Probability in the Engineering and Informational Sciences, **29**, 2015, 131–145.
doi:10.1017/S0269964814000254

MODELING AND OPTIMIZATION OF GENETIC SCREENS VIA RNA INTERFERENCE AND FACS

Yair Goldberg and Yuval Nov
Department of Statistics, University of Haifa, Israel

We study mathematically a method for discovering which gene is related to a cell characteristic ("phenotype") of interest. The method is based on RNA interference – a molecular process for gene deactivation – and on coupling the phenotype with cell fluorescence. A small number of candidate genes are thus isolated, and then tested individually. We model probabilistically this process, prove a limit theorem for its outcome, and derive operational guidelines for maximizing the probability of successful gene discovery.

1. INTRODUCTION

A fundamental problem in science is discovering which genes are related to an organism's phenotype of interest (a phenotype is an observable trait or characteristic, such as eye color or susceptibility for a certain disease). Revealing unknown gene–phenotype relationships advances our understanding of biological systems, and often paves the way for developing novel therapeutics.

A widely used experimental approaches for studying the gene–phenotype relationship is based on RNA interference (RNAi) – a natural biochemical process, in which small RNA molecules deactivate (or "silence") genes inside cells. Biotechnological advances of the past decade allow scientists to exploit the RNAi mechanism and deactivate specific genes of their choosing, by introducing into cells appropriately designed RNAi molecules. RNAi technology has thus become a powerful tool that revolutionized biomedical and genetic research. For reviews, see Dykxhoorn and Lieberman [6], Mohr, Bakal, and Perrimon [12].

When attempting to discover which (supposedly single) gene – henceforth termed *the target gene* – is related to a specific cell phenotype of interest, a researcher can deactivate a candidate gene using the appropriate RNAi construct, and then observe whether the phenotype is altered; when it is, a relationship between the two is established. However, since the organisms under study typically have many thousands of genes, genome-wide RNAi experiments of this type are often too expensive and laborious to be carried out on a gene-by-gene basis.

An alternative is *pooled* RNAi screens, whereby a large number of RNAi constructs of various types (i.e., corresponding to various genes) are inserted randomly into a large population of cells. We refer to a cell with at least one construct deactivating the target gene as a *target cell.* All cells then undergo selection based on the phenotype, and the abundance of each RNAi construct type among the selected cells is measured. If the selection favors cells

0269-9648/15 $25.00

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

not exhibiting the phenotype (so-called negative selection), it will result in enrichment of the target cells, and hence in a relatively high count of the RNAi constructs corresponding to the target gene. A small number of genes exhibiting high RNAi counts (say, the three genes corresponding to the three highest counts) can then be validated in separate, individual (i.e., not pooled) RNAi experiments, in the hope that the target gene is among them. Conversely, under positive selection, some low-count genes need to be validated.

In most pooled RNAi screens, the phenotype of interest directly influences the survivability of the cells, so that the desired selection takes place automatically as a result of inserting the RNAi constructs into the cells. When this is not the case, it is sometimes possible to couple the phenotype with fluorescence, as measured in a flow cytometry experiment (e.g., Bassik et al. [1], Fellmann et al. [7]). In such an experiment, the cells are processed by a fluorescence-activated cell sorter (FACS), which first excites fluorescent-labeled molecules harbored in them, and then sorts the cells into two categories according to the resulting fluorescence intensity. The coupling means that the cells in which the target gene was deactivated (the target cells) tend to exhibit stronger fluorescence, so the entire process results in enrichment of the target cells. The relative abundance of the various construct types among the selected cells can then be measured, and a small number of genes corresponding to the top counts can be further validated, as described above.

In this work, we model probabilistically and optimize this FACS-aided pooled RNAi experiment. The main decision point in our analysis concerns the FACS selection criterion: if too many cells pass the selection, no detectable signal for the target gene will emanate from the construct counts; if the selection is too stringent, few or no target cells will be selected, resulting again in a failure to discover the target gene.

Several studies have dealt with statistical aspects of RNAi experiments (König et al. [10], Birmingham et al. [2], Bassik et al. [1], Hao et al. [9]). Our approach differs from those pursued in these studies, in that rather than being data driven, it models probabilistically the experiment starting from its fundamentals (e.g., the distribution of the number of constructs inserted to a cell, and the distribution of a cell's fluorescence intensity). In a broad sense, our work belongs to a line of works tackling problems in biotechnology and bioinformatics, using tools of operations research and applied probability (Piau [13], Strickland, Barnes, and Sokol [15], Blazewicz et al. [3], Łukasiak, Błażewicz, and Miłostan [11], Blazewicz et al. [4], Caserta and Voß [5]).

2. MODEL AND NOTATION

Let r be the number of genes considered. These genes may constitute the entire genome of the organism under study, or some sizable subset thereof (e.g., all genes related to signaling pathways). We index the genes by $i = 1, \ldots, r$, and, without loss of generality, designate the index $i = 1$ to the target gene. Let n be the number of cells sorted by FACS, indexed by $k = 1, \ldots, n$. In a typical experiment, r is in the thousands and n is in the millions.

Define $N_{k,i}$ to be the number of constructs of type i inserted into cell k, so that the target cells (those in which the target gene was deactivated) are those satisfying $N_{k,1} \geq 1$. Let F_k be the fluorescence intensity of cell k, and denote by G_1 and G_2 the cumulative distribution functions (CDFs) of the fluorescence intensity of the target and non-target cells, respectively. Then,

$$P(F_k \leq a) = \begin{cases} G_1(a) & N_{k,1} \geq 1, \\ G_2(a) & N_{k,1} = 0. \end{cases} \tag{1}$$

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

It is assumed that G_1 is larger than G_2 in some sense (e.g., via the usual stochastic order, whereby $G_1(a) \leq G_2(a)$ for all a). Also define

$$\overline{G}_1(a) = 1 - G_1(a), \quad \overline{G}_2(a) = 1 - G_2(a).$$

The experimenter may define the FACS selection criterion either through a percentile (i.e., the selected cells are the top t percents of the cells, in terms of their fluorescence intensity, for some $t \in (0,1)$) or through a fixed threshold (i.e., the selected cells are those whose fluorescence intensity exceeds some threshold α). The latter criterion is more tractable mathematically, so we adopt it henceforth. The resulting *construct count* corresponding to gene i is therefore

$$X_i = \sum_{k=1}^{n} N_{k,i} I_{\{F_k > \alpha\}}, \quad i = 1, \ldots, r.$$

The analysis below relies on the behavior of $M_{k,i} = N_{k,i} I_{\{F_k > \alpha\}}$, which is the contribution of cell k to the construct count X_i.

We consider two models for the process of inserting the RNAi constructs into the cells: the multinomial model, which is simpler to analyze, and the Poisson model, which is more realistic.

2.1. The Multinomial Model

In the multinomial model it is assumed that each cell always admits a single construct, so that $N_k = \sum_{i=1}^{r} N_{k,i} = 1$ for each cell $k = 1, \ldots, n$. The type of the construct in each cell is equally likely to be any of the r possible construct types, independent of other cells.

Define $X_0 = \sum_{k=1}^{n} I_{\{F_k < \alpha\}}$ to be the number of cells not selected by the FACS. Then, the joint distribution of the X_i is multinomial:

$$(X_0, X_1, X_2, \ldots, X_r) \sim \mathrm{Mult}(n, p_0, p_1, p_2, \ldots, p_r),$$

where

$$p_0 = \frac{1}{r} G_1(\alpha) + \frac{r-1}{r} G_2(\alpha), \quad p_1 = \frac{1}{r} \overline{G}_1(\alpha), \quad p_i = \frac{1}{r} \overline{G}_2(\alpha), \quad i \geq 2.$$

It is easily verified that under the multinomial model, for any power $m > 0$,

$$E[(M_{k,1})^m] = \frac{1}{r} \overline{G}_1(\alpha), \quad E[(M_{k,i})^m] = \frac{1}{r} \overline{G}_2(\alpha), \quad i \geq 2 \tag{2}$$

and

$$\mathrm{Var}(M_{k,1}) = \frac{1}{r} \overline{G}_1(\alpha) \left(1 - \frac{1}{r} \overline{G}_1(\alpha)\right) \tag{3}$$

$$\mathrm{Var}(M_{k,i}) = \frac{1}{r} \overline{G}_2(\alpha) \left(1 - \frac{1}{r} \overline{G}_2(\alpha)\right), \quad i \geq 2 \tag{4}$$

$$\mathrm{Cov}(M_{k,1}, M_{k,i}) = -\frac{1}{r^2} \overline{G}_1(\alpha) \overline{G}_2(\alpha), \quad i \geq 2 \tag{5}$$

$$\mathrm{Cov}(M_{k,i}, M_{k,j}) = -\frac{1}{r^2} \overline{G}_2(\alpha)^2, \quad i, j \geq 2, \quad i \neq j. \tag{6}$$

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

2.2. The Poisson Model

Under the Poisson model, the process of preparing the cells with the constructs for the FACS is comprised of two steps. In the first step, each cell admits a Poisson number of constructs, with parameter λ that is called "multiplicity of infection." The value of λ is typically low, in the range 0.1–1, and we treat it as exogenously given, rather than as a decision variable. As in the multinomial model, the type of each construct is assumed to be drawn uniformly from $\{1, \ldots, r\}$, independent of other constructs. Because the support of the Poisson distribution includes the value 0, a cell may contain no constructs, in which case it will contribute no useful information for the experiment. To avoid this, in the second step, all cells having no constructs are eliminated, and only those with at least one construct are processed by the FACS machine. Thus, the total number of constructs per cell has a Poisson distribution truncated below 1.

Proposition 1: *In the Poisson model, the contribution $M_{k,i}$ of cell k to the construct count X_i satisfies*

$$\begin{aligned}
E(M_{k,1}) &= \frac{\lambda c_1}{r(1-e^{-\lambda})},\\
E(M_{k,i}) &= \frac{\lambda c_2}{r(1-e^{-\lambda})}, \quad i \geq 2,\\
\mathrm{Var}(M_{k,1}) &= \frac{c_1}{1-e^{-\lambda}}\left(\frac{\lambda}{r}+\frac{\lambda^2}{r^2}\right)-\left(\frac{\lambda c_1}{r(1-e^{-\lambda})}\right)^2,\\
\mathrm{Var}(M_{k,i}) &= \frac{c_2}{1-e^{-\lambda}}\left(\frac{\lambda}{r}+\frac{\lambda^2}{r^2}\right)-\left(\frac{\lambda c_2}{r(1-e^{-\lambda})}\right)^2, \quad i \geq 2,\\
\mathrm{Cov}(M_{k,1}, M_{k,i}) &= \frac{c_1\lambda^2}{r^2(1-e^{-\lambda})}-\frac{c_1c_2\lambda^2}{r^2(1-e^{-\lambda})^2} \quad i \geq 2,\\
\mathrm{Cov}(M_{k,i}, M_{k,j}) &= \frac{c_2\lambda^2}{r^2(1-e^{-\lambda})}-\frac{c_2^2\lambda^2}{r^2(1-e^{-\lambda})^2} \quad i,j \geq 2, \quad i \neq j,\\
E[(M_{k,1})^3] &= \frac{c_1}{1-e^{-\lambda}}\left[\frac{\lambda}{r}+3\left(\frac{\lambda}{r}\right)^2+\left(\frac{\lambda}{r}\right)^3\right],\\
E[(M_{k,i})^3] &= \frac{c_2}{1-e^{-\lambda}}\left[\frac{\lambda}{r}+3\left(\frac{\lambda}{r}\right)^2+\left(\frac{\lambda}{r}\right)^3\right] \quad i \geq 2.
\end{aligned}$$

where

$$\begin{aligned}
c_1 &= c_1(\alpha) = \overline{G}_1(\alpha),\\
c_2 &= c_2(\alpha) = \overline{G}_2(\alpha)e^{-\lambda/r} + \overline{G}_1(\alpha)(1-e^{-\lambda/r}).
\end{aligned}$$

Proof: Because of the elimination of cells having zero constructs, the counts $N_{k,1}, \ldots, N_{k,r}$ at each cell k satisfy

$$(N_{k,1}, \ldots, N_{k,r}) \overset{d}{=} (\widehat{N}_{k,1}, \ldots, \widehat{N}_{k,r}) \mid \widehat{N}_k \geq 1, \quad k = 1, \ldots, n,$$

where $\overset{d}{=}$ denotes equality in distribution, $\widehat{N}_{k,1}, \ldots, \widehat{N}_{k,r}$ are independent Poisson (λ/r) random variables, and $\widehat{N}_k = \sum_{i=1}^r \widehat{N}_{k,i} \sim \mathrm{Poisson}(\lambda)$.

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

The distribution of a cell's fluorescence depends only on the presence of constructs of type 1 (recall Eq. (1)), so for $x \geq 1$ we have

$$P(F_k > \alpha \,|\, \widehat{N}_{k,1} = x) = \overline{G}_1(\alpha) = c_1, \tag{7}$$

and for $i \geq 2$, since $\widehat{N}_{k,1}$ and $\widehat{N}_{k,i}$ are independent,

$$\begin{aligned} P(F_k > \alpha \,|\, \widehat{N}_{k,i} = x) &= P(F_k > \alpha \,|\, \widehat{N}_{k,i} = x, \widehat{N}_{k,1} = 0) P(\widehat{N}_{k,1} = 0 \,|\, \widehat{N}_{k,i} = x) \\ &\quad + P(F_k > \alpha \,|\, \widehat{N}_{k,i} = x, \widehat{N}_{k,1} \geq 1) P(\widehat{N}_{k,1} \geq 1 \,|\, \widehat{N}_{k,i} = x) \\ &= P(F_k > \alpha \,|\, \widehat{N}_{k,1} = 0) P(\widehat{N}_{k,1} = 0) \\ &\quad + P(F_k > \alpha \,|\, \widehat{N}_{k,1} \geq 1) P(\widehat{N}_{k,1} \geq 1) \\ &= \overline{G}_2(\alpha) e^{-\lambda/r} + \overline{G}_1(\alpha)(1 - e^{-\lambda/r}) \\ &= c_2. \end{aligned} \tag{8}$$

Using (7), we have for $x \geq 1$,

$$\begin{aligned} P(M_{k,1} = x) &= P(N_{k,1} = x, F_k > \alpha) \\ &= P(\widehat{N}_{k,1} = x, F_k > \alpha \,|\, \widehat{N}_k \geq 1) \\ &= \frac{P(\widehat{N}_{k,1} = x, F_k > \alpha)}{P(\widehat{N}_k \geq 1)} \\ &= \frac{P(\widehat{N}_{k,1} = x) P(F_k > \alpha \,|\, \widehat{N}_{k,1} = x)}{P(\widehat{N}_k \geq 1)} \\ &= \frac{e^{-\lambda/r} (\lambda/r)^x c_1}{x!(1 - e^{-\lambda})}. \end{aligned}$$

Similarly, using (8), for $i, j \geq 2$, $i \neq j$ and $x, y \geq 1$ we get

$$\begin{aligned} P(M_{k,i} = x) &= \frac{e^{-\lambda/r} (\lambda/r)^x c_2}{x!(1 - e^{-\lambda})}, \\ P(M_{k,1} = x, M_{k,i} = y) &= \frac{e^{-2\lambda/r} (\lambda/r)^{x+y} c_1}{x!y!(1 - e^{-\lambda})}, \\ P(M_{k,i} = x, M_{k,j} = y) &= \frac{e^{-2\lambda/r} (\lambda/r)^{x+y} c_2}{x!y!(1 - e^{-\lambda})}. \end{aligned}$$

Computing now the moments through their basic definitions – e.g., $E(M_{k,1}) = \sum_{x=1}^{\infty} x P(M_{k,1} = x)$ – the proposition is proved. ■

When λ is near zero, there is a low probability that a cell will admit two constructs or more. Because cells with no constructs are eliminated, the probability in this case of having eventually a single construct is close to 1, similar to the multinomial model, in which there is always a single construct in each cell. The next proposition formalizes this observation, and asserts that the entire Poisson model converges in distribution to the multinomial model as $\lambda \to 0$. This proposition is the only place in this work in which the multinomial and the Poisson models are considered simultaneously; to distinguish between them notationally, we attach a superscript (λ) to all random variables related to the Poisson model.

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

Proposition 2: *Let $X_1, X_2, \ldots, X_r$ denote the construct counts under the multinomial model, and let $X_1^{(\lambda)}, X_2^{(\lambda)}, \ldots, X_r^{(\lambda)}$ denote the construct counts under the Poisson model. Then,*

$$(X_1^{(\lambda)}, X_2^{(\lambda)}, \ldots, X_r^{(\lambda)}) \Rightarrow (X_1, X_2, \ldots, X_r) \quad \textit{as } \lambda \to 0.$$

Proof: As in the proof of Proposition 1, let $\widehat{N}_k^{(\lambda)} = \sum_{i=1}^{r} \widehat{N}_{k,i}^{(\lambda)} \sim \text{Poisson}(\lambda)$ be the total number of constructs inserted into cell k *before* eliminating the empty cells, under the Poisson model. Using l'Hospital's rule, we have

$$\begin{aligned}\lim_{\lambda \to 0} P(N_k^{(\lambda)} = 1) &= \lim_{\lambda \to 0} P(\widehat{N}_k^{(\lambda)} = 1 \,|\, \widehat{N}_k^{(\lambda)} \geq 1) \\ &= \lim_{\lambda \to 0} \frac{P(\widehat{N}_k^{(\lambda)} = 1)}{P(\widehat{N}_k^{(\lambda)} \geq 1)} \\ &= \lim_{\lambda \to 0} \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} \\ &= \lim_{\lambda \to 0} \frac{e^{-\lambda} - \lambda e^{-\lambda}}{e^{-\lambda}} \\ &= 1.\end{aligned}$$

Thus, $N_k^{(\lambda)} \Rightarrow 1$ as $\lambda \to 0$ for each cell k. The result now follows from the Continuous Mapping Theorem. ■

2.3. Maximizing the Probability of Discovery

Under either the multinomial or the Poisson model, let $X_{[1]} \geq X_{[2]} \geq \cdots \geq X_{[r-1]}$ be the order statistics of the $r-1$ non-target construct counts $X_2, \ldots, X_r$, sorted from largest to smallest. Also let v denote the number of genes to be validated; this number is assumed to be exogenously given, according to budget constraints. The target gene is discovered in the experiment if it is among the v genes that are validated, an event that occurs if the construct count of the target gene is among the v top counts. Mathematically, the *probability of discovery* is

$$p_{\text{disc}}(\alpha) = P(X_{[v]} < X_1). \tag{9}$$

Our goal is to find a threshold α^* that maximizes this probability, that is, that satisfies

$$\alpha^* = \arg\max_{\alpha} p_{\text{disc}}(\alpha).$$

3. ASYMPTOTIC ANALYSIS

Let n, the number of cells, approach infinity, and assume that the number of genes grows to infinity with n, that is, $r = r(n) \to \infty$ as $n \to \infty$. We attach a superscript n to all random variables defined above, so that $X_i^n = \sum_{k=1}^{n} M_{k,i}^n$ is the construct count corresponding to

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

gene i in the nth system. Let

$$Y_i^n = \frac{X_i^n - E(X_i^n)}{\sqrt{\mathrm{Var}(X_i^n)}}, \quad i = 1, 2, \ldots, r(n), \tag{10}$$

be the normalized construct counts, and define the process

$$\mathbf{Y}^n = (Y_1^n, Y_2^n, \ldots, Y_{r(n)}^n, 0, 0, \ldots). \tag{11}$$

The following result asserts that if $r(n)$ approaches infinity slower than n, then the scaled construct counts are asymptotically normal and independent.

Proposition 3: *Under both the multinomial and the Poisson models, and for fixed α, if $r(n) \to \infty$ and $n/r(n) \to \infty$ as $n \to \infty$, then*

$$\mathbf{Y}^n \Rightarrow (Z_1, Z_2, \ldots) \quad \textit{as } n \to \infty,$$

where the Z_i are independent standard normal random variables.

Proof: By Theorem 1.4.8 of van der Vaart and Wellner [16], it is enough to prove finite-dimensional convergence, that is, to show that for each $d \in \mathbb{N}$,

$$Y^n = (Y_1^n, \ldots, Y_d^n)^T \Rightarrow (Z_1, \ldots, Z_d)^T.$$

By the Cramér–Wold theorem, it needs to be shown that for each $a \in \mathbb{R}^d$,

$$a^T Y^n \Rightarrow N(0, a^T a). \tag{12}$$

Define

$$Q_{k,i}^n = \frac{M_{k,i}^n - E(M_{k,i}^n)}{\sqrt{n\mathrm{Var}(M_{k,i}^n)}}.$$

Then,

$$E(Q_{k,i}^n) = 0, \tag{13}$$

$$\mathrm{Var}(Q_{k,i}^n) = 1/n, \tag{14}$$

$$\mathrm{Cov}(Q_{k,1}^n, Q_{k,i}^n) = \frac{\mathrm{Cov}(M_{k,1}^n, M_{k,2}^n)}{n\sqrt{\mathrm{Var}(M_{k,1}^n)\mathrm{Var}(M_{k,2}^n)}}, \quad i \geq 2 \tag{15}$$

$$\mathrm{Cov}(Q_{k,i}^n, Q_{k,j}^n) = \frac{\mathrm{Cov}(M_{k,2}^n, M_{k,3}^n)}{n\mathrm{Var}(M_{k,2}^n)}, \quad i, j \geq 2, \ \ i \neq j. \tag{16}$$

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

Define $V_k^n = \sum_{i=1}^d a_i Q_{k,i}^n$. Then,

$$\begin{aligned}\sum_{k=1}^n V_k^n &= \sum_{k=1}^n \sum_{i=1}^d a_i Q_{k,i}^n \\ &= \sum_{i=1}^d a_i \sum_{k=1}^n \frac{M_{k,i}^n - E(M_{k,i}^n)}{\sqrt{n\mathrm{Var}(M_{k,i}^n)}} \\ &= \sum_{i=1}^d a_i \frac{X_i^n - nE(M_{k,i}^n)}{\sqrt{n\mathrm{Var}(M_{k,i}^n)}} \\ &= a^T Y^n.\end{aligned}$$

Thus, proving (12) is equivalent to proving $\sum_{k=1}^n V_k^n \Rightarrow N(0, a^T a)$. Now consider the triangular array $\{V_k^n,\ k = 1, \ldots, n,\ n = 1, 2, \ldots\}$. Using (13)–(16), we have that $E(V_k^n) = 0$ and

$$\begin{aligned}\mathrm{Var}(V_k^n) &= \sum_{i=1}^d \mathrm{Var}(a_i Q_{k,i}^n) + \sum_{i \neq j} \mathrm{Cov}(a_i Q_{k,i}^n, a_j Q_{k,j}^n) \\ &= \frac{1}{n}\sum_{i=1}^d a_i^2 + \frac{2}{n}\left[\frac{\mathrm{Cov}(M_{k,1}^n, M_{k,2}^n)}{\sqrt{\mathrm{Var}(M_{k,1}^n)\mathrm{Var}(M_{k,2}^n)}} \sum_{i=2}^d a_1 a_i + \frac{\mathrm{Cov}(M_{k,2}^n, M_{k,3}^n)}{\mathrm{Var}(M_{k,2}^n)} \sum_{\substack{i,j \geq 2 \\ i<j}} a_i a_j\right].\end{aligned}$$

Therefore,

$$\begin{aligned}s_n^2 &= \mathrm{Var}\left(\sum_{k=1}^n V_k^n\right) \\ &= n\mathrm{Var}(V_1^n) \\ &= \sum_{i=1}^d a_i^2 + 2\left[\frac{\mathrm{Cov}(M_{k,1}^n, M_{k,2}^n)}{\sqrt{\mathrm{Var}(M_{k,1}^n)\mathrm{Var}(M_{k,2}^n)}} \sum_{i=2}^d a_1 a_i + \frac{\mathrm{Cov}(M_{k,2}^n, M_{k,3}^n)}{\mathrm{Var}(M_{k,2}^n)} \sum_{\substack{i,j \geq 2 \\ i<j}} a_i a_j\right]. \qquad \textbf{(17)}\end{aligned}$$

Using Eqs. (3)–(6) for the multinomial model, or Proposition 1 for the Poisson model, we have that the coefficients before both sums in the square brackets in the last expression converge to zero as $n \to \infty$, as the numerator in each is $O(1/r^2)$, and the denominator is $O(1/r)$. Therefore, $s_n^2 \to \sum_{i=1}^d a_i^2$ as $n \to \infty$.

By the Lindeberg–Feller Central Limit Theorem, a sufficient condition for

$$\frac{1}{s_n}\sum_{k=1}^n V_k^n \Rightarrow N(0,1) \qquad \textbf{(18)}$$

is the Lyapunov condition:

$$\text{there exists } \delta > 0 \text{ such that} \quad \frac{1}{s_n^{2+\delta}} \sum_{k=1}^n E\left(|V_k^n|^{2+\delta}\right) \to 0 \quad \text{as } n \to \infty.$$

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

We now show that the condition holds for $\delta = 1$. Since $V_1^n, \ldots, V_n^n$ are identically distributed for each n, the Lyapunov condition in the case $\delta = 1$ reduces to

$$\frac{n}{s_n^3} E\left(|V_1^n|^3\right) \to 0 \quad \text{as } n \to \infty.$$

Using Minkowski inequality and the fact that the $M_{k,i}^n$ are non-negative, we have that

$$\begin{aligned}\frac{n}{s_n^3} E\left(|V_1^n|^3\right) &= \frac{n}{s_n^3} E\left(\left|\sum_{i=1}^d a_i \frac{M_{1,i}^n - E(M_{1,i}^n)}{\sqrt{n\mathrm{Var}(M_{1,i}^n)}}\right|^3\right)\\ &= \frac{1}{s_n^3} E\left(\left|\sum_{i=1}^d \frac{a_i[M_{1,i}^n - E(M_{1,i}^n)]}{n^{1/6}\sqrt{\mathrm{Var}(M_{1,i}^n)}}\right|^3\right)\\ &\leq \frac{1}{s_n^3}\left[\sum_{i=1}^d \left(E\left|\frac{a_i[M_{1,i}^n - E(M_{1,i}^n)]}{n^{1/6}\sqrt{\mathrm{Var}(M_{1,i}^n)}}\right|^3\right)^{1/3}\right]^3\\ &\leq \frac{1}{s_n^3}\left[\sum_{i=1}^d \left(E\left|\frac{a_i M_{1,i}^n}{n^{1/6}\sqrt{\mathrm{Var}(M_{1,i}^n)}}\right|^3\right)^{1/3}\right]^3.\end{aligned}$$

Recall that s_n^2 converges to a constant. Thus, since the sum in the last expression involves a finite and fixed number of summands, to show that the last expression converges to zero, it is enough to show that for each i, the expression

$$E\left|\frac{a_i M_{1,i}^n}{n^{1/6}\sqrt{\mathrm{Var}(M_{1,i}^n)}}\right|^3 = \frac{|a_i^3| E[(M_{1,i}^n)^3]}{n^{1/2}[\mathrm{Var}(M_{1,i}^n)]^{3/2}} \tag{19}$$

converges to zero. Indeed, using Eqs. (2)–(6) for the multinomial model, or Proposition 1 for the Poisson model, we have that $E[(M_{1,i}^n)^3]$ converges to zero at rate $1/r$, whereas $[\mathrm{Var}(M_{1,i}^n)]^{3/2}$ does so at rate $1/r^{3/2}$. Thus, the entire right-hand side of the last displayed equation is of order $(r/n)^{1/2}$, and since we assumed that $n/r \to \infty$, the Lyapunov condition is satisfied.

We have shown that (18) holds. What we need to show is (12), which may be written equivalently as

$$\frac{1}{\|a\|_2} \sum_{k=1}^n V_k^n \Rightarrow N(0,1).$$

However, since $s_n \to (\sum_{i=1}^d a_i^2)^{1/2} = \|a\|_2$, the proposition is proved. ■

4. APPROXIMATING THE PROBABILITY OF DISCOVERY

Under either the multinomial or the Poisson model, evaluating the exact probability of discovery $p_{\text{disc}}(\alpha)$ in (9) is difficult, because of the dependence among the X_i. We therefore

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

use the asymptotic result of the previous section to derive an approximation to $p_{\text{disc}}(\alpha)$. For fixed n and r, let $\widetilde{X}_1, \ldots, \widetilde{X}_r$ be independent normal random variables, with $E(\widetilde{X}_i) = E(X_i)$ and $\text{Var}(\widetilde{X}_i) = \text{Var}(X_i)$. By Proposition 3, when both n and r are large, but $r \ll n$, these $\widetilde{X}_1, \ldots, \widetilde{X}_r$ may serve as approximations to $X_1, \ldots, X_r$. Let ϕ_1 be the density function of $\widetilde{X}_1$, and Φ_2 be the CDF of $\widetilde{X}_i$ for $i \geq 2$ (recall that $X_2, \ldots, X_r$ are identically distributed); note that both ϕ_1 and Φ_2 depend on α. Also let $\widetilde{X}_{[1]} \geq \widetilde{X}_{[2]} \geq \cdots \geq \widetilde{X}_{[r-1]}$ be the order statistics of $\widetilde{X}_2, \ldots, \widetilde{X}_r$, and define

$$\widetilde{p}_{\text{disc}}(\alpha) = P(\widetilde{X}_{[v]} < \widetilde{X}_1) \tag{20}$$

to be the approximation to the probability of discovery $p_{\text{disc}}(\alpha)$ in (9).

Proposition 4:

$$\widetilde{p}_{\text{disc}}(\alpha) = \int_{-\infty}^{\infty} \sum_{j=r-v}^{r-1} \binom{r-1}{j} [\Phi_2(x)]^j [1 - \Phi_2(x)]^{r-j-1} \phi_1(x)\, dx.$$

Proof: Because $\widetilde{X}_2, \ldots, \widetilde{X}_r$ are iid with CDF Φ_2, the CDF of $\widetilde{X}_{[v]}$ is

$$P(\widetilde{X}_{[v]} \leq x) = \sum_{j=r-v}^{r-1} \binom{r-1}{j} [\Phi_2(x)]^j [1 - \Phi_2(x)]^{r-j-1}.$$

See p. 87 of Serfling [14]. Conditioning on the value of $\widetilde{X}_1$ in the right-hand side of (20) and integrating with respect to its density ϕ_1, the proposition is proved. ■

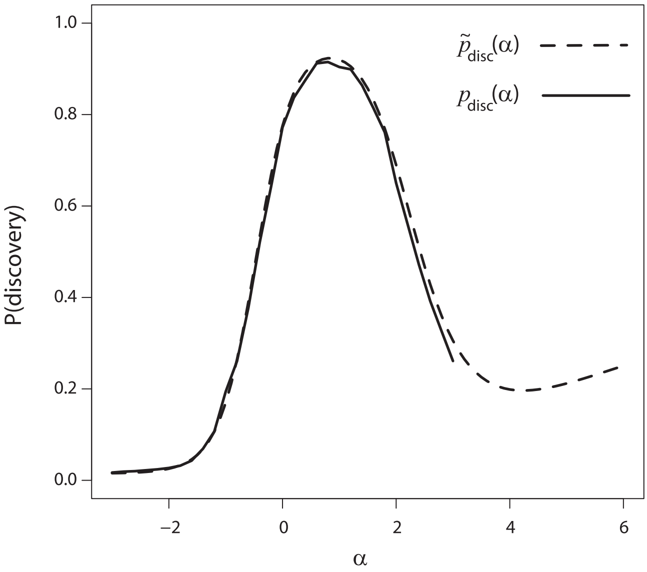

FIGURE 1. The probability of discovering the target gene as a function of the selection threshold α. Solid curve is the true probability of discovery for the exact system, $p_{\text{disc}}(\alpha)$, as estimated by simulation. The dashed curve is the approximate probability of discovery, $\widetilde{p}_{\text{disc}}(\alpha)$.

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

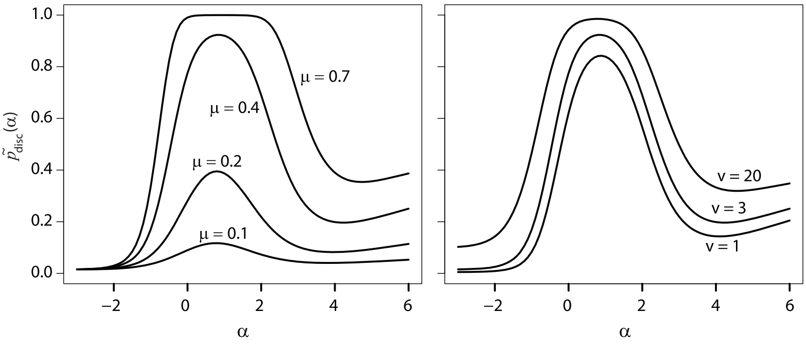

FIGURE 2. The approximate probability of discovery $\widetilde{p}_{\text{disc}}(\alpha)$ as a function of α, for various system parameters. Left panel: $G_1 = N(\mu, 1)$ for various mean values μ, and $G_2 = N(0, 1)$. Right panel: various values of v, the number of genes to be validated.

The dashed curve in Figure 1 shows $\widetilde{p}_{\text{disc}}(\alpha)$, the approximate probability of discovery, as a function of α. The solid curve is the true probability of discovery, $p_{\text{disc}}(\alpha)$, as estimated by simulation. The system parameters are $r = 200$ genes, $n = 40,000$ cells, $v = 3$ genes to be validated, fluorescence distributions $G_1 = N(0.4, 1)$ and $G_2 = N(0, 1)$, and the multinomial model.

The main feature of Figure 1 is that the approximate curve follows the exact curve very closely. Thus, the asymptotics-based approximation works well in practice. The optimal threshold α^* for this system parameters is about 0.8. Note also that the curve of $p_{\text{disc}}(\alpha)$ is plotted only for $\alpha \leq 3$; the reason for this is that for $\alpha > 3$ the selection criterion is too stringent, so that in practice no cells satisfy it, and the construct counts – which are always integer in the exact system – are all zero. In contrast, the counts in the approximate system are continuous random variables, which may assume near-zero values, and so $\widetilde{p}_{\text{disc}}(\alpha)$ can be computed for any α.

Figure 2 shows how the curve $\widetilde{p}_{\text{disc}}(\alpha)$ changes with the system parameters. The left panel shows the influence of the separation between the two fluorescence distributions G_1 and G_2, and the right panel the influence of v, the number of genes to be validated. As expected, better separation and higher v result in higher probabilities of discovery.

5. THE TWO-STAGE DISCOVERY PROBLEM

After enriching the target cells by FACS, it is possible to enrich them further, by first growing the selected cells until their population is large enough, and then processing them in a second FACS round.

We model this two-stage process as follow. We let n be the number of cells processed by FACS in the first stage. As before, $N_{k,i}$ denotes the number of constructs of type i inserted into cell k (according to either the multinomial model or the Poisson model), and F_k denotes the fluorescence intensity of cell k, which is determined according to (1). The selection criterion for cell k in the first stage is $F_k > \alpha$, for some threshold α. Each selected cell from the first stage gives rise to L descendant cells, to be processed in the second stage; all L descendants of the same ancestor cell inherit the RNAi construct content of their ancestor. The fluorescence intensity of the lth descendant of cell k is measured in the second FACS stage, and is denoted by $F_{k,l}$; the distribution of $F_{k,l}$ is the same as that of

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

the ancestor cell, that is,

$$P(F_{k,l} \leq a) = \begin{cases} G_1(a) & N_{k,1} \geq 1, \\ G_2(a) & N_{k,1} = 0. \end{cases}$$

For each ancestor cell k, the L fluorescence intensities $F_{k,1}, \ldots, F_{k,L}$ of the L descendant cells are conditionally independent given $N_{k,1}$. The selected cells in the second stage are those satisfying $F_{k,l} > \beta$, for some threshold β. The final construct counts are given by

$$X_i = \sum_{k=1}^{n} \sum_{l=1}^{L} N_{k,i} I_{\{F_k > \alpha,\ F_{k,l} > \beta\}}, \quad i = 1, \ldots, r.$$

Let $T_{k,i} = N_{k,i} I_{\{F_k > \alpha\}} \sum_{l=1}^{L} I_{\{F_{k,l} > \beta\}}$, so that $X_i = \sum_{k=1}^{n} T_{k,i}$. The $T_{k,i}$ are thus the counterparts of the $M_{k,i}$ from the single-stage problem.

Proposition 5: *Under the multinomial model, the $T_{k,i}$ satisfy*

$$\begin{aligned}
E(T_{k,1}) &= \frac{1}{r} L \overline{G}_1(\alpha) \overline{G}_1(\beta), \\
E(T_{k,i}) &= \frac{1}{r} L \overline{G}_2(\alpha) \overline{G}_2(\beta) \quad i \geq 2, \\
\mathrm{Var}(T_{k,1}) &= \frac{1}{r} \overline{G}_1(\alpha) \left[L G_1(\beta) \overline{G}_1(\beta) + L^2 \overline{G}_1(\beta)^2 \right] - \frac{1}{r^2} \overline{G}_1(\alpha)^2 L^2 \overline{G}_1(\beta), \\
\mathrm{Var}(T_{k,i}) &= \frac{1}{r} \overline{G}_2(\alpha) \left[L G_2(\beta) \overline{G}_2(\beta) + L^2 \overline{G}_2(\beta)^2 \right] - \frac{1}{r^2} \overline{G}_2(\alpha)^2 L^2 \overline{G}_2(\beta) \quad i \geq 2, \\
\mathrm{Cov}(T_{k,1}, T_{k,i}) &= -\frac{1}{r^2} L^2 \overline{G}_1(\alpha) \overline{G}_1(\beta) \overline{G}_2(\alpha) \overline{G}_2(\beta) \quad i \geq 2, \\
\mathrm{Cov}(T_{k,i}, T_{k,j}) &= -\frac{1}{r^2} L^2 \overline{G}_2(\alpha)^2 \overline{G}_2(\beta)^2 \quad i, j \geq 2, \quad i \neq j, \\
E[(T_{k,1})^3] &= \frac{1}{r} L \overline{G}_1(\alpha) \overline{G}_1(\beta) \left(L^2 \overline{G}_1(\beta)^2 - 3L \overline{G}_1(\beta)^2 + 2\overline{G}_1(\beta)^2 \right. \\
&\quad \left. + 3L \overline{G}_1(\beta) - 3\overline{G}_1(\beta) + 1 \right), \\
E[(T_{k,i})^3] &= \frac{1}{r} L \overline{G}_2(\alpha) \overline{G}_2(\beta) \left(L^2 \overline{G}_2(\beta)^2 - 3L \overline{G}_2(\beta)^2 + 2\overline{G}_2(\beta)^2 \right. \\
&\quad \left. + 3L \overline{G}_2(\beta) - 3\overline{G}_2(\beta) + 1 \right), \quad i \geq 2,
\end{aligned}$$

Proof: Under the multinomial model, $T_{k,i}$ is binomial conditional on $N_{k,i} I_{\{F_k > \alpha\}} = 1$, and zero otherwise:

$$\begin{aligned}
T_{k,1} \,|\, N_{k,1} I_{\{F_k > \alpha\}} = 1 &\sim \mathrm{Bin}(L, \overline{G}_1(\beta)), \\
T_{k,i} \,|\, N_{k,i} I_{\{F_k > \alpha\}} = 1 &\sim \mathrm{Bin}(L, \overline{G}_2(\beta)), \quad i \geq 2.
\end{aligned}$$

The moments of $T_{k,i}$ are then the well known moments of the binomial distribution, multiplied either by $P(N_{k,1} I_{\{F_k > \alpha\}} = 1) = \overline{G}_1(\alpha)/r$ for $i = 1$, or by $P(N_{k,i} I_{\{F_k > \alpha\}} = 1) = \overline{G}_2(\alpha)/r$ for $i \geq 2$. ■

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

As in the single-stage problem, we let both n and $r = r(n)$ approach infinity, and define the normalized construct count Y_i^n through (10), and the process $\mathbf{Y}^n$ through (11). The following result is the two-stage counterpart of Proposition 3, and asserts that the scaled construct counts are again asymptotically normal and independent.

Proposition 6: *Under both the multinomial and the Poisson models, and for fixed thresholds α and β, if $r(n) \to \infty$ and $n/r(n) \to \infty$ as $n \to \infty$, then*

$$\mathbf{Y}^n \Rightarrow (Z_1, Z_2, \ldots) \quad \text{as } n \to \infty,$$

where the Z_i are independent standard normal random variables.

Proof: For brevity, we prove the proposition only for the multinomial model. The proof follows the exact same steps as that of Proposition 3, with the $T_{k,i}$ replacing the $M_{k,i}$. Only two points need to be reestablished: the first is that the coefficients before both sums in the square brackets in equation (17) converge to zero as $n \to \infty$; this is true since by Proposition 5, we again have that the numerator of each is $O(1/r^2)$, whereas the denominator is $O(1/r)$. The second point is that for each i, the expression at the right-hand side of Eq. (19) converge to zero as $n \to \infty$; this is again true since by Proposition 5, that expression is $O((r/n)^{1/2})$, and by assumption, $n/r \to \infty$. ■

The decision variables in the two-stage model are the thresholds α and β. Clearly, it is desirable to enrich the first-stage selected cells as much as possible, and this can be done by raising α. However, as in the single-stage problem, setting α too high may result in no target cells (and hence no constructs of type 1) among the selected cells. We resolve this conflict by maximizing α subject to a constraint that ensures that the number of selected target cells is high enough. Let

$$W_1 = W_1(\alpha) = \sum_{k=1}^{n} I_{\{F_k > \alpha,\ N_{k,1} \geq 1\}}$$

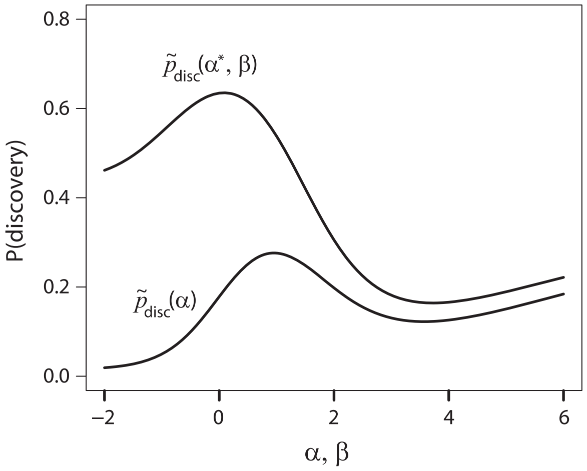

FIGURE 3. The probability of discovery as a function of α in a single-stage system (lower curve) and as a function of β in a two-stage system (upper curve).

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

be the number of target cells selected by the FACS. Under the multinomial model, for example, $W_1 \sim \text{Bin}(n, \overline{G}_1(\alpha)/r)$. We may then set α to be maximal subject to $E(W_1) \geq b$, or to $P(W_1 \geq b) \geq 1 - \epsilon$, for some b and ϵ.

As in the single-stage problem, we let $\widetilde{p}_{\text{disc}}(\alpha, \beta)$ be the approximate probability of discovery, based on the normal approximation from Proposition 6. The solid curve in Figure 3 depicts $\widetilde{p}_{\text{disc}}(\alpha^*, \beta)$ as a function of β, where $\alpha^* = 0.55$ is the maximal α satisfying $E(W_1) \geq 10$, and for a multinomial system with parameters $r = 200$, $n = 5000$, $v = 3$, $G_1 = N(0.3, 1)$, and $G_2 = N(0, 1)$. The parameter L was set to 4, the value required so that the expected number of cells processed in the second stage is roughly 5000. The dashed curve is $\widetilde{p}_{\text{disc}}(\alpha)$ as a function of α for a single-stage system with the same parameters, except for $n = 10,000$ (so that the total number of cells processed by FACS in the two systems is roughly same). Dividing the screening between two stages improves significantly the probability of discovering the target gene: the maximal probability of discovery in the two-stage system is 0.64 (achieved by $\beta^* = 0.1$), whereas in the single-stage system, it is 0.28 (achieved by $\alpha^* = 0.9$).

6. DISCUSSION

In this paper, we modeled and analyzed probabilistically FACS-based RNAi genetic screening experiments. The key decision variable in the analysis is the FACS selection threshold α, which needs to be set optimally so as to maximize the probability of discovering the target gene. This probability of discovery is determined by two factors: the number of the selected cells, and the enrichment level (the proportion of the target cells among the selected cells). The strong law of large numbers guarantees that when the enrichment level is fixed, the probability of discovery approaches 1 as the number of the selected cells increases; clearly, when the number of selected cells is fixed, increasing the enrichment level also results in a higher probability of discovery. Raising α, therefore, has two contradicting effects on the probability of discovery, as it both decreases the number of selected cells, and increases the enrichment level. The optimal α^* balances these opposing requirements, and can be determined through our normal approximation.

The two fluorescence distributions G_1 and G_2 were not assumed to be of any specific type in our analysis. Furusawa et al. [8] advocate using a log-normal distribution to model FACS fluorescence readings. However, since the FACS selection process is ordinal, the entire analysis is invariant under monotonically increasing transformations of the fluorescence distributions. Log-normal distributions may thus be converted to normal ones, as we used in our simulations.

In Section 5 of this paper, we studied a two-stage version of the discovery problem. In principle, it is possible to repeat the enrichment–reproduction process multiple times, rather than just two, to increase further the probability of discovery. However, each such repetition increases the likelihood of introducing a contamination into the cell population, in which case the entire experiment is lost. We follow therefore Bassik et al. [1], and study only the single- and two-stage versions of the problem.

This paper is concerned with the stochastic modeling and analysis of RNAi experiments, and the statistical aspects of the problem are beyond its scope. These aspects, however, deserve study: for example, the uncertainty resulting from estimating G_1 and G_2 can be accounted for in a more detailed analysis, and so is the noise inherent to measuring the construct counts. We plan to study such statistical aspects, in conjunction with the above model, in a sequel to this work.

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press

Acknowledgments

We thank Ben-Zion Levi, Nitsan Fourier, and Ofer Barnea-Yizhar from the Faculty of Biotechnology and Food Engineering at the Technion, Israel, for introducing us to the research problem and for useful discussions. The research of Yair Goldberg was supported in part by the Israeli Science Foundation grant 1308/12. The research of Yuval Nov was supported in part by the Israeli Science Foundation grant 286/13.

REFERENCES

1. Bassik, M.C., Lebbink, R.J., Churchman, L.S., Ingolia, N.T., Patena, W., LeProust, E.M., Schuldiner, M., Weissman, J.S. & McManus, M.T. (2009). Rapid creation and quantitative monitoring of high coverage shRNA libraries. *Nature Methods*, 6(6): 443–445.
2. Birmingham, A., Selfors, L.M., Forster, T., Wrobel, D., Kennedy, C.J., Shanks, E., Santoyo-Lopez, J., Dunican, D.J., Long, A., Kelleher, D., *et al.* (2009). Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods*, 6(8): 569–575.
3. Blazewicz, J., Oguz, C., Swiercz, A. & Weglarz, J. (2006). DNA sequencing by hybridization via genetic search. *Operations Research*, 54(6): 1185–1192.
4. Blazewicz, J., Burke, E.K., Kendall, G., Mruczkiewicz, W., Oguz, C. & Swiercz, A. (2013). A hyper-heuristic approach to sequencing by hybridization of dna sequences. *Annals of Operations Research*, 207(1): 27–41.
5. Caserta, M. & Voß, S. (2014). A hybrid algorithm for the DNA sequencing problem. *Discrete Applied Mathematics*, 163: 87–99.
6. Dykxhoorn, D.M. and Lieberman, J. (2005). The silent revolution: RNA interference as basic biology, research tool, and therapeutic. *Annual Review of Medicine*, 56: 401–423.
7. Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., et al. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Molecular Cell*, 41(6): 733–746.
8. Furusawa, C., Suzuki, T., Kashiwagi, A., Yomo, T. & Kaneko, K. (2005). Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *Biophysics*, 1(0): 25–31.
9. Hao, L., He, Q., Wang, Z., Craven, M., Newton, M.A. & Ahlquist, P. (2013). Limited agreement of independent RNAi screens for virus-required host genes owes more to false-negative than false-positive factors. *PLoS Computational Biology*, 9(9): e1003235.
10. König, R., Chiang, C.-y., Tu, B.P., Yan, S.F., DeJesus, P.D., Romero, A., Bergauer, T., Orth, A., Krueger, U., Zhou, Y., et al. (2007). A probability-based approach for the analysis of large-scale RNAi screens. *Nature Methods*, 4(10): 847–849.
11. Łukasiak, P., Błażewicz, J. & Miłostan, M. (2010). Some operations research methods for analyzing protein sequences and structures. *Annals of Operations Research*, 175(1): 9–35.
12. Mohr, S., Bakal, C. & Perrimon, N. (2010). Genomic screening with RNAi: results and challenges. *Annual Review of Biochemistry*, 79: 37.
13. Piau, D. (2004). Immortal branching Markov processes: averaging properties and PCR applications. *Annals of Probability*, 32: 337.
14. Serfling, R.J. (1981). *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.
15. Strickland, D.M., Barnes, E. & Sokol, J.S. (2005). Optimal protein structure alignment using maximum cliques. *Operations Research*, 53(3): 389–402.
16. van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Series in Statistics. Springer.

https://doi.org/10.1017/S0269964814000254 Published online by Cambridge University Press