as follows. Note that if, for some l , j_l equals zero, the corresponding optimum incidence probability equals unity, and if, for some l , $n_l - j_l$ equals zero, the corresponding optimum incidence probability equals zero. Therefore, without loss of generality we assume that $0 < j_l$ and $0 < n_l - j_l$ for all l .

We consider a very simple design—a *random incidence design*: each object in part l occurs in each group test with incidence probability r_l , independently of the other incidences in group tests. A random incidence design contains a specified number, v , of group tests. Our goal is to find a collection of r_l 's globally optimizing the performance of random incidence designs.

Our performance criterion is $\mathbb{E}(\tilde{N})$, the expected number of unresolved negative objects, which we proceed to determine for our model and the random incidence design. The averaging is over two types of randomness: the incidences in group tests and the identities of the positive objects. A group test is negative if and only if no positive objects are included, which occurs with probability R :

$$R = \prod_{l=1}^L (1 - r_l)^{j_l}.$$

Because the random incidence group test results are independent of one another, the number of negative group tests is binomially distributed with parameters R and v . Given i negative group tests, a negative group l object is unresolved if and only if

it does not belong to any of the i negative group tests, which occurs with probability $(1 - r_l)^i$. Therefore, the probability that a negative group l object is unresolved is

$$\sum_{i=0}^v \binom{v}{i} R^i (1 - R)^{v-i} (1 - r_l)^i = (1 - Rr_l)^v.$$

Therefore,

$$(1) \quad \mathbb{E}(\tilde{N}) = \sum_{l=1}^L (n_l - j_l)(1 - Rr_l)^v.$$

The further characterization of the distribution of \tilde{N} through its variance might be of interest.

$$\begin{aligned} \text{Var}(\tilde{N}) = & \sum_{k,l=1}^L (n_k - j_k)(n_l - j_l) \{ (1 - R(r_k + r_l) + Rr_k r_l)^v - (1 - Rr_k)^v (1 - Rr_l)^v \} \\ & + \sum_{k=1}^L (n_k - j_k) \{ (1 - Rr_k)^v - (1 - 2Rr_k + Rr_k^2)^v \}. \end{aligned}$$

In this manuscript we determine optimum incidence probabilities which globally minimize $\mathbb{E}(\tilde{N})$ for fixed parameters: v, L , and all j_l and n_l . The domain for the incidence probabilities is the closure of the regular, unit measure-polytope, or hypercube, $\mathcal{C}_L \stackrel{\text{def}}{=} [0, 1]^L$. The optimum incidence probabilities are denoted by $r_1^*, r_2^*, \dots, r_L^*$, or collectively by \mathbf{r}^* . In section 3 we demonstrate that there is a unique global minimum, and in section 4 we characterize an asymptotic power series for the optimum incidence probabilities, in reciprocal powers of v . No special treatment is afforded the simple case $v = 1$ because $\mathbb{E}(\tilde{N})$ would ordinarily be unacceptably large.

Before proceeding, it might be reasonable to hazard that it should follow from our performance criterion that objects in groups with more positives should have smaller incidence probabilities, so that all objects would cast comparable “shadows” over the negatives. In fact, among the surprises in store, the optimum incidence probabilities are, approximately, a logarithmic function of the reciprocal of the prior probability of being positive. Full details are given in section 4.

3. The global minimum, \mathbf{r}^* , of $\mathbb{E}(\tilde{N})$. In this section we find a unique global minimum of $\mathbb{E}(\tilde{N})$, considered as a function of the L incidence probabilities, r_1, r_2, \dots, r_L . In the spirit of Lagrange multipliers, an auxiliary variable, Λ , is introduced to facilitate the solution of the equations for a stationary point. When Λ is sufficiently large, these equations have one solution, and, as Λ is decreased, there remains only one solution—a continuous function of Λ . The definition of a stationary point is extended so that boundary points of \mathcal{C}_L might also be stationary points. Λ is decreased until all conditions for a stationary point are met. This is the global minimum point. We begin with two introductory lemmas which bear upon the minimization.

LEMMA 1. *If every r_l increases by an infinitesimal increment equal to ϵr_l , $0 < \epsilon$, every $(1 - Rr_l)^v$ of $\mathbb{E}(\tilde{N})$ changes, respectively, by $\epsilon v R r_l (1 - Rr_l)^{v-1} (\sum_{l=1}^L j_l r_l / (1 - r_l) - 1)$.*

This result is obtained by taking the total differential of $(1 - Rr_l)^v$. Therefore, the only points of \mathcal{C}_L which might be the global minimum point of $\mathbb{E}(\tilde{N})$, given by (1),

are the origin, $(0, 0, \dots, 0)$, and the points of a manifold, \mathcal{H} , defined by the following equation:

$$(2) \quad \sum_{l=1}^L \frac{j_l r_l}{1 - r_l} = 1.$$

The global minimum, evidently, occurs in \mathcal{H} . For future reference, we also denote the manifold in which the r_l are nonnegative, $1 \leq l \leq L$, and where $\sum_{l=1}^L \frac{j_l r_l}{1 - r_l} \leq 1$ by \mathcal{D} .

Recall that $(r_1^*, r_2^*, \dots, r_L^*)$ denotes the global minimum point of $\mathbb{E}(\tilde{N})$. Our second result is that the r_l^* are all strictly between zero and unity when v is sufficiently large.

LEMMA 2. *The r_l^* 's are bounded away from 0 and 1 as v tends to infinity: $0 < \liminf_{v \rightarrow \infty} r_l^* \leq \limsup_{v \rightarrow \infty} r_l^* < 1, \quad 1 \leq l \leq L$.*

Proof. We prove this lemma by reductio ad absurdum. The $r_l^*(v)$ are bounded between zero and unity. Therefore, the Bolzano–Weierstrass theorem affirms the existence of a subsequence v_i of increasing values of v such that $\lim_{i \rightarrow \infty} r_l^*(v_i)$ exists for $1 \leq l \leq L$. Suppose there exists an l , such that $\lim_{v_i \rightarrow \infty} r_l^*(v_i) = 0$; without loss of generality, take l to be 1. Let $\hat{r} \stackrel{\text{def}}{=} 1/(1 + j_T)$ and $\hat{R} \stackrel{\text{def}}{=} (1 - \hat{r})^{j_T}$, in which j_T denotes $\sum_{l=1}^L j_l$. Then,

$$\lim_{i \rightarrow \infty} \frac{1 - R r_1^*(v_i)}{1 - \hat{R} \hat{r}} = \frac{1}{1 - \hat{R} \hat{r}} > 1.$$

Therefore

$$\lim_{i \rightarrow \infty} \left(\frac{1 - R r_1^*(v_i)}{1 - \hat{R} \hat{r}} \right)^{v_i} = \infty.$$

Thus, from (1),

$$\lim_{i \rightarrow \infty} \frac{E(\tilde{N}(r_1^*(v_i), r_2^*(v_i), \dots, r_L^*(v_i)))}{E(\tilde{N}(\hat{r}, \hat{r}, \dots, \hat{r}))} = \lim_{i \rightarrow \infty} \sum_{l=1}^L \frac{n_l - j_l}{n_T - j_T} \left(\frac{1 - R r_l^*(v_i)}{1 - \hat{R} \hat{r}} \right)^{v_i} = \infty,$$

in which n_T denotes $\sum_{l=1}^L n_l$. This is self-contradictory as $\mathbf{r}^*(v_i)$ is supposed to be the global minimum point of $\mathbb{E}(\tilde{N})$. Similarly, $\limsup_{v \rightarrow \infty} r_l^* < 1$. \square

All admissible minima of $\mathbb{E}(\tilde{N})$ must occur either in the interior or upon the boundary of \mathcal{C}_L , the measure-polytope. We will modify the definition of a stationary point so that all minima must occur at a *candidate-minimum stationary point*. In the interior of \mathcal{C}_L , the necessary and sufficient conditions for a stationary point of $\mathbb{E}(\tilde{N})$ are the vanishing of the first partial derivative with respect to r_1, r_2, \dots , and r_L . Employing (1), these L conditions are written as follows.

$$(3) \quad A_k(1 - r_k)(1 - R r_k)^{v-1} = \Lambda,$$

$k = 1, 2, \dots, L$, in which A_k denotes $(n_k - j_k)/j_k$, and with

$$(4) \quad \Lambda \stackrel{\text{def}}{=} \sum_{l=1}^L A_l j_l r_l (1 - R r_l)^{v-1}.$$

By adopting (3) and using (4) to determine Λ , we exclude other stationary solutions which are evidently not minima, with R equal to zero—or an r_k equal to unity. Note

that equations (3) are identical when A_k is the same for all k . In this case we can substitute \hat{r} for all r_k and obtain

$$(5) \quad \hat{r} \stackrel{\text{def}}{=} \frac{1}{1 + j_T},$$

where, as above, j_T denotes $\sum_{l=1}^L j_l$. As the value which (1) takes at \hat{r} is clearly smaller than at $r = 0$ or at $r = 1$, this stationary point is the global minimum of $\mathbb{E}(\tilde{N})$. This is, of course, the same minimum found for $L = 1$ with j_T positives [2]. When the A_k 's are unequal, a closed-form solution for a candidate-minimum stationary point is, in general, unobtainable. Nevertheless, the salient aspect of stationary solutions—that there is a unique solution yielding a global minimum point for $\mathbb{E}(\tilde{N})$ —is always true, as we proceed to demonstrate.

For sufficiently large Λ , those equations (3) with the smallest values of A_k cannot have a solution in $[0, 1]$. It is useful to set the corresponding r_k 's equal to zero, and to define a solution of the remaining equations (3) and (2) to be a *candidate-minimum stationary point* because this facilitates the uniform treatment for all points in the domain: boundary and interior. Note that, for any point $\mathbf{r} \in \mathcal{H}$, Λ is given by (4), as demonstrated at the beginning of the proof of Theorem 1.

DEFINITION. A candidate-minimum stationary point is a point $\mathbf{r} \in \mathcal{H} \cap \mathcal{C}_L$ whose components $r_k, 1 \leq k \leq L$, satisfy the following condition: if $\Lambda < A_k$ then r_k solves (3), and otherwise, r_k equals zero.

It is easy to see from (3) and (4) that $\mathbb{E}(\tilde{N})$ would increase if any of the variables, equal to zero, of a candidate-minimum stationary point were to increase from zero. Thus, this definition is consistent with the requirements upon a minimum which is attained upon the boundary of \mathcal{C}_L . We now state our main result.

THEOREM 1. There is precisely one candidate-minimum stationary point of $\mathbb{E}(\tilde{N})$, at which $\mathbb{E}(\tilde{N})$ achieves its global minimum in \mathcal{C}_L .

Proof. First, recall that, from Lemma 1, all minima of $\mathbb{E}(\tilde{N})$ must occur in \mathcal{H} . To prove this theorem, we establish the uniqueness of the candidate-minimum stationary point in Lemmas 3 and 4. To this end, we will use Λ as an auxiliary variable, deferring its functional dependence upon the r_l 's given in (4)—the arrow in (4) indicating that equality will not obtain until the candidate-minimum stationary point is found. If (4) were used to eliminate Λ from (3), multiplying both sides by $j_k r_k / (1 - r_k)$ and summing over k would yield (2). Thus, we pursue the alternative strategy of solving the $L + 1$ equations (2) and (3) in $L + 1$ unknowns—the r_l and Λ . The same ends are, of course, achieved, but the use of an auxiliary variable will be advantageous. In fact, this use of Λ is analogous to the use of Lagrange multipliers for constrained optimization—our optimum is constrained to lie in \mathcal{H} . The uniqueness of the candidate-minimum stationary point is a corollary of (2) and the following lemma.

LEMMA 3. The component variables of a collection of r_k s, which solve (3) for all k such that $\Lambda < A_k$ and which equal zero otherwise, are all continuous functions of Λ within \mathcal{D} . Furthermore, all nonzero r_k 's are strictly decreasing functions of Λ .

Proof. From (3), the total differential of Λ can be written (in L different ways) as

$$(6) \quad d\Lambda = \sum_{l=1}^L f_{kl} dr_l, \quad k = 1, 2, \dots, L,$$

in which

$$f_{kl} \stackrel{\text{def}}{=} A_k \partial \{ (1 - r_k)(1 - Rr_k)^{v-1} \} / \partial r_l.$$

Performing the partial differentiation yields

$$(7) \quad f_{kl} = b_k \rho_l - \delta_{kl} a_k,$$

in which δ_{kl} is the Kronecker delta—unity if k equals l and zero otherwise—and

$$a_k \stackrel{\text{def}}{=} A_k(1 - r_k)(1 - Rr_k)^{v-2} \left\{ (v - 1)R + \frac{1 - Rr_k}{1 - r_k} \right\},$$

$$b_k \stackrel{\text{def}}{=} (v - 1)RA_k(1 - r_k)(1 - Rr_k)^{v-2}r_k,$$

$$\rho_l \stackrel{\text{def}}{=} j_l/(1 - r_l).$$

The L linear equations (6) can be solved for all dr_l in terms of a ratio of determinants of two matrices of order L . The determinant in the denominators is of the matrix f with elements f_{kl} given by (7), and, for each dr_l , the determinant in the numerator is of the matrix $g^{(l)}$: the matrix f with the elements of the l th column replaced by $d\Lambda$. It will suffice to study $|f|$ and $|g^{(1)}|$: establishing that dr_1 must have a sign opposite to $d\Lambda$ implies that the same holds for all dr_l because a unit cyclic permutation of the first $l - 1$ columns followed by a unit cyclic permutation of the first $l - 1$ rows transforms $g^{(l)}$ to the form of $g^{(1)}$, with the a_k 's on the diagonal, and the determinant of $g^{(l)}$ is invariant to this transformation.

The “outer product” form contributing to the matrix elements of f results in extensive cancellation in $|f|$ and $|g^{(1)}|$. Explicitly,

$$|f| = (-1)^L \left\{ 1 - \sum_{l=1}^L \left(\frac{j_l r_l}{1 - r_l} \right) / \left(1 + \frac{1 - Rr_l}{(v - 1)R(1 - r_l)} \right) \right\} \prod_{k=1}^L a_k.$$

When v equals unity, the value of the summation is taken to be zero. As Rr_l is strictly less than unity, it is evident that the sign of $|f|$ is $(-1)^L$ in \mathcal{D} . It is perhaps easiest to find the sign of $|g^{(1)}|$ by performing a Laplace expansion of $|g^{(1)}|$, with the elements $g_{11}^{(1)} = d\Lambda$ in its first column, and evaluating the cofactors. This yields that the sign of $|g^{(1)}|$ is $(-1)^{L-1}$ in this domain, as can be seen from the following straightforward considerations. Once it has been shown that $|f|$ and $|g^{(1)}|$ have opposite signs in \mathcal{D} , Lemma 3 will be proven because the latter is proportional to $d\Lambda$.

The cofactor of $g_{11}^{(1)}$ has the sign $(-1)^{L-1}$ because it has the form of an $|f|$ of order $L - 1$. The cofactors of the remaining $g_{l1}^{(1)}$, $2 \leq l \leq L$, are put in the same form as one another, with the $(L - 2)a_k$'s on the diagonal, with a unit cyclic permutation of the first $l - 2$ columns. This permutation ensures that all cofactors will have the same sign as the cofactor of $g_{21}^{(1)}$. As $g_{21}^{(1)}$ reduces to only one term, the product of -1 with $(L - 2)(-a_k)s, b_1$, and $j_2/(1 - r_2)$, its sign is $(-1)^{L-1}$, and, recalling that this applies to all cofactors, this is the sign of $|g^{(1)}|$. Thus, Lemma 3 is proven. \square

The formulas for the derivatives follow.

$$(8) \quad \frac{dr_l}{d\Lambda} = \frac{|g^{(l)}|}{|f|} = - \frac{1 + \sum_{k=1}^L (b_l - b_k)\rho_k/a_k}{a_l(1 - \sum_{k=1}^L b_k\rho_k/a_k)}.$$

We use both Proposition 1 and Lemma 3 to establish Lemma 4.

LEMMA 4. *There is a unique candidate-minimum stationary point.*

Proof. Our proof relies on the existence of a unique solution to (8), passing through each point in \mathcal{D} . A. L. Cauchy was the first of many to consider these issues [9, 7, 11].

PROPOSITION 1. *The continuity of $\frac{dr_k}{d\Lambda}$ and of $\frac{\partial dr_k}{\partial r_l d\Lambda}$, $1 \leq k, l \leq L$, is sufficient to establish the uniqueness of the solution to (8) taking given initial values, or, equivalently, the uniqueness of the solution to the L integral equations:*

$$r_k(\Lambda_1) = r_k(\Lambda_0) + \int_{\Lambda_0}^{\Lambda_1} d\Lambda' \frac{dr_k}{d\Lambda} \Big|_{\Lambda'}.$$

This proposition is readily established. Therefore, as the conditions are evidently met, integration could yield all the $r_l(\Lambda)$ throughout \mathcal{D} , given an initial “point.”

The following conceptual algorithm describes a technique for obtaining \mathbf{r}^* through integration, employing Lemma 3 and Proposition 1. It is convenient to index the r_l 's according to the decreasing value of the A_l 's. Thus, A_1 is taken to be the largest of the A_l 's, A_2 is taken to be the next largest, etc. Integration commences with all the r_l 's equal to zero and Λ equal to A_1 . As Λ decreases, relevant r_l increase and \mathcal{H} is approached monotonically. Integration ceases when \mathcal{H} is encountered because the candidate-minimum stationary point is found there. \mathcal{H} will, evidently, be encountered before Λ decreases to zero. In detail, as Λ decreases from A_1 to A_2 , all variables other than r_1 are maintained at zero, and r_1 increases according to the integral of $\frac{dr_1}{d\Lambda}$, integrating (8) as if L were unity. As Λ decreases from A_2 to A_3 , all variables other than r_1 and r_2 are maintained at zero, and r_1 and r_2 both increase according to the integrals of $\frac{dr_1}{d\Lambda}$ and $\frac{dr_2}{d\Lambda}$, integrating (8) as if L were two. Similarly, an additional $\frac{dr_l}{d\Lambda}$ would be integrated as an additional interval between A_l 's is traversed by Λ .

The possibility that there is more than one candidate-minimum stationary point is eliminated as a corollary of Proposition 1. If there were two candidate-minimum stationary points in \mathcal{H} , integration back into \mathcal{D} from both points, using (8)—reversing the conceptual algorithm given above and increasing both values of Λ to A_1 —would yield the point $(0, 0, \dots, 0)$ in both cases. This would be self-contradictory because Proposition 1 establishes the uniqueness of the solution of (8), passing through a point of \mathcal{D} . Therefore, Lemma 4 is proven. \square

Recalling that the r_l are indexed according to decreasing value of the A_l 's, the following is a corollary of Lemmas 3 and 4 and their proofs.

COROLLARY 1. *For the candidate-minimum stationary point \mathbf{r} , $r_k \leq r_l$ if $l < k$. In particular, if any r_l equals zero, all r_k with $l < k$ equal zero. Because all j_l are at least unity, it can be seen from (2) that all r_l are less than unity.*

This corollary is easily established from (8), because the only dependence of the right-hand side involves the ratio b_l/a_l . From their definitions it follows that $r_l < r_{l'}$ implies $b_l/a_l < b_{l'}/a_{l'}$, at least for $r_{l'} \leq 1/2$, which holds in \mathcal{D} because $j_{l'}$ is at least unity.

According to Weierstrass's theorem, $\mathbb{E}(\tilde{N})$ must achieve its global minimum somewhere in \mathcal{C}_L . This minimum cannot occur on any part of the boundary with any r_l equal to unity because $\mathbb{E}(\tilde{N})$ takes its maximum value there. If a minimum were to occur on any other part of the boundary, with some r_l 's equal to zero, it would be necessary that $\mathbb{E}(\tilde{N})$ be increasing with these r_l 's, evaluated at zero, and also that $\mathbb{E}(\tilde{N})$ be stationary with respect to the remaining r_l 's. However, this is directly seen to be equivalent to our definition of a candidate-minimum stationary point, and by Lemma 4, this point is unique in \mathcal{C}_L . Thus, by process of elimination, the unique

candidate-minimum stationary point is the point at which $\mathbb{E}(\tilde{N})$ attains its global minimum. This completes the proof of Theorem 1. \square

4. Asymptotic power series for \mathbf{r}^* . The optimum incidence probabilities are further characterized in this section, focusing on their evaluation for large v . We demonstrate that these probabilities are representable by an asymptotic series in reciprocal powers of v , and we derive the three leading coefficients of the series. The coefficients of this series are tabulated near the end of this section. Inspection shows, as expected, that the series affords a useful approximation even if v is $\mathcal{O}(10)$.

Lemma 2 establishes that $0 < r_l^* < 1$ for all l , in the limit that v tends to infinity. This result is essential in our proof of the following theorem. Recall that $(r_1^*, r_2^*, \dots, r_L^*)$ denotes the global minimum point of $\mathbb{E}(\tilde{N})$.

THEOREM 2.

$$\lim_{v \rightarrow \infty} r_l^* = \hat{r}, \quad 1 \leq l \leq L,$$

in which $\hat{r} = 1/(1 + j_T)$.

Proof. The point \mathbf{r}^* must satisfy (3). Therefore,

$$(9) \quad \frac{n_k - j_k}{j_k} (1 - Rr_k^*)^{v-1} (1 - r_k^*) = \frac{n_l - j_l}{j_l} (1 - Rr_l^*)^{v-1} (1 - r_l^*), \quad 1 \leq k < l \leq L.$$

Taking the $(v - 1)$ th root on both sides yields

$$(10) \quad \left(\frac{n_k - j_k}{j_k} \right)^{1/(v-1)} (1 - Rr_k^*) (1 - r_k^*)^{1/(v-1)} = \left(\frac{n_l - j_l}{j_l} \right)^{1/(v-1)} (1 - Rr_l^*) (1 - r_l^*)^{1/(v-1)},$$

$k < l.$

From Lemma 2 we see that, in the limit that v goes to infinity, any global minimum point must be in $(0, 1)^L$. Because $(1 - r_l^*)$ is bounded away from 0 and 1, we have

$$\lim_{v \rightarrow \infty} (1 - r_l^*)^{1/(v-1)} = 1$$

and

$$\lim_{v \rightarrow \infty} \left(\frac{n_l - j_l}{j_l} \right)^{1/(v-1)} = 1.$$

We now prove $\lim_{v \rightarrow \infty} r_l^*$ exists and equals \hat{r} for every l . From Lemma 2 and the assertions made in its proof, there exists a subsequence v_i of increasing values of v such that $\lim_{i \rightarrow \infty} r_l^*(v_i) = \hat{r}_l$ exists and $0 < \hat{r}_l < 1$ for $1 \leq l \leq L$. Substituting v_i for v and letting i tend to infinity in (10), we obtain

$$1 - R' \hat{r}_k = 1 - R' \hat{r}_l,$$

where $R' = \prod_{l=1}^L (1 - \hat{r}_l)^{j_l} \neq 0$. Therefore

$$\hat{r}_k = \hat{r}_l, \quad 1 \leq k < l \leq L.$$

Next we prove all $\hat{r}_l = \hat{r}$, as defined in the theorem and in (5). From (9) with $k = 1$, we have

$$(n_l - j_l)(1 - Rr_l^*)^{v-1} = \frac{(n_1 - j_1)j_l}{j_1} \frac{1 - r_1^*}{1 - r_l^*} (1 - Rr_1^*)^{v-1}, \quad 1 \leq l \leq L.$$

Substituting the above equations into (3) with $k = 1$ and dividing both sides by $\frac{n_1 - j_1}{j_1} (1 - Rr_1^*)^{v-1}$ we obtain (2). Because the \hat{r}_l are all equal, we obtain the equation for \hat{r} appearing in the theorem. \square

Theorem 2 gives the limit for \mathbf{r}^* as v tends to infinity. Corrections to this result for finite v will prove to be nonnegligible for most group testing applications. The next theorem gives the first-order correction of r_l^* .

THEOREM 3.

$$r_l^{(1)} \stackrel{\text{def}}{=} \lim_{v \rightarrow \infty} v(r_l^* - \hat{r}) = \frac{1 - \hat{R}\hat{r}}{\hat{R}} \log \left(\frac{n_l - j_l}{j_l} / \prod_i \left(\frac{n_i - j_i}{j_i} \right)^{j_i/j_T} \right), \quad 1 \leq l \leq L,$$

in which $\hat{r} = 1/(1 + j_T)$ and $\hat{R} = (1 - \hat{r})^{j_T}$.

Proof. We first prove that $\lim_{v \rightarrow \infty} v(r_l^* - \hat{r})$ exists for every l . Transposing (9) with $k = 1$ gives

$$\frac{(1 - Rr_l^*)^{v-1}}{(1 - Rr_1^*)^{v-1}} = \frac{(n_1 - j_1)j_l (1 - r_1^*)}{(n_l - j_l)j_1 (1 - r_l^*)}.$$

From Theorem 2, the limit of the left-hand side of the equation exists and it is finite and nonzero. Therefore,

$$\frac{1 - Rr_l^*}{1 - Rr_1^*} = 1 + \frac{Rc_l}{v} + o(1/v),$$

where c_l does not depend on v . The factor of R on the right simplifies subsequent calculations. From this equation we can easily show

$$r_l^* = r_1^* + c_l'/v + o(1/v),$$

where $c_l' = (\hat{r}\hat{R} - 1)c_l$. Let $r_1^* = \hat{r} + \epsilon(v)$, where $\epsilon(v)$ tends to zero as v tends to infinity from Theorem 2. Then

$$\begin{aligned} \frac{r_l^*}{1 - r_l^*} &= \frac{\hat{r} + \epsilon(v) + c_l'/v + o(1/v)}{1 - (\hat{r} + \epsilon(v) + c_l'/v + o(1/v))} \\ &= \frac{\hat{r}}{1 - \hat{r}} \left(\frac{1 + (\epsilon(v) + c_l'/v)/\hat{r} + o(1/v)}{1 - (\epsilon(v) + c_l'/v)/(1 - \hat{r}) + o(1/v)} \right) \\ &= \frac{\hat{r}}{1 - \hat{r}} \left(1 + \frac{\epsilon(v) + c_l'/v}{\hat{r}} + o(1/v) \right) \left(1 + \frac{\epsilon(v) + c_l'/v}{1 - \hat{r}} + o(\max(\epsilon(v), 1/v)) \right) \\ &= \frac{\hat{r}}{1 - \hat{r}} \left(1 + \left(\frac{1}{\hat{r}} + \frac{1}{1 - \hat{r}} \right) \left(\epsilon(v) + \frac{c_l'}{v} \right) + o(\max(\epsilon(v), 1/v)) \right). \end{aligned}$$

Substituting these expansions into (2) yields

$$\sum_{l=1}^L j_l \left(\epsilon(v) + \frac{c_l'}{v} \right) + o(\max(\epsilon(v), 1/v)) = 0.$$

Multiplying by v on both sides and letting v tend to infinity yields

$$\lim_{v \rightarrow \infty} \left(\sum_{l=1}^L j_l (v\epsilon(v) + c_l') + o(\max(v\epsilon(v), 1)) \right) = 0.$$

If $v\epsilon(v)$ were unbounded, then there would be a subsequence v_i of increasing values of v such that $v_i\epsilon(v_i)$ tends to infinity. Because the last term tends to infinity at a lower rate than $v_i\epsilon(v_i)$, the left-hand side will tend to infinity. This is self-contradictory. Therefore, $v\epsilon(v)$ is bounded and the last term tends to zero as v tends to infinity. Furthermore, $\lim_{v \rightarrow \infty} v\epsilon(v)$ exists and is finite. From the above argument, we see that $\lim_{v \rightarrow \infty} v(r_l^* - \hat{r})$ exists and is finite for every l . We now obtain this limit.

Let $r_l^* = \hat{r} + r_l^{(1)}/v + o(1/v)$. Using (2), we find

$$(11) \quad \sum_{l=1}^L j_l r_l^{(1)} = 0,$$

and, therefore, $R = \hat{R} + o(1/v)$. Substituting these formulas into (9) and letting v tend to infinity yields

$$\frac{n_k - j_k}{j_k} \exp(-\hat{R}r_k^{(1)}/(1 - \hat{R}\hat{r})) = \frac{n_l - j_l}{j_l} \exp(-\hat{R}r_l^{(1)}/(1 - \hat{R}\hat{r})), \quad 1 \leq k < l \leq L.$$

That is, there is a constant $C^{(1)}$, which does not depend on l , satisfying

$$\frac{n_l - j_l}{j_l} \exp(-\hat{R}r_l^{(1)}/(1 - \hat{R}\hat{r})) = C^{(1)}, \quad 1 \leq l \leq L.$$

Thus

$$r_l^{(1)} = \frac{1 - \hat{R}\hat{r}}{\hat{R}} \log \left(\frac{n_l - j_l}{C^{(1)}j_l} \right), \quad 1 \leq l \leq L.$$

We next calculate $C^{(1)}$. From (11)

$$\sum_{l=1}^L j_l \log \left(\frac{n_l - j_l}{C^{(1)}j_l} \right) = 0,$$

which can be solved for $C^{(1)}$:

$$C^{(1)} = \prod_{l=1}^L \left(\frac{n_l - j_l}{j_l} \right)^{j_l/j_T}.$$

These results complete the proof of the theorem. \square

As a corollary of Theorem 3, $r_l^* = \hat{r} + r_l^{(1)}/v + o(1/v)$. The next theorem gives the second-order correction of r_l^* .

THEOREM 4.

$$r_l^{(2)} \stackrel{\text{def}}{=} \lim_{v \rightarrow \infty} v^2(r_l^* - \hat{r} - r_l^{(1)}/v) = C^{(2)} - \frac{\hat{R}}{2(1 - \hat{R}\hat{r})} r_l^{(1)2} - \frac{1 - \hat{R}}{\hat{R}(1 - \hat{r})} r_l^{(1)},$$

in which \hat{r} , \hat{R} , and $r_l^{(1)}$ are defined in Theorem 3 and

$$C^{(2)} = \frac{1}{j_T} \left(\frac{\hat{R}}{2(1 - \hat{R}\hat{r})} - \frac{1}{1 - \hat{r}} \right) \sum_{l=1}^L j_l r_l^{(1)2}.$$

Proof. In the following, we use the notation $A_l = (n_l - j_l)/j_l$, $\hat{B} = (1 - \hat{R}\hat{r})/\hat{R}$, and $C^{(1)} = \prod_{l=1}^L \binom{n_l - j_l}{j_l}^{j_l/j^r}$. As in the proof of Theorem 3, we first establish the existence of the posited limits. From (9), we have

$$\begin{aligned} \left(\frac{1 - Rr_l^*}{1 - Rr_1^*}\right)^{v-1} &= \frac{A_l(1 - r_l^*)}{A_1(1 - r_1^*)} \\ &= \frac{A_l(1 - \hat{r} - r_l^{(1)}/v + o(1/v))}{A_1(1 - \hat{r} - r_1^{(1)}/v + o(1/v))} \\ &= \frac{A_l}{A_1} \left(1 - \frac{r_l^{(1)} - r_1^{(1)}}{v(1 - \hat{r})} + o(1/v)\right) \\ &= \frac{A_l}{A_1} \left(1 - \hat{B} \frac{\log(A_l/A_1)}{v(1 - \hat{r})} + o(1/v)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{1 - Rr_l^*}{1 - Rr_1^*} &= \left(\frac{A_l}{A_1}\right)^{1/(v-1)} \left(1 - \hat{B} \frac{\log(A_l/A_1)}{v(1 - \hat{r})} + o(1/v)\right)^{1/(v-1)} \\ &= \exp\left(\frac{1}{v} \log\left(\frac{A_l}{A_1}\right) - \frac{\log(A_l/A_1)}{v^2} \left(\frac{\hat{B}}{1 - \hat{r}} - 1\right) + o(1/v^2)\right) \\ &= 1 + \frac{1}{v} \log\left(\frac{A_l}{A_1}\right) - \frac{\log(A_l/A_1)}{v^2} \left(\frac{\hat{B}}{1 - \hat{r}} - 1\right) + \frac{1}{2v^2} \log^2\left(\frac{A_l}{A_1}\right) + o(1/v^2). \end{aligned}$$

Transposing and simplifying yields:

$$\begin{aligned} 1 - Rr_l^* &= (1 - Rr_1^*) \left(1 + \frac{1}{v} \log\left(\frac{A_l}{A_1}\right) - \frac{\log(A_l/A_1)}{v^2} \left(\frac{\hat{B}}{1 - \hat{r}} - 1\right) \right. \\ &\quad \left. + \frac{1}{2v^2} \log^2\left(\frac{A_l}{A_1}\right) + o(1/v^2)\right) \\ &= 1 - Rr_1^* + (1 - \hat{R}\hat{r}) \frac{1}{v} \log\left(\frac{A_l}{A_1}\right) + \frac{d_l}{v^2} + o(1/v^2), \end{aligned}$$

in which d_l does not depend on v . Therefore,

$$r_l^* = r_1^* - \frac{\hat{B}}{v} \log\left(\frac{A_l}{A_1}\right) - \frac{d_l}{v^2} + o(1/v^2).$$

Thus we obtain

$$r_l^* - \hat{r} - \frac{r_l^{(1)}}{v} = r_1^* - \hat{r} - \frac{r_1^{(1)}}{v} - \frac{d_l}{v^2} + o(1/v^2).$$

As in the proof of Theorem 3, we can assume

$$r_1^* = \hat{r} + \frac{r_1^{(1)}}{v} + \epsilon(v)$$

and prove that $\lim_{v \rightarrow \infty} v^2 \epsilon(v)$ exists. Therefore, the limits in Theorem 4 exist, and, from the above argument, we can write

$$r_l^* = \hat{r} + \frac{r_l^{(1)}}{v} + \frac{r_l^{(2)}}{v^2} + o(1/v^2).$$

Using Taylor series expansions, it is easily found that

$$\frac{r_l^*}{1 - r_l^*} = \frac{\hat{r}}{1 - \hat{r}} \left(1 + \frac{r_l^{(1)}}{\hat{r}(1 - \hat{r})v} + \frac{1}{v^2 \hat{r}(1 - \hat{r})} \left(\frac{r_l^{(1)2}}{1 - \hat{r}} + r_l^{(2)} \right) + o(1/v^2) \right).$$

From (2), it follows that

$$(12) \quad \frac{\sum_{l=1}^L j_l r_l^{(1)2}}{1 - \hat{r}} + \sum_{l=1}^L j_l r_l^{(2)} = 0.$$

Using these results, it is straightforward to obtain

$$\frac{A_l(1 - Rr_l^*)^{v-1}}{C^{(1)}(1 - \hat{R}\hat{r})^{v-1}} = \exp(-e_l/v + o(1/v)),$$

in which

$$e_l = \frac{(r_l^{(2)} - r_l^{(1)})\hat{R}}{1 - \hat{R}\hat{r}} + \frac{\hat{R}^2 r_l^{(1)2}}{2(1 - \hat{R}\hat{r})^2} + D,$$

and D does not depend on l . Dividing both sides of (9), with $k = 1$, by $(1 - \hat{R}\hat{r})^{v-1}(1 - \hat{r})$ yields, as v tends to infinity,

$$e_l + r_l^{(1)}/(1 - \hat{r}) = e_1 + r_1^{(1)}/(1 - \hat{r}).$$

Therefore, $e_l + r_l^{(1)}/(1 - \hat{r})$ does not depend on l , and there is another constant, $C^{(2)}$, independent of l :

$$C^{(2)} = r_l^{(2)} + \frac{\hat{R}}{2(1 - \hat{R}\hat{r})} r_l^{(1)2} + \frac{1 - \hat{R}}{\hat{R}(1 - \hat{r})} r_l^{(1)}.$$

Multiplying this last equation by j_l and summing over l , using (12), yields the values given in Theorem 4. \square

Using the methods described in the proofs of Theorems 2, 3, and 4, the following theorem is established.

THEOREM 5. *The r_l^* must have an asymptotic power series expansion of the form*

$$r_l^* = \hat{r} + \frac{r_l^{(1)}}{v} + \frac{r_l^{(2)}}{v^2} + \dots + \frac{r_l^{(k)}}{v^k} + o(1/v^k).$$

We omit the proof of this theorem, which is constructed along the lines of some of the foregoing proofs. We do not know whether the infinite series is convergent for any finite v . We summarize the results of this section in Table 1.

TABLE 1

$r_l^{(0)} = \hat{r} = (1 + j_T)^{-1}$	$r_l^{(1)} = \frac{1 - \hat{R}\hat{r}}{\hat{R}} \log\left(\frac{n_l - j_l}{C^{(1)}j_l}\right)$	$r_l^{(2)} = C^{(2)} - \frac{\hat{R}}{2(1 - \hat{R}\hat{r})} r_l^{(1)2} - \frac{1 - \hat{R}}{\hat{R}(1 - \hat{r})} r_l^{(1)}$
$\hat{R} = (1 - \hat{r})^{j_T}$	$C^{(1)} = \prod_{l=1}^L \left(\frac{n_l - j_l}{j_l}\right)^{j_l/j_T}$	$C^{(2)} = \frac{1}{j_T} \left(\frac{\hat{R}}{2(1 - \hat{R}\hat{r})} - \frac{1}{1 - \hat{r}}\right) \sum_{l=1}^L j_l r_l^{(1)2}$

5. Discussion. In this manuscript we optimized random incidence designs for nonadaptive group testing of mixtures of heterogeneous objects, in which the heterogeneity is manifest through stratified prior probabilities of being positive. We used a reasonable and simple criterion for optimization: the expected number of unresolved negative objects. Progress was aided by the analytic simplicity of random incidence designs and by our model, with fixed numbers of uniformly distributed, positive objects. We established the existence of a unique global minimum point of $\mathbb{E}(\tilde{N})$ in \mathcal{C}_L , the domain of admissibility.

The coefficients in the asymptotic power series expansion of the optimum incidence probabilities have a number of salient features. Theorem 2 establishes that the optimum incidence probabilities converge to $\hat{r} = (1 + j_T)^{-1}$ in the limit that v tends to infinity. Theorem 3 contains the first correction terms, proportional to $\frac{1}{v} \log\left(\frac{n_l - j_l}{C^{(1)}j_l}\right)$. Each is a function of the ratio $(n_l - j_l)/j_l$: essentially, the probability that a group l object is negative divided by the probability that it is positive. The sign of this correction term is also noteworthy. The constant $C^{(1)}$, equal to $\prod_i \left(\frac{n_i - j_i}{j_i}\right)^{j_i/j_T}$, is the geometric mean of $(n_l - j_l)/j_l, l = 1, 2, \dots, L$, with $(n_l - j_l)/j_l$ repeated j_l times. From Theorem 3, we see that when $(n_l - j_l)/j_l$ is less than $C^{(1)}$, $r_l^{(1)}$ is negative, and when $(n_l - j_l)/j_l$ is greater than $C^{(1)}$, $r_l^{(1)}$ is positive. In the extreme cases that there are either no positive objects, $j_l = 0$, or no negative objects, $n_l - j_l = 0$, $r_l^{(1)}$ would be undefined, corroborating the necessity of excluding these situations—where the optimum incidence probabilities are, respectively, unity and zero. Corollary 1 is consistent with the functional form of $r_l^{(1)}$. That is, if any groups have $r_l^* = 0$, these will have the smallest ratios $(n_l - j_l)/j_l$. The exponential dependence of $\mathbb{E}(\tilde{N})$ on v ensures that, for modestly large v , optimum incidence probabilities will not equal zero. However, if some of the approximate values of the r_l^* were negative, a remedy might be to repeat the calculation with the r_l^* 's of the k smallest A_l 's equal to zero, sequentially for $k = 1, 2, \dots$, until all r_l^* are positive.

Because the first corrections of the r^* 's are exponentiated in the calculation of $\mathbb{E}(\tilde{N})$, they cannot be neglected when obtaining the leading order estimate: $\mathbb{E}(\tilde{N}) \approx j_T C^{(1)} (1 - \hat{R}\hat{r})^v + o(1/v)$, employing various approximations such as $j_l \ll n_l$ and $1 \ll j_T$. The ratio of this last result to $\sum_{l=1}^L (n_l - j_l) (1 - \hat{R}\hat{r})^v$ is indicative of whether it might be advantageous to employ distinct incidence probabilities. From the approximate formula, the value of v at which $\mathbb{E}(\tilde{N})$ equals unity can be found: $v \approx e \times \sum_{l=1}^L j_l \log(n_l)$. This result might be compared with the information-theory bound, which asserts that the minimum number of group tests guaranteeing resolution of all objects' status exceeds $\frac{1}{\log(2)} \times \sum_{l=1}^L j_l \log(n_l)$. However, it is always remarkable for a design to achieve the information-theory bound, and it should, therefore, be reassuring that a reasonable random incidence design could be constructed using approximately 1.9 times as many group tests as the theoretical minimum. Furthermore, the approximate total number of group tests needed to achieve similar performance

if, instead, the objects of each part were included in separate, optimized, random-incidence designs, is $e \times j_T \log(n_T)$. All of these considerations should encourage the use of nonadaptive designs for screening heterogeneous mixtures of objects.

The second-order correction terms yield first-order corrections to the minimum $\mathbb{E}(\tilde{N})$. Higher-order coefficients could be obtained canonically from the Taylor series expansions, and this might be preferable to direct numerical optimization, whenever high accuracy is required. Antipodally, we are prepared to sacrifice accuracy to allow extrapolation of our results to situations which exceed our model, such as the experiments on clones which do not have fixed numbers of positives. Only qualitative insights might be gained because our model does not precisely capture the key attributes of a collection of clones. Therefore, the performance of candidate group testing designs should always be characterized, through computer simulation, in advance of their implementation.

Although it might be noted that our results could be applied to extremely heterogeneous collections of clones, such as mixtures of clones from different cloning hosts, we focus upon a key collection of 33,000 clones for the human genome, having clones between 10^5 to 10^6 basepairs in length [10]. Assume the coverage, c —the average number of times a “point” within the human genome occurs in this collection—equals 10. Suppose we had decided to use 250 group tests, a number which has been predicted to give good performance for a related stochastic design [8]. How might the probabilities for a random incidence design be chosen for this collection of clones? Furthermore, is there a substantial difference between the optimum incidence probabilities for the shortest and longest clones? A natural way to use our results to provide answers would be to substitute relevant expectations for the numbers of positives.

Thus, we replace j_T by c in the coefficients of the series approximations of the optimum random incidence probabilities.² The leading order coefficient, $1/(1+c)$, would place each clone into 23 group tests, on the average, with c equal to 10. Similarly, if p is the probability a clone is positive, then $(1-p)/p$ replaces $(n_i - j_i)/j_i$ in $r^{(1)}$. As the extent of a haploid human genome is 3×10^9 basepairs, the probability that a uniformly placed, short clone, of length 10^5 basepairs, covers any “point” within the genome is approximately equal to $(3 \times 10^4)^{-1}$, and the comparable probability for a long clone, of length 10^6 basepairs, is approximately equal to $(3 \times 10^3)^{-1}$. These probabilities are denoted $p_<$ and $p_>$, respectively. Making the indicated substitutions, the difference between the first-order corrections to the optimum incidence probabilities for short and long clones,

$$\frac{r_{<}^{(1)} - r_{>}^{(1)}}{v} \approx \frac{1 - \hat{R}\hat{r}}{v\hat{R}} \{ \log((1 - p_{<})/p_{<}) - \log((1 - p_{>})/p_{>}) \}.$$

Substituting $(1+c)^{-1}$ for \hat{r} and $(c/(1+c))^c$ for \hat{R} , and letting v equal 250 yields that $\frac{1-\hat{R}\hat{r}}{v\hat{R}}$ is approximately equal to 0.01. Substituting the probabilities given above, the term within the curly braces is approximately equal to $\log(10)$. The leading-order term, \hat{r} , equals $1/11$. The ratio of the difference between the two first-order corrections to \hat{r} is approximately equal to 0.25, indicating that these corrections to the incidence probabilities should not be neglected. If $C^{(1)}$ were known, this would complete the estimation of the $r^{(1)}$'s. In fact, the distribution of clone lengths is required to estimate

²In practice, it would probably be more advantageous to replace j_T with an approximate upper bound on the number of positives, such as $c + 2\sqrt{c}$. This would result in clones occurring in a smaller proportion of the group tests, consistent with previous results [8].

$C^{(1)} \approx \prod_{i=1}^{n_T} ((1 - p_i)/p_i)^{\frac{p_i}{c}}$, in which the product is over all n_T clones and p_i is the probability that clone i is positive, which is, as above, proportional to its measured length.

It is apparent that our model, with fixed numbers of positives, is not fully adequate for modeling the screening of collections of objects with individual priors. Substantial additional technical difficulties would necessarily arise in the analysis of the more realistic model. For our model and our criterion, optimum designs are “well behaved” in the limit that the number of group tests goes to infinity, facilitating asymptotic analysis. For example, the optimum proportion of group tests each object occurs in converges rapidly. On the other hand, if the number of positives were variable, the optimum proportion of group tests each object would occur in would slowly converge to a number most suited to the case with maximum possible number of positives. This peculiarity arises because an optimum design would be expected to resolve the status of most objects, provided the number of positives is not too large, and to perform abysmally for larger numbers of positives. Thus, if the number of group tests were increased, the characteristics of an optimum design would alter in the process of achieving better performance for larger numbers of positives. Therefore, in the limit that the number of group tests increases indefinitely, optimum designs would be unduly influenced by the case with the maximum possible number of positives. It follows that a novel limiting procedure is required for many models with a randomly distributed number of positives. For instance, let the number of objects increase with the number of group tests, v , as a^v , a being an arbitrary positive number, targeting the optimum designs upon the central part of the probability distribution. It is also possible to achieve this objective by adopting another performance criterion [8].

In conclusion, our model yields helpful insights for actual applications of group testing, but there remains plenty of room for the analysis of more comprehensive models. A practical, intermediate goal might involve modeling a probability distribution which is adequately characterized by expectations and covariances. Wouldn't it be nice to know, ultimately, which nonadaptive designs are optimum for an arbitrary joint probability distribution of the positives?

Acknowledgments. George I. Bell, of Los Alamos National Laboratory, fostered D. Torney's interest in optimization.

REFERENCES

- [1] N. ARNHEIM, H. LI, AND X. CUI (1990), *PCR analysis of DNA sequences in single cells: Single sperm gene mapping and genetic disease diagnosis*, Genomics, 8, pp. 415–419.
- [2] D. J. BALDING, W. J. BRUNO, E. KNILL, AND D. C. TORNEY (1996), *A comparative survey of non-adaptive pooling designs*, in Genetic Mapping and DNA Sequencing, T. P. Speed and M. S. Waterman, eds., IMA Vol. Math. Appl., Springer-Verlag, New York, pp. 133–154,
- [3] D. J. BALDING AND D. C. TORNEY (1991), *Statistical analysis of DNA fingerprint data for ordered clone physical mapping or human chromosomes*, Bull. Math. Biol., 53, pp. 853–879.
- [4] D. J. BALDING AND D. C. TORNEY (1996), *Optimal pooling designs with error detection*, J. Combin. Theory Ser. A, 74, pp. 131–140.
- [5] E. BARILLOT, B. LACROIX, AND D. COHEN, (1991), *Theoretical analysis of library screening using a n-dimensional pooling strategy*, Nucl. Acids Res., 19, pp. 6241–6247.
- [6] T. BETH, D. JUNGNIKEL, AND H. LENZ (1985), *Design Theory*, Bibliographisches Institut-Wissenschaftsverlag.
- [7] W. E. BOYCE AND R. C. DiPRIMA (1969), *Elementary Differential Equations and Boundary Value Problems*, 2nd Ed., John Wiley, New York.

- [8] W. J. BRUNO, E. KNILL, D. C. BRUCE, N. A. DOGGETT, W. W. SAWHILL, R. L. STALLINGS, C. C. WHITTAKER, AND D. C. TORNEY (1995), *Efficient pooling designs for library screening*, *Genomics*, 26, pp. 21–30.
- [9] A. L. CAUCHY (1855), *Sur la nature des Intégrales d'un système d'équations différentielles du premier ordre*, *C. R. Acad. Sci. Paris*, 40, pp. 376–381.
- [10] I. CHUMAKOV, P. RIGAUT, S. GUILLOU, P. OUGEN, A. BILLAUT, G. GUASCONI, P. GERVY, I. LEGALL, P. SOULARUE, L. GRINAS, L. BOGUELERET, C. BELLANÉ-CHANTELOT, B. LACROIX, E. BARILLOT, P. GESNOUIN, S. POOK, G. VAYSSEIX, G. FRELAT, A. SCHMITZ, J.-L. SAMBUCY, A. BOSCH, X. ESTIVILL, J. WEISSENBACH, A. VIGNAL, H. REITHMAN, D. COX, D. PATTERSON, K. GARDINER, M. HATTORI, Y. SATAKI, H. ICHIKAWA, M. OHKI, D. LE PASLIER, R. HEILIG, S. ANTONARAKIS, AND D. COHEN (1992), *Continuum of overlapping clones spanning the entire human chromosome 21q*, *Nature (London)*, 359, pp. 380–387.
- [11] E. A. CODDINGTON (1961), *An Introduction to the Theory of Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ.
- [12] N. A. DOGGETT, L. A. GOODWIN, J. G. TESMER, L. J. MEINCKE, D. C. BRUCE, L. M. CLARK, M. R. ALTHERR, A. A. FORD, H.-C. CHI, B. L. MARRONE, J. L. LONGMIRE, S. A. LANE, S. A. WHITMORE, M. G. LOWENSTEIN, R. D. SUTHERLAND, M. O. MUNDT, E. H. KNILL, W. J. BRUNO, C. A. MACKEN, D. C. TORNEY, J.-R. WU, J. GRIFFITH, G. R. SUTHERLAND, L. L. DEAVEN, D. F. CALLEN, AND R. K. MOYZIS (1995), *An integrated physical map of human chromosome 16*, *Nature*, 377, pp. 335S–365S.
- [13] E. D. GREEN AND M. V. OLSON (1990), *Systematic screening of yeast artificial chromosome libraries by the use of the polymerase chain reaction*, *Proc. Nat. Acad. Sci. USA*, 87, pp. 1213–1217.
- [14] R. L. GRAHAM, M. GRÖTSCHEL, AND L. LOVÁSZ, EDS. (1995), *Handbook of Combinatorics*, Elsevier and MIT Press.
- [15] E. KNILL, A. SCHLIEP, AND D. C. TORNEY (1996), *Interpretation of pooling experiments using the Markov chain Monte Carlo method*, *J. Comp. Biol.*, 3, pp. 395–406.
- [16] M. OLSON, L. HOOD, C. CANTOR, AND D. BOTSTEIN (1989), *A common language for physical mapping of the human genome*, *Science*, 245, pp. 1434–1435.
- [17] F. SUN (1995), *The polymerase chain reaction and branching processes*, *J. Comput. Biol.*, 2, pp. 63–86.
- [18] S. M. ULAM, *Adventures of a Mathematician*, University of California Press, Berkeley, 1991.