medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

A strategy for finding people infected with SARS-CoV-2: optimizing pooled testing at low prevalence

Leon Mutesa,[1,2] Pacifique Ndishimye,[3,2] Yvan Butera,[2] Annette Uwineza,[2] Robert Rutayisire,[2] Emile Musoni,[2] Nadine Rujeni,[2] Thierry Nyatanyi,[2] Edouard Ntagwabira,[2] Muhammed Semakula,[2] Clarisse Musanabaganwa,[2] Daniel Nyamwasa,[2] Maurice Ndashimye,[3,2] Eva Ujeneza,[3] Ivan Emile Mwikarago,[2] Claude Mambo Muvunyi,[2] Jean Baptiste Mazarati,[2] Sabin Nsanzimana,[2] Neil Turok,[3,4,*] and Wilfred Ndifon[3,†]

[1] *Centre for Human Genetics, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda*
[2] *Rwanda Joint Task Force COVID-19, Kigali, Rwanda*
[3] *African Institute for Mathematical Sciences, Kigali, Rwanda*
[4] *Perimeter Institute for Theoretical Physics, 31 Caroline St N, Ontario, Canada*

Suppressing SARS-CoV-2, the causative agent of COVID-19, requires that infected individuals are rapidly identified and isolated, on an ongoing basis. However, regular testing of every individual is expensive. We present an approach to pooled testing, based on the geometry of a hypercube, which enables fast, cost-effective identification of infected people when the prevalence of the virus is low. Field trials of this approach are now under way in Rwanda.

* nturok@perimeterinstitute.ca
† wndifon@nexteinstein.org

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

SARS-CoV-2 represents a major threat to global health. Rapidly identifying and quarantining infected individuals is one of the most important strategies available to contain the virus. However, each diagnostic SARS-CoV-2 test costs 30-50 US dollars [1]. Therefore, testing every individual in a population regularly, as may be essential to eliminating the virus, is very expensive. The costs are unaffordable for most low-income countries with limited resources for massive SARS-CoV-2 testing. It is therefore important to ask: are there more efficient ways to find infected people?

The first step in testing, swab collection, is labour intensive but does not require expensive chemicals or equipment. It may therefore be feasible to collect swabs regularly from everyone. The next step involves RT-PCR machines [2]. These require expensive chemical reagents, currently in short supply, as well as skilled personnel. Reducing the cost requires that we minimize the number of tests. Testing rapidly is also vital because SARS-CoV-2 is so infectious. Any time wasted during testing results in a higher prevalence of the virus, which spreads quickly [3].

To find infected individuals, the naive approach is to test everyone. For a group of size N, this takes N tests. However, far fewer tests are actually needed, especially at low prevalence. It is much more efficient to pool (or combine) samples and test them together. Pooling samples has already been discussed as a way to efficiently estimate the prevalence of the virus [4, 5]. Our focus here is, instead, on how to use pooling of subsamples to identify infected individuals.

The idea of group testing dates back to a paper of Dorfman in 1943 [6]. Dorfman's algorithm reduces the number of tests per person, required to find all infected individuals, to $\approx 2\sqrt{p}$ at low prevalence p (see Appendix A). We shall describe an algorithm which requires only $\approx ep\ln(1/p)$ tests per person at low p, substantially improving on Dorfman's. Our algorithm is largely parallel, so that finding an infected individual usually happens in a single step. There are algorithms which require fewer tests, such as binary searches. However, they are adaptive and hence serial. As the prevalence falls, the optimal group size N grows, and adaptive searches take longer and longer. During this time, the disease continues to spread. We detail below how, at low prevalence (see, *e.g.*, Ref. [7]), this effect disfavours adaptive searches. Among parallel algorithms, ours scales similarly to the best ones known in the literature (see Appendix B).

Group testing is most obviously effective when there are no infected individuals at all. When the samples from a group are pooled and tested together, just one test suffices to show that no-one is infected. Our algorithm takes full advantage of this powerful result. As we shall see, the optimal group size N is chosen so that the first test, conducted on the whole group, is usually negative.

Now consider the case where only one individual is infected. The idea behind our algorithm is geometrical: we pool subsamples, prior to testing, in the pattern suggested by a high-dimensional hypercube. The group of individuals to be tested is represented by a set of N points on a cubic lattice in D dimensions, organized in the form of a hypercube of side L, with

$$L^D = N. \tag{1}$$

Instead of directly testing the sample taken from each individual, we first divide it into D equal subsamples. These DN subsamples are recombined as follows. Slice the hypercube into L planar slices, perpendicular to one of the principal directions on the lattice. Form such a set of slices in each of the D principal directions. Altogether, there are DL slices, each consisting of $N/L = L^{D-1}$ points. We combine the L^{D-1} subsamples corresponding to each slice. If there is just one infected individual, then one slice out of the L slices, taken in each of the D directions, will yield a positive result. That slice directly indicates the coordinate of the point corresponding to the infected individual, along the associated principal direction.

Therefore the number of tests required to uniquely identify the infected individual is

$$DL = DN^{1/D}, \tag{2}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

where we used (1). Treating D as a continuous variable, the right hand side of (2) diverges at both small and large D, possessing a minimum at

$$D = \ln N, \tag{3}$$

corresponding to $L = e$ and a total of $e \ln N$ trials. In reality, D and L must be integers, but using $L = 3$ achieves almost the same efficiency. With no further constraint, finding one infected person in a group of 10^6 individuals would require only 38 tests, performed in just one round of testing.

In practice, we are limited by the capacity of the testing machine. A typical swab yields 10^5 viral RNA molecules/ml [8]. For each slice of the hypercube, we combine N/L subsamples of the virus, each of volume v. If the volume of each combined sample, $V = Nv/L$, exceeds the capacity of the PCR we will have to only use a portion of it. We should also keep in mind that at least 1 viral RNA is needed for an unambiguous result, and we must remain well above this limit.

Setting $L = 3$ and $N = 3^D$, we find

$$D = \frac{\ln(V/v)}{\ln 3} + 1 \tag{4}$$

For example, if $V/v = 100$ then (4) yields $D \approx 5$, from which (3) yields $N = 243$. If v is a microliter, then V is 100 microliters. In a positive combined sample, there would be 100 viral RNA molecules. Even if only 10 microliters are used in the PCR machine, it would contain 10 viral RNA molecules, sufficient for a positive result. The typical number of tests required to find the infected individual is then only $3 \ln N \approx 17$, an efficiency gain of 14.

As a proof of concept, using oropharyngeal swab specimens collected during COVID-19 surveillance in Rwanda, we have shown that positive specimens can still be detected even after they are diluted by up to 100 fold (i.e. $V/v = 100$; Figure 1). Sample pooling is now used for cost-effective, large scale testing in Rwanda to understand the spatial spread of SARS-Cov-2 nationally and identify new infection hotspots to enable a rapid response by public health officials.

Note that the viral load found in a swab specimen is relatively low if collected during the early stages of viral replication [8]. Therefore, swabs taken during this period may contain insufficient virus to yield a positive result. The sensitivity of the test is typically increased by testing specimens collected at sequential time points. Methods like the one we describe here facilitate such sequential testing on a massive scale by drastically reducing the associated costs. In view of the large potential efficiency gains, it is worth exploring whether testing machines could be engineered to accommodate larger test volumes.

We have so far assumed there is just one infected individual in the group to be tested. However, we cannot know this ahead of time. What if there are 2, 3 or more infected people present? The number of infected individuals present in a group of size N, at prevalence p, should be described by Poisson statistics with mean $\lambda = pN$. For λ significantly below unity, the probability to find m infected individuals decreases rapidly with m. As we shall see, this is the regime in which our algorithm is most efficient.

We summarize the algorithm as follows:

(i) For a given initial estimate of the prevalence p, select the group size N to be the optimal power of 3, discussed below. Test the whole sample to see whether one or more infected individuals are present. A negative test indicates everyone in the group is clear of infection.

(ii) If the test is positive, run one round of testing according to the hypercube algorithm with $L = 3$. The distribution of results for the LD slices should, for large N, accurately indicate the number m of infected individuals, with many consistency checks including a new estimate of the prevalence. If m=1, the infected individual is immediately identified.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

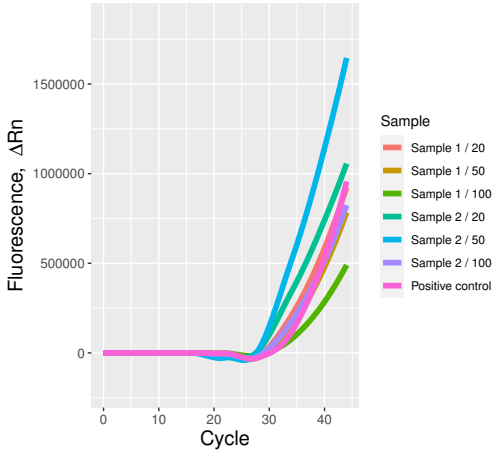

Figure 1: **Positive samples are detected after a 100-fold dilution.**
Shown are representative RT-PCR fluorescence curves obtained by amplifying SARS-COV-2 E-gene in two positive samples diluted in ratios of 1:20, 1:50, and 1:100 using equal volumes of 19, 49, and 99 negative samples, respectively. For comparison, a fluorescence curve from a positive control is also shown.

(iii) If $m = 2$, select a principal direction in which two slices are positive. Take one of these, itself a hypercube of dimension $D - 1$, and run the hypercube algorithm again. The coordinates of the corresponding infected individual are then uniquely identified, and those of the second infected individual are inferred by elimination.

(iv) If $m > 2$, select a principal direction in which all three slices are positive. Taking one of these, run the hypercube algorithm again. If that slice contains one infected individual, it is immediately identified. If the slice contains 2 or 3 infected individuals, their coordinates are discovered in 1 or 2 additional hypercube runs. The last infected individual is found by elimination. So, for $m = 1, 2$ or 3, only m runs are needed.

(v) For $m > 3$, it is an exercise in combinatorics to work out the expected number of runs. This turns out to be a number slightly higher than m. The corrections to m for $m = 4, 5, 6$ are given in Appendix B. For reasons given there, these corrections turn out to be relatively unimportant.

In Appendix B we show that, for given prevalence p, the number of tests per person is minimized for a group size $N \approx 0.350/p$, and that its minimal value is $\approx ep\ln(0.734/p)$. For prevalences p in the ranges $\{(2.3, 0.78), (0.78, 0.26), (0.26, 0.09), (0.09, 0.03)\}$ per cent, the optimal group sizes N are $\{27, 81, 243, 729\}$ and the efficiency gains over naive testing are in the respective ranges$\{4, 10), (10, 24), (24, 62), (62, 160)\}$.

We now compare our search algorithm with those already considered, for diverse purposes, in computer science or mathematical statistics. Information theory sets a lower bound on the number of tests per person required to identify an infected individual (see Appendix B). At low p, this bound is $\approx p\log_2(1/p)$. Binary searches approach this limit by performing an iterated series of tests (see, *e.g*, Refs. [9, 10]). These algorithms require fewer tests than ours by a factor of $e\ln 2 \approx 1.88$. However, when considering a rapidly spreading infectious disease like COVID-19, saving time is important because infected individuals who are still at large can infect others. A parallel, or largely parallel, search such as ours reduces this risk. In contrast, binary searches are adaptive and must be performed serially. At low prevalence, the optimal group size N scales as $1/p$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(see Appendix A). A binary search takes $\log_2 N \sim \log_2(1/p)$ steps to find an infected individual [9]. Each PCR test takes several hours, to which must be added any time taken for subsample sorting and selection. So, multiple of rounds of testing would consume significant time. During this period, the prevalence p of the virus grows exponentially. The doubling time for the virus is somewhat uncertain but it has been recently estimated, using data from China [3], to be $\tau_2 \approx 2.4$ days. If each PCR test takes τ days, the viral prevalence grows by $\sim (1/p)^{\tau/\tau_2}$ during an adaptive search. It follows that, at fixed τ/τ_2, the adaptive algorithms do worse at low p. For example, if we assume that at most 3 rounds of adaptive PCR tests may be performed in a working day, then $\tau = 1/3$ day. For prevalences p below $(e \ln 2)^{-\tau_2/\tau} \approx 1$ per cent, a parallel approach like ours is then preferred. Reducing the costs of staff and lab time also favours a largely parallel strategy.

Another key consideration is robustness to error. A search method such as ours allows for many consistency checks which will help to eliminate false positives or negatives. In contrast, binary searches rely on repeated testing of the positive sample, making errors harder to identify.

The most striking consequence of our approach is that the cost of testing the whole population falls rapidly as the prevalence declines. This should incentivize decision-makers to act firmly in the early phases of the pandemic. Although driving the prevalence down initially is costly, maintaining a low prevalence thereafter and, indeed, eliminating COVID-19 altogether, will become progressively more affordable.

Acknowledgments: We thank the Rwanda Ministry of Health through RBC for stimulating discussions and correspondence. Kendrick Smith and Corinne Squire provided valuable encouragement and helpful references. Research at AIMS is supported in part by the Carnegie Corporation of New York and by the Government of Canada through the International Development Research Centre and Global Affairs Canada. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities.

---

[1] Medicare, "Medicare administrative contractor (MAC) COVID-19 test pricing,". `https://www.cms.gov/files/document/mac-covid-19-test-pricing.pdf`.

[2] V.M. Corman, O. Landt, M. Kaiser *et al.* , "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Euro. Surveill.* **25** (2020) 2000045.

[3] S. Sanche, Y.T. Lin, C. Xu, E. Romero-Severson, M. Hengartner, R. Ke, "High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus," *Emerg. Infect. Dis.* **26** (July, 2020) . `https://doi.org/10.3201/eid2607.200282`.

[4] Stanford Medicine Center, "Testing pooled samples for COVID-19 helps Stanford researchers track early viral spread in Bay Area,". `http://med.stanford.edu/news/all-news/2020/04/testing-pooled-samples-to-track-early-spread-of-virus.html`.

[5] R. Hanel, S. Thurner, "Boosting test-efficiency by pooled testing strategies for SARS-CoV-2,". `arXiv:2003.09944[q-bio.PE]`.

[6] R. Dorfman, "The Detection of Defective Members of Large Populations," *Ann. Math. Stat.* **14** (1943) 436.

[7] D.F. Gudbjartsson *et. al.*, "Spread of SARS-CoV-2 in the Icelandic Population," *New Eng. J. Med.* **DOI: 10.1056/NEJMoa2006100** (April 14, 2020) . `https://www.nejm.org/doi/full/10.1056/NEJMoa2006100`.

[8] R. Wölfel, V.M. Corman, W. Guggemos *et al.* , "Virological assessment of hospitalized patients with COVID-2019," *Nature* . `https://doi.org/10.1038/s41586-020-2196-x`.

[9] F.K. Hwang, "A method for detecting all defective members in a population by group testing," *J. Amer. Stat. Assoc.* **67** (1972) 605.

[10] A. Allemann, "An efficient algorithm for combinatorial group testing," in Aydinian H., Cicalese F., Deppe C. (eds), *Information Theory, Combinatorics, and Search Theory,* Lecture Notes in Computer

medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Science, **7777**, Springer, 2013 .

[11] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: an information theory perspective," `arXiv:1902.06002`. `http://arxiv.org/abs/1902.06002`.

Appendix A: Comparison with Dorfman

In a landmark paper in 1943, R. Dorfman considered the problem of searching for infected individuals by grouping (or pooling) samples. Consider a population of N individuals, broken up into k groups of N/k members each. If the probability that any individual is infected (the prevalence) is p, the probability that a group is free of infection is $(1-p)^{N/k}$. Conversely, the probability that at least one member is infected is $p' = 1-(1-p)^{N/k}$. Dorfman's strategy was to test all groups, and then to test every member of every infected group. The expected number of tests is then

$$\langle T \rangle = k + p'N, \tag{A1}$$

and the number of tests required per person is

$$\frac{\langle T \rangle}{N} = \frac{k}{N} + p' = \frac{k}{N} + 1 - (1-p)^{N/k} \approx \frac{k}{N} + p\frac{N}{k}. \tag{A2}$$

In the last step, we assumed the prevalence p to be small. The number of tests per person is minimised for $k = \sqrt{p}N$, and the optimal group size is $1/\sqrt{p}$. The expected number of tests per person, at the optimal group size, is approximately $2\sqrt{p}$.

Let us compare these results with those obtained using our hypercube algorithm. Assuming Poisson statistics, the expected number of tests per person is given by

$$\frac{\langle T \rangle}{N} = \frac{e^{-\lambda}}{N}\left(1 + e\ln N \sum_{m=1}^{\infty} \frac{\lambda^m}{m!} r_m\right) \approx \frac{e^{-pN}}{N} + pe\ln N \tag{A3}$$

where r_m denotes the number of runs of our hypercube algorithm needed to identify all infected individuals. Each run consists of $e\ln N$ tests. We shall ignore any changes in the argument of the logarithm due to fewer than N tests sometimes being needed. These changes are unimportant at large N. For $m = 1, 2$ or 3, $r_m = m$. For larger m, more than m runs may be needed. Setting $r_m = m + c_m$ for $m > 3$, the c_m denote the average excess number of runs. They are tedious but straightforward to compute. The first few are: $c_4 = \frac{1}{3}$, $c_5 = \frac{16}{33}$, $c_6 = \frac{37}{54}$. In any case, they turn out to be unimportant, because the minimum of (A3) is dominated by contributions from low m.

The last expression in (A3) represents the approximation that the c_m corrections are neglected. The first term diverges at small N, and the second diverges at large N. Thus, a minimum exists. It is located at $\lambda = pN \approx 0.350$, independent of p. For such a small value of λ, contributions from higher m are strongly suppressed. One subtlety is that, at very low p, since $N = \lambda/p$, for the relevant values of λ the argument of the logarithm becomes very large, so its derivative with respect to λ becomes suppressed with respect to its value. This effect means that, for extremely small p, the derivative of the corrections can compete with the derivative of the logarithm. We have checked that only for $p < 10^{-10}$ does this alter the minimum value of λ by more than 0.01. Hence, for all practical purposes, we can safely ignore the c_m corrections.

The optimal group size and the corresponding minimal number of tests per person are thus given, to an excellent approximation, by

$$N \approx \frac{0.350}{p}; \qquad \frac{\langle T \rangle}{N} \approx ep\ln\left(\frac{0.734}{p}\right). \tag{A4}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.02.20087924; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

For a prevalence p of 1 per cent, Dorfman's approach yields a group size of approximately 10 and approximately 0.2 tests per person, whereas ours yields a group size of 35 and 0.12 tests per person. For a prevalence of 0.1 per cent, Dorfman's optimal group size is 32 whereas ours is 350. His approach requires 0.06 tests per person, whereas ours needs only 0.018. For all lower, but still reasonable, values of p our approach prevails by an increasing margin.

Appendix B: Information theory bound

Information theory sets a lower bound on the number of tests required to uniquely identify all infected individuals. The uncertainty in who is infected is associated with an entropy,

$$S = -\sum_i p_i \ln p_i, \tag{B1}$$

where the sum is over all possible states and the p_i are the corresponding probabilities. If a test outputs a zero or a one then for t tests the number of possible test outputs is 2^t and the corresponding information gained is at most $t \ln 2$. In order to learn everything about the system, one requires an information gain of at least S, hence

$$t > \frac{S}{\ln 2}. \tag{B2}$$

Consider a sample of size n, with k infected individuals chosen at random. The number of such states is $\binom{n}{k}$. Therefore, from (B2), the minimum number of tests required is $\log_2 \binom{n}{k} \sim k \log_2(n/k)$ for $k \ll n$. Assuming a binomial distribution with prevalence p, and replacing k with its expectation $p\,n$, we find the expected number of tests per person is $\sim p \log_2(1/p)$. Binary searches can approach this limit by performing an adaptive series of tests (see, *e.g*, Refs. [9, 10]). However, the number of tests in each search scales as $\log_2 N \sim \log_2(1/p)$ for the optimal value of N, at low p. In contrast, the most efficient known parallel searches, called "noiseless, nonadaptive" tests in Ref. [11], require a factor of $e \ln 2$ more tests (see their equations (2.8) and (2.10)), just as our algorithm does.