Optimally Pooled Viral Testing

Dor Ben-Amotz*

Purdue University, Department of Chemistry, West Lafayette, IN 47907

E-mail: bendor@purdue.edu

Abstract

It has long been known that pooling samples may be used to minimize the total number of tests required in order to identify each infected individual in a population. Pooling is most advantageous in populations with low infection probability, but is expected to remain better than non-pooled testing in populations with an infection probability up to 30%. Additional testing efficiency may be realized by performing a second round of pooled testing, thus reducing the average number of tests required to uniquely identify each infected individual in a population with 1% infection from 20 to 14 out of 100, and from 6 to 4 when the infection probability is 0.1%. These best case predictions, obtained assuming perfect test accuracy and specificity, provide a quantitative measure of the optimal pool size and expected testing efficiency gains in populations with infection probabilities ranging from 0.1% to 30%, and are supported by recent COVID-19 empirical detection sensitivity and optimized pool size studies. Although large pools are most advantageous for testing populations with very low infection probabilities, they are predicted to become highly non-optimal with increasing infection probability, while pool sizes smaller than 10 remain near-optimal over a broader range of infection probabilities.

Introduction

The advantages of pooled testing in applications ranging from disease screening to manufacturing quality assurance have long been appreciated.[1] Efficiently and practically containing viral outbreaks requires minimizing the total number of tests needed in order to uniquely identify every positive individual. This may be achieved using pooled testing, given a prior estimate of the probability of infection as well as the availability of a sufficiently sensitive diagnostic test with an acceptably low false-negative detection probability. When applicable, pooled testing a population consisting of a large number N of individuals can be achieved with significantly fewer than N tests, by initially screening pools containing a mixture of n

samples, followed by further testing of only the positive pools. The optimal pool size n is expected to increase with decreasing infection probability p, and is no longer expected to be advantageous when p>0.3 (30%). The present predictions are obtained assuming that there is no correlation between the infections within a pool and that all pools have approximately the same infection probability. Additionally, the predictions are obtained assuming perfect test accuracy and specificity, and thus represent a best case scenario. However, the predictions are in reasonable agreement with available COVID-19 pooled testing results,[2,3] as well as recent single round pooled testing simulations that include the effects of imperfect test sensitivity.[4] Thus, the present results are expected to accurately approximate the testing efficiency gains obtainable using actual COVID-19 tests performed using either one or two rounds of pooling, as well as to aid in the selection of a fixed pool size that remains near optimal over the range of infection probabilities in a given population.

The optimal value of n, as well as the expected number of infections per pool, may be determined using the binomial distribution,[5] whose optimal n predictions are equivalent to those first obtained by Dorfman.[1] Here these results are extended to include predictions of the range of infection percentages over which a give fixed pool size remains nearly optimal, as well as the expected number of infected individuals per pool as a function of p. Moreover, the significant additional efficiency that may be obtainable from a second round of pooling is determined. These predictions imply that optimal pooling may be used in populations with a very low infection probability of 0.1%, using an optimal pool size of 32. The practicality of using pools this large has recently been demonstrated by Prof. Idan Yelin and co-workers at Technion and Rambam Health Care Campus in Haifa, Israel, who showed that a standard RT-qPCR test for COVID-19 may be used to detect a single positive individual in pools as large as 32, with an estimated false negative rate of 10%.[2] However, it is also important to note that large pool testing is only predicted to be beneficial for populations with a very low and narrow range of infection percentages, and becomes highly non-optimal for populations with infection percentages exceeding 1%.

The practical relevance of the present results is further supported by the recent finding of Dr. Manoj Jain of the Baptist Memorial Hospital and Methodist Hospitals in Memphis, using COVID-19, who found that hundreds of test samples obtained from firefighters, police officers and city workers could be most efficiently tested using a pool size of 7,[3] which is consistent with the present predictions pertaining to an infection percentage of 2.4%. In practice, these predictions may be used by initially employing a pool size obtained by roughly estimating the expected infection probability in the population of interest, and subsequently adjusting the pool size to better match the actual infection probability. These predictions are expected to be most useful in facilitating large scale screening and continuous testing of populations with low infection probabilities for early detection of COVID-19 outbreaks, to enhance both public safety and economic productivity.

Results

The binomial distribution yields the following expression for the probability that there will be k infected individuals in a pool of sized n, drawn from a population with an infection probability of p.[5]

$$P(k) = \frac{n! p^k (1-p)^{n-k}}{k!(n-k)!} \tag{1}$$

When $k = 0$ this reduces to the following expression for the fraction of pools that are expected to contain no infected individuals, in keeping with Dorfman's original predictions.[1]

$$P(0) = (1-p)^n \tag{2}$$

This yields the following expression for the total number of tests N_{tests} required in order to exhaustively test a population of size N, with an infection probability of p, using a pool size of n.

$$N_{tests} = \frac{N}{n} + N\left[1 - (1-p)^n\right] \tag{3}$$

Thus, the average percentage of tests that must be performed in order to identify every infected individual in a population of size N is $T\% = 100 \times (N_{tests}/N)$. In other words, $T\%$ represents the average number of tests required to identify each infected individual in a population of size 100, or equivalently $T\%\times$1,000 is the number of tests required to do so in a population of 100,000.

$$T\% = 100 \left[1 + \frac{1}{n} - (1-p)^n \right] \tag{4}$$

The optimal value of n is that which minimizes $T\%$, and thus may be obtained by finding the roots of the following expression for the partial derivative of $T\%$ with respect to n, pertaining to a given value of p.

$$-\frac{1}{100} \left(\frac{\partial T\%}{\partial n} \right)_p = \frac{1}{n^2} + (1-p)^n \ln(1-p) = 0 \tag{5}$$

The above expression may be solved numerically using Newton's method. Alternatively, the optimal pool size may also be obtained iteratively, using an initial guess for the pool size n_0, inserted into the right-hand-side of the following expression, to obtain a better estimate of n (where the "Round" operation rounds the result to the nearest positive integer).

$$n \approx \text{Round} \left\{ \left[\ln \left(\frac{1}{1-p} \right) (1-p)^{n_0} \right]^{-1/2} \right\} \tag{6}$$

If n_0 is not very similar to n, then one may set $n_0 = n$ and repeat the process to obtain a better estimate of n. This iterative procedure typically converges within a few cycles (whose convergence can be most accurately quantified by removing the Round operation from the right-hand-side of Eq. 6). The resulting optimal values of n are given in the 3rd column in Table 1, whose first two columns represent the average infection probability p and the corresponding infection percentage $100p$ in the population of interest. The 4th column in Table 1 contains the resulting $T\%$ predictions, obtained when using a single round of pooled testing. Note that these values correspond to averages over large populations. For example,

for a population with an infection probability of $p = 0.001$ (0.1%), the optimal pool size is 32 and the vast majority of such pools will contain no infected individuals. More specifically, any such pool of size 32 is predicted to have no infections 97% of the time, and the remaining 3% of the pools of size 32 are predicted to contain only one infected individual.

Figure 1 contains more detailed predictions pertaining to the average number of infected individuals in pools of optimal size, when the overall infection probability ranges from 1% to 30%. Note that at (and below) an infection probability of 1%, essentially all of the positive pools are predicted to contain only one infected individual. At higher infection probabilities a non-negligible number of positive pools are predicted to contain more than one infected individual, but nevertheless most positive pools are predicted to contain only one infected individual. For example, even in a population with an infection probability of 30%, about 34% of the pools of size 3 are predicted to contain no infected individuals, while 45% contain one, and only 21% contain more than one infected individual. However, at this high rate of infection there is no longer any significant advantage to pool testing, relative to exhaustively testing every single individual, as indicated by the 4th column in Table 1, which indicates that an average of 99 tests would have to be performed when optimally pool testing a population of 100 individuals that has an infection percentage of 30%.

Figures 2 and 3 contain graphical predictions pertaining to tests performed using either one or two rounds of optimal pooling, respectively. Figure 2 shows the resulting optimal first round pool size n (**a**) and testing percentage T% (**b**) predictions. The insert panels in each figure contain an expanded view of the predictions pertaining to populations with infection percentages less than 1%, and the solid curves are optimal pooled testing predictions. The optimal pool size values shown in Table 1 are obtained by rounding the graphical results to the nearest positive integer. The dotted curves in Figure 2**b** show the testing efficiency predictions obtained when using various fixed pool size estimates n_0, indicating that near-optimal pool sizes produce results that are essentially the same as those obtained using an optimal pool size, as long as the actual infection percentage is not too far from its

Table 1: Optimized Pool Testing Results

Infection probability (p)	Infection percent	Pool Size 1st round (n)	Tests Needed 1 round (T%)	Pool Size 2nd round (n_2)	Tests Needed 2 rounds (T%)
0.001	0.1%	32	6	8	4
0.002	0.2%	23	9	7	6
0.003	0.3%	19	11	7	7
0.004	0.4%	16	12	6	8
0.005	0.5%	15	14	6	9
0.006	0.6%	13	15	6	11
0.007	0.7%	12	16	5	12
0.008	0.8%	12	18	5	12
0.009	0.9%	11	19	5	13
0.01	1%	11	20	5	14
0.02	2%	8	27	4	20
0.03	3%	6	33	4	26
0.04	4%	6	38	4	30
0.05	5%	5	43	4	35
0.06	6%	5	47	3	38
0.07	7%	4	50	3	43
0.08	8%	4	53	3	45
0.09	9%	4	56	3	48
0.1	10%	4	59	3	51
0.15	15%	3	72	3	66
0.2	20%	3	82	3	77
0.25	25%	3	91	3	88
0.3	30%	3	99	3	99

estimated value. More specifically, these predictions indicate that pool sizes of 5, 6, and 7 are expected to produce nearly optimal testing efficiency in populations with average infection percentages ranging from 2% to 12%, 1% to 8%, and 0.7% to 6%, respectively (as determined by requiring that T% remain within 3% of its optimal value). Larger pool sizes are predicted to remain near optimal over a narrower range of infection probabilities, and to rapidly become significantly non-optimal with increasing infection probability, as exemplified by the dotted curves in Figure 2**b**.

Figure 3, as well as the last two columns in Table 1, contain predictions obtained if two rounds of optimal pooling are performed on the same population of test samples. Specifically, the 5th column in Table 1 indicates the optimal second round pool size and the last

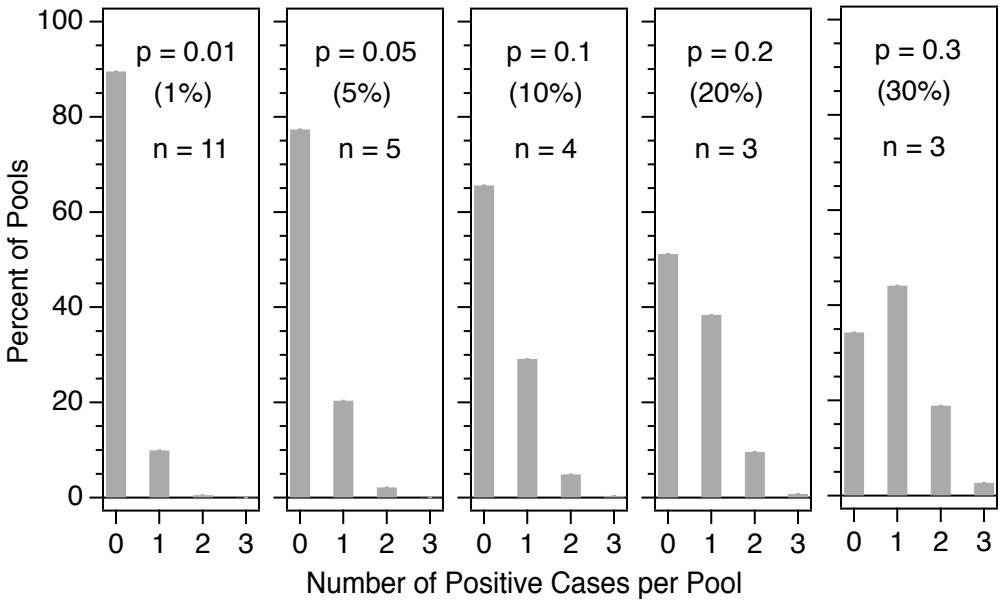

Figure 1: Predicted number of infected individuals in optimally sized pools obtained from populations with average infection percentages ranging from 1% to 30%.

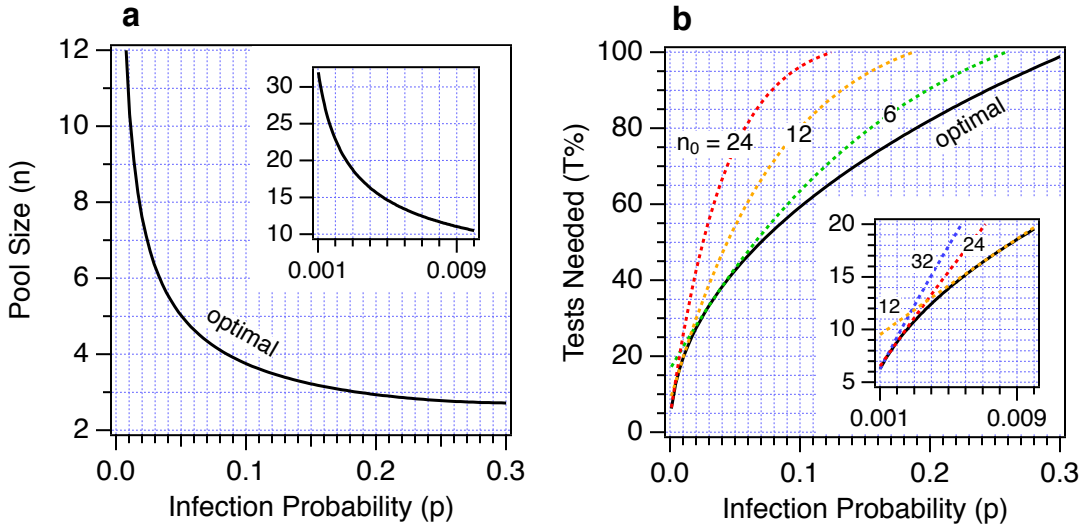

Figure 2: Optimal pool size **a** and testing percentage **b** predictions obtained when applying a single round of pooled testing. The dashed lines in **b** represent the testing percentages obtained using three different fixed pool sizes.

column indicates the predicted average number of tests required to determine all the infected individuals in a population of size 100 when using two rounds of optimal pooling. The second round of optimal pooling is performed by limiting the second round tests to individuals in

the positive first round pools, as further described below.

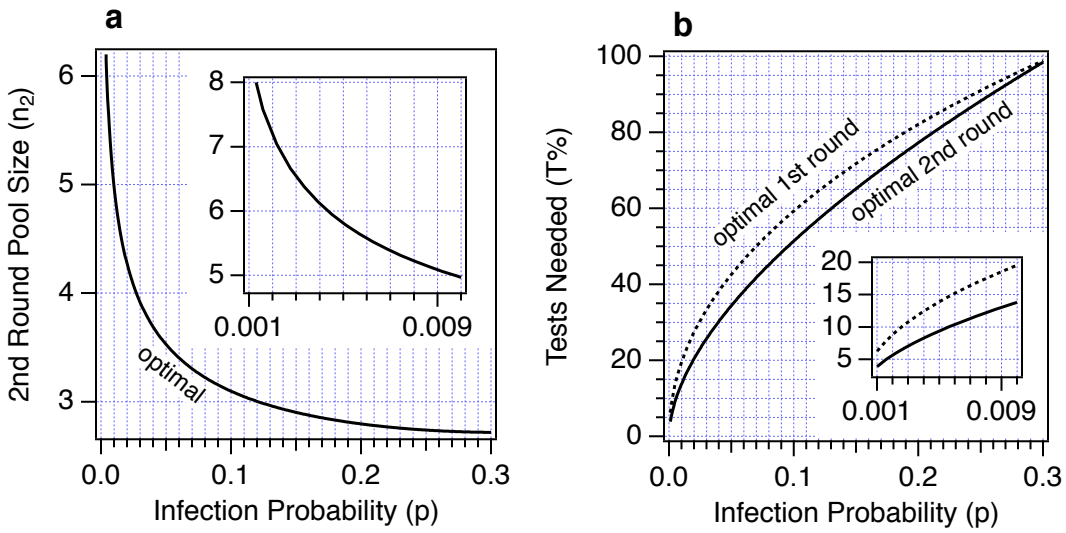

Figure 3: Predicted optimal pool sizes **a** and testing percentages **b** obtained when applying two rounds of pooled testing. The dotted and solid curves in **b** compare the predicted testing efficiencies obtainable using one or two rounds of optimized pool testing, respectively.

The solid curves in Figure 3 are optimal second round pool testing predictions. Figure 3**a** shows the predicted optimal pool size, n_2, that should be used in order to efficiently re-test all the samples from the positive first round pools. More specifically, the average infection probability p_2 in all the positive first round pools is higher than that in the original population because all the non-infected individuals in the negative first round pools have been removed from the population of second round test samples. Thus, the optimal second round pool size n_2 is obtained as follows, where $T\%(p, n)$ is the first round testing percentage (obtained using Eq. 4) and $n(p_2)$ is the optimal pool size pertaining to an infection percentage of p_2.

$$p_2 = \frac{100\,p}{T\%(p, n)} \tag{7}$$

$$n_2 = n(p_2) \tag{8}$$

Thus, the following equation predicts the total number of tests required to identify every

infected individual when using two rounds of pooling.

$$T\%(\text{2 round total}) = \frac{100}{n} + \left[1 - (1-p)^n\right] T\%(p_2, n_2) \tag{9}$$

Note that $100/n$ is the number of pools that were tested in the first round (expressed as a percent of total number of tested individuals N), and $1-(1-p)^n$ is the fraction of positive first round pools, and thus $\left[1 - (1-p)^n\right] T\%(p_2, n_2)$ is the number of tests required to identify all the positive individuals in those pools, where $T\%(p_2, n_2)$ is again obtained using Eq. 4.

The dotted and solid curves in Figure 3**b**, as well as the 4th and 6th columns in Table 1, compare the first and second round optimal testing percentage predictions. These results indicate that there is a significant advantage to performing two rounds of pooled testing for populations with infection probabilities less than 30%. For example, in a population with an infection percentage of 1%, one round of pooling is predicted to require an average of 20 tests per 100 individuals, while two rounds of pooling reduces that to 14 tests per 100 individuals. The fractional gain in testing efficiency increases as the infection percentage decreases, and decreases from 6 to 4 tests per 100 individuals in a population with an infection percentage of 0.1%. Again, note that these predictions represent the average number of tests required per 100 individuals, and the value of 4 arises because approximately 97% of the first round pools of size 32 drawn from such a population are predicted to contain no infected individuals.

Summary and Discussion

Optimal pooled testing is expected to be useful in improving the efficiency of COVID-19 diagnostics in populations with infection percentages below 30%, and becomes more advantageous with decreasing infection probability, as long as the sensitivity of each test is sufficient to detect one infected individual diluted in a pool of optimal size. In a populations with an infection probability of 0.1% the predited optimal pool size is 32, which is consistent with recently reported COVID-19 testing sensitivities achievable using a standard RT-qPCR

test.[2] At lower infection percentages, pool sizes of 32 may continue to be used, although larger pool sizes would become optimal if the testing sensitivity were sufficiently high.[2] It is also important to note that such large pool sizes are only optimal for use in populations with very low infection probabilities, and rapidly become non-optimal with increasing infection probability. On the other hand, smaller pool sizes tend to remain nearly optimal over a larger range of infection probabilities (as illustrated by the dotted curves in Figure 2**b**).

Optimal pooled testing is expected to be most advantageous when applied to asymptomatic or randomly sampled individuals, as symptomatic individuals are likely to have an infection probability near or exceeding 30%. Thus, pooled testing is expected to be most useful for detection of outbreaks in a relatively stable population, so as to prevent a runaway growth of viral infections. However, controlling such outbreaks requires not only optimal pooled testing but also factors, including effective isolation and contact tracing.

Two rounds of pooled testing are expected to be most advantageous in continuous testing of a population with a low infection percentage. For example, in a population of 100,000 with an average infection probability of 0.1% it is predicted that every infected individual could be identified by performing as few as 4,000 tests, or as few as 14,000 tests in a population of the same size with an infection probability of 1%. This relatively low testing load should make it practical to repeatedly test a population in order to identify early warnings of an emerging outbreak. Although the present predictions were obtained assuming that tests are perfectly accurate and specific, the results are also expected to be of relevance to more realistic situations as illustrated, for example, by the fact that a recent study determined that a pool size of 7 was approximately optimal in testing a population of hundreds of volunteers in Memphis, TN,[3] which is consistent with the predicted optimal pool size for a population with an infection percentage of 2.4%. The practical utility of the present results is further supported by the prediction that the actual pool size need not be precisely optimal in order to obtain near-optimal testing efficiency, particularly for infection probabilities that range from 1-10%, which may be nearly optimally tested using pool sizes within the correspond-

ing optimal pools size range of 4 to 11. Moreover, the accuracy of tests performed using non-optimal pool sizes is expected to be limited primarily by the false negative detection percentage pertaining the chosen pool size, thus favoring the use of smaller rather than larger pools in situations where it is not practical to employ an optimal pool size.

Acknowledgement

This work was supported by the National Science Foundation (Grant Number CHE-109746).

References

(1) Dorfman, R. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics* **1943**, *14*, 436–440.

(2) Yelin, I.; Aharony, N.; Shaer-Tamar, E.; Argoetti, A.; Messer, E.; Berenbaum, D.; Shafran, E.; Kuzli, A.; Gandali, N.; Hashimshony, T. Evaluation of COVID-19 RT-qPCR Test in Multi-Sample Pools. *Clin. Infect. Dis.* **2020**, *DOI:10.1093/cid/ciaa531*, Published online ahead of print, May 2.

(3) Mandavilli, A. Federal Officials Turn to a New Testing Strategy as Infections Surge. *New York Times* **2020**, July 1st.

(4) Cherif, A.; Grobe, N.; Wang, X.; Kotanko, P. Simulation of Pool Testing to Identify Patients With Coronavirus Disease 2019 Under Conditions of Limited Test Availability. *JAMA network open* **2020**, *5*, e2013075.

(5) Wilcox, D. S.; Rankin, B. M.; Ben-Amotz, D. Distinguishing Aggregation from Random Mixing in Aqueous t-Butyl Alcohol Solutions. *Faraday Disc.* **2013**, *167*, 177–190.