Group testing as a strategy for COVID-19 epidemiological monitoring and community surveillance

Vincent Brault, Bastien Mallein, Jean-François Rupprecht

Group testing as a strategy for COVID-19 epidemiological monitoring and community surveillance

Vincent Brault [1,4❍*], Bastien Mallein [2,4❍*], Jean-François Rupprecht [3,4❍*]

1 Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France,
2 Université Sorbonne Paris Nord, LAGA, UMR 7539, F-93430, Villetaneuse, France.
3 Aix Marseille Univ, CNRS, Centre de Physique Théorique, Turing Center for Living Systems, Marseille, France.
4 Members of the GROUPOOL & MODCOV19 initiatives.

❍These authors contributed equally to this work.
* rupprecht@cpt.univ-mrs.fr

Abstract

We propose an analysis and applications of sample pooling to the epidemiologic monitoring of COVID-19. We first introduce a model of the RT-qPCR process used to test for the presence of virus in a sample and construct a statistical model for the viral load in a typical infected individual inspired by large-scale clinical datasets. We then propose a method for the measure of the prevalence in a population, based on group testing, taking into account the increased number of false negatives associated with this method. Finally, we present an application of sample pooling for the prevention of epidemic outbreak in closed connected communities.

Author summary

Sample pooling consists in combining samples from multiple individuals into a single pool that is then tested using a unique test-kit. A positive test means that at least one individual within the pool is infected. Sample pooling could provide the means for regular, massive testing for the presence of SARS-COV2 among asymptomatic individuals. We provide a mathematical framework to estimate the proportion of contaminated individuals in populations that includes a reduced sensitivity of diagnostic tests with the pool size. Within our model, we find that large pools sizes are beneficial in the early detection of epidemic outbreaks within communities (e.g. nursing home, university campus), with an optimal size that depends on the sensing capacity of the diagnostic test.

Introduction   1

Regularly monitoring the *prevalence* of a disease , i.e. the proportion of infected   2
individuals within the general population at a given time, is a key element to prevent   3
the onset of an epidemic wave, to estimate the effect of social distancing policies and to 4
anticipate a potential increase in the demand for hospitalization in intensive care   5
units [1].   6
    In the context of the COVID-19 epidemics, contagious individuals are generally   7
assumed to bear a viral load of SARS-CoV-2 in their respiratory tract [2,3]. Such viral   8

load can be detected and quantified within swab samples using a technique called *reverse transcription quantitative polymerase chain reaction* (RT-qPCR) [4]. With tests performed in priority on symptomatic patients, the proportion of positive tests (which corresponds to an *apparent* prevalence [5]) is larger than the prevalence among the whole population, which we call *overall* prevalence. In principle, the overall prevalence could be deduced from the apparent prevalence based on inferred model estimates for the proportion of tested individuals among the infected population. However, such reliable estimation is challenging given (a) the current large uncertainty regarding the proportion of asymptomatic carriers (estimated to be in $20 - 50\%$ range [6–9]) and (b) the variable delay between the contamination and first symptoms, which varies from 1 to 5 days [10, 11].

Testing a large portion of the population at random would allow for a direct measure of the overall prevalence, including the proportion of asymptomatic individuals. Unfortunately, it appears that the production of reactants used in RT-qPCR diagnostic would not meet a demand in regular large-scale population testing [12, 13].

In such context of a shortage in reactants and/or of RT-qPCR machines, group testing has received renewed interest. The principle of group testing consists in combining samples from multiple individuals into a single pool that is then tested using a single test-kit. The pool sample is considered to be positive if and only if at least one individual in the group is contaminated. The idea of group testing is not new, with a long history that dates back to 1943 [14] in the context of syphilis detection, see [15] for a review. Optimal diagnostic strategies include smart-pooling, whereby pools are organised according to lines and columns on a grid –or hypercube– with overlaps enabling the identification of positive individuals [16–18].

Several teams across the world have developed group testing protocols for SARS-CoV-2 infected individuals using RT-qPCR tests. As early as February 2020, pools of 10 have been used over 2740 patients to detect 2 positive patients over the San Francisco Bay in California [19]. A recent publication from Saarland University, Germany, shows that positive sample with a relatively mild viral load from asymptomatic patients could still be detected within pools of 30 [20]. Further works suggest that RT-qPCR viral detection can been achieved in pools with a number of samples ranging from 5 to 64 [21–33].

In parallel, the theoretical literature on group testing for SARS-CoV-2 diagnostic is growing at a fast pace [12, 34–39]. Most of the emphasis has been put on the binary (positive or negative) outcome of tests, with little or no regard on the viral load quantification [4]. Moreover, if the possibility of false negatives is sometimes considered, the increase in the rate of false negatives with dilution of samples due to group testing is rarely taken into account [18].

In this article, we do not address any diagnostic problems, such as the question of determining optimal strategies to provide individual positive diagnostic to a large population using a minimal number of tests. Rather, we propose to evaluate pooling strategies as a tool for the study of epidemiologic questions. In the pooling strategies we discuss below, no individual will be part of two different pools at the same time, so no information on the infection status of any distinguished individual is obtained.

Here, we instead focus on (i) the measure of the overall prevalence and (ii) on the early detection of contamination in a closed community.

The rest of the article is constructed as follows. We first present a simple protocol for the measure of the prevalence in the population by the use of group testing; we make in Section 1 the assumption of *perfect test*, i.e. that the test used is not subject to any false positive or negative, no matter the size of the pool to which the test is applied. In Section 2, we provide a short description of the RT-qPCR and propose a statistical model for its study, that underlines its limit of detection at small concentrations. Then

in Section 3, we analyse part of the information recovered from the quantified viral charge in patients from the clinical dataset of [40]; this gives a sense of the viral load infected patients should carry in the general population, therefore the effect of dilution on the rate of false negatives. Finally, we show in Section 4 how, using the statistical model and the measure of the error rate discussed above, one can measure the viral prevalence in the general population, or design a protocol allowing an early detection of an epidemic outbreak in a closed vulnerable community (e.g. schools, retirement homes, detention centers).

1   Measuring prevalence with perfect tests

We investigate in this section the measure of the prevalence of the disease in a population using a group testing strategy, under the assumption of *perfect* tests, i.e. with no risks of false negative (or false positive). Our derivation is similar to [41].

We assume that we have n tests at our disposal. Given $N \in \mathbb{N}$, we sample nN individuals at random in the general population, and organize n pools of N individuals. Each of these pools is then tested using the perfect tests. For all $i \leq n$, we write $X_i^{(N)} = 1$ if the ith test is positive (i.e. if and only if at least one of the N individuals in the ith pool is infected), and $X_i^{(N)} = 0$ otherwise. We denote by p the (unknown) proportion of infected individuals in the population, then $(X_i^{(N)}, i \leq n)$ forms an independent and identically distributed (i.i.d.) sequence of Bernoulli random variables with parameter $1 - (1-p)^N$.

Lemma 1.1. *Writing $\overline{X}_n^{(N)} = \frac{1}{n} \sum_{j=1}^N X_j^{(N)}$, the quantity $1 - (1 - \overline{X}_n^{(N)})^{1/N}$ is a strongly consistent and asymptotically normal estimator of p. A confidence interval of asymptotic level $1 - \alpha$ is*

$$\mathrm{CI}_{1-\alpha}(p) = \left[1 - (1 - \overline{X}_n^{(N)})^{1/N} \pm \frac{q_\alpha (1 - \overline{X}_n^{(N)})^{1/N-1} \sqrt{\overline{X}_n^{(N)}(1 - \overline{X}_n^{(N)})}}{\sqrt{n}N} \right], \quad (1)$$

where q_α is the quantile of order $1 - \alpha/2$ of the standard Gaussian random variable.

Proof. Note that $(X_j^{(N)}, j \leq n)$ is a standard Bernoulli model, hence $\overline{X}_n^{(N)}$ is a consistent and asymptotically normal estimator of $f(p) = 1 - (1-p)^N$. Hence, using that f^{-1} is $\mathcal{C}^1$ and Slutsky's lemma, we deduce all the above properties of the estimator $f^{-1}(\overline{X}_n^{(N)})$ of p. $\square$

Remark 1.2. As $\lim_{n\to\infty} 1 - (1 - \overline{X}_n^{(N)})^{1/N} = p$ almost surely, for any $N \in \mathbb{N}$ the width of the confidence interval defined in Lemma 1.1 satisfies

$$\frac{2q_\alpha (1 - \overline{X}_n^{(N)})^{1/N-1} \sqrt{\overline{X}_n^{(N)}(1 - \overline{X}_n^{(N)})}}{\sqrt{n}N} \underset{n\to\infty}{\sim} \frac{2q_\alpha}{\sqrt{n}} \frac{(1-p)}{N} \sqrt{\frac{1 - (1-p)^N}{(1-p)^N}} \quad \text{a.s.} \quad (2)$$

In other words, the precision of the measure of prevalence decays as $n^{-1/2}$, with a prefactor that depend on the prevalence p and the number N of individual per pool. There exists an optimal choice of N that minimizes the value of this prefactor, largely improving the precision of the measure.

A classical computation (cf. e.g. [12]) shows that the prefactor in Eq. (2) is minimal when the number of mixed samples per pool is equal to:

$$N_{\mathrm{opt}}^{(\mathrm{perf})} = -\frac{c_\star}{\log(1-p)} \iff (1-p)^{N_{\mathrm{opt}}^{(\mathrm{perf})}} \approx 0.20, \quad (3)$$
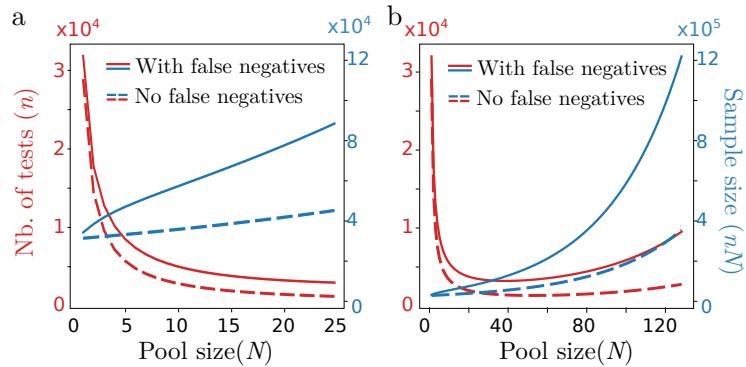
Figure 1. (a,b) Total number of tests (red) and total number of sampled individuals (blue) in order to estimate a prevalence of $p = 3\%$ with a $\pm 0.2\%$ precision with 95% confidence interval as a function of the pool size N for the perfect case (dashed lines) considered in Sec. 1 with no false negative, and the more realistic case (solid lines) considered in Sec. 4 (with false negatives; parameters are defined in Table 2). In (a) N ranges from 0 to 25; in (b) N ranges from 0 to 128; as visible in (b), the valley around the optimal pool size $N_{\mathrm{opt}}^{(\mathrm{perf})} \approx 50$ is large: near optimal savings in tests are achieved even for moderately large pool sizes that require smaller number of individuals to sample.

where $c_\star = 2 + W(-2e^{-2}) \approx 1.59$ and W is the Lambert W function. Specifically, the size of the pools is optimal when approximately 80% of the tests made on the groups turn positive, in sharp contrast with the Dorfman criterium [14].

If we measure the prevalence of the population using group testing, choosing $N = N_{\mathrm{opt}}^{(\mathrm{perf})}$ for the size of the groups, then measuring with a given precision the prevalence will require significantly less tests than if we were to use one test per sampled individual (i.e. if $N = 1$). On the other hand, using this group testing method increases the total number of individuals needed to be sampled, which also has a cost to be considered. However, one can observe that the bottom of the valley of the (red) functions plotted in Fig. 1, that represent the number of tests needed as a function of the size of the pool, is rather wide and flat. There is therefore a large variety of quasi-optimal pool sizes that can be chosen with minimal diminution of the precision in the measure of the prevalence.

Taking the exemple of a prevalence at $p = 3\%$, we expect the pool size minimizing the number of tests needed to read $N_{\mathrm{opt}}^{(\mathrm{perf})} = 50$, which divides by 20 the number of tests needed (see Fig. 1). But to reach similar level of precision than in single testing, the total number of individuals that need to be sampled is more than doubled. Choosing instead a pool size of $N = 20$ requires almost the same number of tests, yet at a cost of a 30% in the total number Nn of sampled individuals (c.f. Fig. 1). The same observation holds for different values of the prevalence, see SI. Fig. S1.

2 Statistical modelling of the PCR

The RT-qPCR technique has been extensively used to estimate the concentration of viral material in samples [42]. We briefly sketch the main steps of an RT-qPCR diagnostic protocol in Box 1. The qPCR typically returns a C_t value, which corresponds to $-\log_2$ of the initial number of DNA copies in the sample, up to an additive constant and measure error. It is measured as an estimated number of cycles needed for the intensity of the fluorescence of the sample to reach a target value (see Fig. 2).

Combined measures of two viral RNA strands with a control of a human RNA strand are recommended in order to detect defective sampling that could induce false negatives, but also to improve precision of the measure as well as to normalize the number of virus copies by the quantity of human DNA [4]. Such combined measure can also improve precision of the measure as well as to normalize the number of virus copies by the quantity of human. We do not intend to include such features in our model. Furthermore, we do not model here the possible errors at the reverse transcription stage, which could lead to some biased measure of the viral load distribution

PCR tests are prone to amplify non-specific DNA sequences [42,43] that can trigger an onset of fluorescence in a samples with no viral SARS-CoV-2 load. These events, called *artefacts* [43], will typically occur beyond a relatively large critical number of cycles, thus imposing the following condition on the diagnosis: a reliable positive result can only be made if the C_t value is lower than a critical value, denoted d_{cens}. Here, the onset of fluorescence in which the virus is absent will be modelled as if triggered by a vanishingly small artificial concentration, denoted ϵ_1.

2.1 Statistical model for the cycle threshold value for a fixed viral concentration

We propose to model the number of cycles threshold value C_t as a random variable, denoted by Y, that depends on the viral load c in the measured sample as

$$Y = -\log_2(c + \epsilon_1) + \epsilon_2, \tag{4}$$

where ϵ_1 is the law of the artefact, modelled as a log-normal distribution with parameters (ν, τ^2); ϵ_2 is an intrinsic measurement error on the threshold value C_t measurement, modelled as a centered Gaussian random variable with variance ρ^2.

As mentioned above, tests are considered to be reliably positive when $Y \leq d_{\mathrm{cens}}$. To avoid false positives, the threshold d_{cens} (with cens for censoring) is chosen such that $\mathbb{P}(\epsilon_1 > 2^{-d_{\mathrm{cens}}}) \ll 1$. Thus, using that as long as a and b are of different orders of magnitude, we have $\log(a + b) \approx \log(\max(a, b))$, we deduce that

$$Y \approx \min(-\log_2(c), d_{\mathrm{cens}}) + \epsilon_2, \tag{5}$$

which obeys the law of a Gaussian random variable with variance ρ^2 and mean $-\log_2(c)$, censored at d_{cens}.

In the idealized no artefact limit ($\epsilon_1 \to 0$), the PCR threshold intensity of a negative patient ($c = 0$) would never be reached ($Y \to \infty$ as well as $d_{\mathrm{cens}} = \infty$).

2.2 Model of the cycle threshold values for pooled samples

We now consider what happens when constructing a pooled sample of N samples. For each $i \leq N$, we write $Z_i = 1$ if the sample i contains a viral RNA load with concentration $C_i > 0$, and $Z_i = C_i = 0$ otherwise. In the rest of the paper, we assume that, in a combined sample created from a homogeneous mixing of the individual samples, the viral concentration reads:

$$C^{(N)} = \frac{1}{N} \sum_{j=1}^{N} Z_j C_j. \tag{6}$$

This assumption relies on the fact that contaminated individuals should have a sufficiently high number of viral copies per sample, so that taking a portion $1/N$ of a virus bearing sample brings a fraction $1/N$ of its viral charge. The result of the

RT-qPCR measure of a grouped test with N individuals is then given by Eq. (4), with $c = C^{(N)}$, hence reads

$$Y^{(N)} = \min\left[\log_2 N - \log_2\left(\sum_{j=1}^{N} Z_j C_j\right), d_{\text{cens}}\right] + \epsilon_2. \tag{7}$$

where $(Z_i, i \leq N)$ are i.i.d. Bernoulli random variables whose parameter is the prevalence of the disease in the population; $(C_i, i \leq N)$ are i.i.d. random variables corresponding to the law of the viral concentration within samples taken from a typical infected individual in the overall population.

Our model Eq. (7) is consistent with the experimental result of [21] as well as [32], whereby linear relations are found between the logarithm of the pool size and the measured C_t that are sufficiently distant from the identified detection threshold.

Remark 2.1. If it were possible to combine samples without dilution (e.g. following the protocol of [33], whereby the exact same volume of each sample is added to the buffer solution as if the sample were being tested individually), Eq. (7) would then be replaced by

$$Y^{(N)} = \min\left[-\log_2\left(\sum_{j=1}^{N} Z_j C_j\right), d_{\text{cens}}\right] + \epsilon_2, \tag{8}$$

in which case, theoretically, pool testing would never loose precision when the pool size increases. That case was treated in Section 1. Hoewever, if the dilution effect occurs for pool sizes exceeding a thresold size K, Eq. (7) would be replaced by

$$Y^{(N)} = \min\left[\log_2(\frac{N}{K})_+ - \log_2\left(\sum_{j=1}^{N} Z_j C_j\right), d_{\text{cens}}\right] + \epsilon_2; \tag{9}$$

where $\log_2(N/K)_+ = 0$ if $N < K$ and $\log_2(N/K))$ otherwise; the analysis would then be similar to what is presented in the rest of the paper, yet with a lower false negative rate.

In order to determine the statistics of the measured cycle $Y^{(N)}$ in a group test of N individuals, we need a distribution for the value of C_j, the viral distribution of infected individuals in the population; this is the objective of the next section.
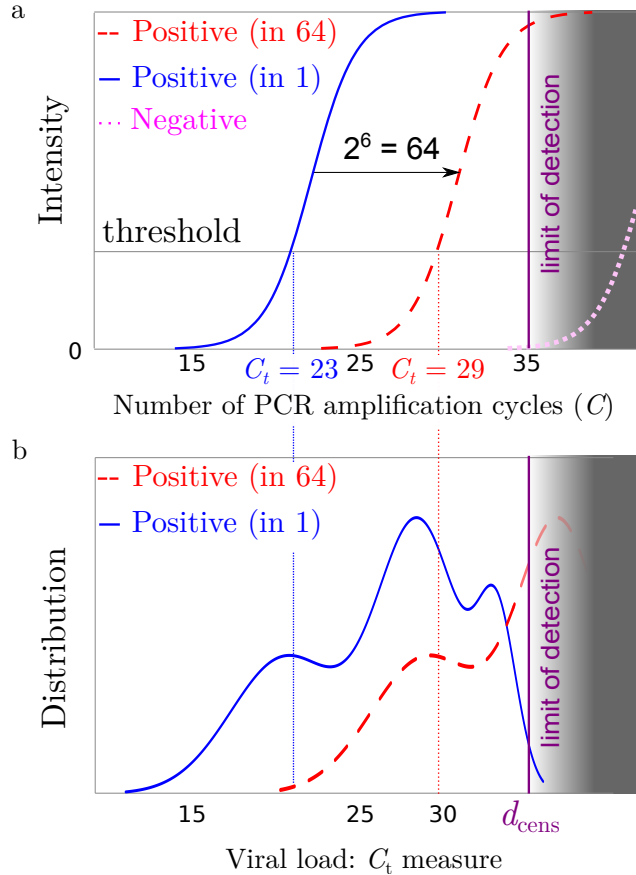
Figure 2. (a) Sketch of an RT-qPCR fluorescence intensity signal for a positive patient without pooling (solid red line) a single positive patient in a pool of 64 patients (dashed red curve) and for a negative sample representing the response of an artefact (dotted magenta curve); as pooling dilutes the initial concentration, the pooled response (dashed red curve) is expected to be close to the translation $x \to x + 6$ from that of a single patient (solid red line). (b) Sketch of the distribution of threshold values for qPCR, either for individual testing (solid blue line) or in pools 64 (dashed red curve); part of the distribution crosses the limit of detection of the test (figured as the grey area) at the detection threshold d_{cens}.

3 Statistical analysis of the viral load measured in positive samples

In [44], the authors propose an analysis of SARS-CoV-2 viral load by patient age. Their analysis is based on the clinically measured viral load of a series of 3,712 infected patients. In particular, in Fig. 1, an histogram representing the frequency distribution of the viral load estimated using RT-qPCR testing. We present in this section a method to analyse these data as a mixture of Gaussian random variables. As RT-qPCR does not allow the measure of viral load below a certain value, we take this fact into account in our statistical modelling by considering the Gaussian in the mixture to be censored after a given threshold d_{cens}.

Unfortunately, as the clinical dataset of [44] is not available, we created a simulated set of measures which reproduces the distribution described by the histogram in Fig. 1 [44]. As the precise distribution of data points within each class of the histogram
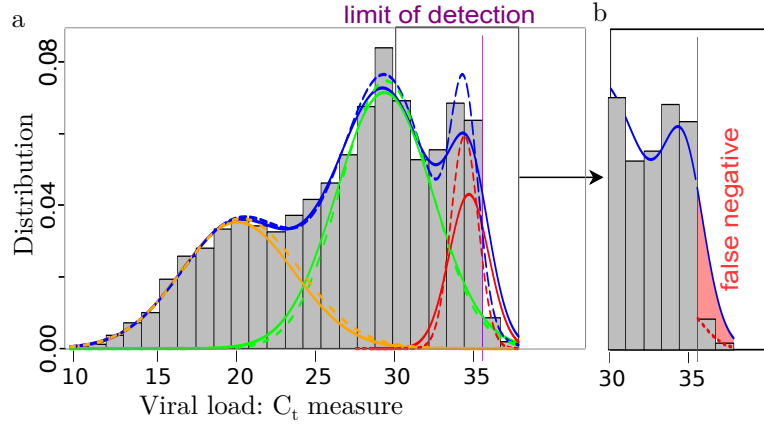
Figure 3. (a) Representation of the density for the classical mixing Gaussian model (dashed lines) and the partially censored model (solid lines) each composed as a sum of 3 components for the Gaussian model (orange/green/red dashed lines) and the partially censored model (orange/green/red solid lines); (purple vertical line) location of the threshold $d_{\text{cens}} \approx 35.6$. (b) Focus on the false negative region, with the estimated false negative probability in the partially censored model (solid line) due to the defect of detection above the threshold d_{cens} (red color filled area).

of Fig. 1 [44] is unknown, we assumed that points were distributed uniformly in their class. We verified the robustness of our estimator for several realizations for the simulated datasets, which lead to consistent values for our model parameters (see SI Sec. 2. B).

Furthermore, as we expect the measure error ϵ_2 of the qPCR to be small with respect to the width of the histogram classes, we set $\rho = 0$ in the rest of the section.

The histogram of our simulated data is plotted in Fig. 3. It is similar to the histogram [44, Fig. 1] but with the x-axis graduated as $-\log_2$ of the viral concentration (to obtain an estimation of the attended C_t value). We observe the presence of a wide and rather low first bump, centered around the value of 20. Additionally, two taller but less wide bell shapes seem to be present, centered around the values 29 and 34 respectively. After the value of about 35.5, there is a significant drop in the number of registered values, very likely due to a sharp detection limit.

The presence of three well-marked distinct bell shapes in the histogram suggests to model the density of $-\log_2$ of the viral load of contaminated patients as a mixture of Gaussian distributions. Heuristically, the patient population can be divided into several groups, which we call clusters, such that patients in cluster i have a C_t value distributed according to a normal distribution with parameters (μ_i, σ_i).

In a first paragraph A., we use a classical statistical tool to estimate the number of clusters and their parameters (frequencies, means, variances). However, this first Gaussian modelling does not take into account the sudden drop observed for values around $C_t \sim 35.5$, hence the estimated density does not fit well to such drop. This seems to be due to the sudden loss of sensitivity of qPCR measure for low viral load.

To better represent the variables, we thus introduce a (partially) censored Gaussian model in Section 3.2, which, to our knowledge, was not previously proposed.

Finally in Section 3.3, we construct a model consisting of a mixing of censored Gaussian variables, that we apply to the data of [44]. This model fits better the data with stable estimation of the parameters and provides further validation of the previous three-cluster analysis of the tested population.

3.1 Mixture model

The shape of the histogram Fig. 3 suggests that the law of the viral load is distributed according to a mixture of three or more Gaussian distributions. The typical decomposition into three clusters is represented in dashed lines in Fig. 3(a). We also plotted in SI. Fig. S4 the histogram with an example of the estimated density with 3 and 4 clusters.

We observe that the cluster associated to the lowest viral load, i.e. the highest C_t value (red curve in Fig. 3) has a small variance. However, as recalled by [44], there is a loss of sensibility of the measure for very small viral loads, which can explain the drop in the number of detected cases around $d_{\mathrm{cens}} = 35.6$. This motivates the introduction of the censored Gaussian model in the next section.

3.2 Censored model and partially censored model

To model a partial lack of detection of low viral load (C_t higher than a threshold d_{cens}), we introduce the partially censored Gaussian variable as a building block for the representation of the density of the viral load in infected patients. However, depending on the machine tuning, the C_t value of the sample can shift by an additive constant, which is computed by measuring the C_t value of a standard solution of viral DNA to tare the measure. Then, some tests might allow the detection of lower viral loads than others.

In view of the shape of the histogram Fig. 3, it is reasonable to assume that the C_t value of patients in a cluster follows a Gaussian distribution. To model this partial censorship phenomenon, we assume that if the sampled C_t value is lower than the detection threshold d_{cens}, the measure is always made for samples detected as positive. If the value is higher than d_{cens}, the sample will be detected with probability q, and its measure will be registered. Otherwise, it will be discarded as a (false) negative, with probability $1 - q$. The parameter q represents the probability of detection of a viral load that falls below the detection threshold of some PCR measures.

Remark 3.1. The assumption that the probability of detection only depends on whether the C_t value is higher than a fixed threshold is of course an important simplification, as one would expect lower viral loads to be more difficult to detect than higher ones. However, the simplicity of this model allows us to study it as a three parameters statistical model, and to construct simple estimators for these parameters. Additionally, it fits rather well the available data, and fitting a more complicate censorship model would require a lot of measures of C_t values close to the detection threshold d_{cens}.

We call this statistical model a *partially censored Gaussian model*, denoted by $\mathcal{CN}_{d_{\mathrm{cens}}}(\mu, \sigma, q)$, with μ and σ the mean and standard deviation of the Gaussian variable before censorship and q the detection probability above the threshold. If we denote by X the random variable, $f_{\mu,\sigma}$ (resp. $F_{\mu,\sigma}$) the density (resp. the cumulative distribution function) of a Gaussian law $\mathcal{N}(\mu, \sigma)$ then the density of X is defined for every $x \in \mathbb{R}$ by:

$$f_X(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{\mathrm{cens}})} \times \left\{ \begin{array}{l} 1 \text{ if } x \leq d_{\mathrm{cens}}, \\ q \text{ otherwise.} \end{array} \right. \tag{10}$$

Remark 3.2. In the absence of censorship (i.e. in the limits $q \to 1$ or $d_{\mathrm{cens}} \to +\infty$), we check that Eq. (10) converges to a Gaussian density distribution.

To avoid the problem of modelling of the partial censorship described in Remark 3.1, a solution that we implement here as a comparison tool, is *to forget* the values after the threshold and to fit a *completely censored* model (i.e. with $q = 0$) to the remaining data, that we denote by $\mathcal{CN}_{d_{\mathrm{cens}}}(\mu, \sigma) = \mathcal{CN}_{d_{\mathrm{cens}}}(\mu, \sigma, 0)$, with density defined for every $x \in \mathbb{R}$

by

$$f_X(x) = \frac{f_{\mu,\sigma}(x)}{F_{\mu,\sigma}(d_{\text{cens}})} \mathbb{1}_{\{x \leq d_{\text{cens}}\}}, \tag{11}$$

where $\mathbb{1}_{\{x \leq d_{\text{cens}}\}}$ is the indicator function equal to 1 if $x \leq d_{\text{cens}}$, and 0 otherwise.

Due to the presence of the cumulative distribution function of a Gaussian law in the denominator in the normalization constant, it is not possible to obtain analytical forms of the parameter estimators. Nevertheless, we can estimate the parameters using an optimization algorithm like the R function nlm (available in [45]) which implements a Newton-type algorithm. The following Theorem 3.3 guarantees the quality of the maximal likelihood estimators, hence of the estimations if the maximization procedure is done correctly.

Theorem 3.3. *The estimators* $(\widehat{\mu}, \widehat{\sigma}, \widehat{q})$ *of* (μ, σ, q) *obtained by maximisation of the likelihood ratio are strongly consistent and asymptotically normal.*

The properties of the maximum likelihood estimators is a consequence of the fact that the (partially) censored Gaussian model belongs to the family of exponential laws (c.f. [46, Chapter 9] and SI B.). To check the quality of the approximation of the estimators by nlm, we simulate variable sizes of samples distributed according to the censored Gaussian model. The values of these estimations are plotted in SI B.3.

3.3 Censored mixture model

We apply here the statistical analysis described in the previous section to simulated data based on the values for the viral load distribution found in [44] with a mixture model and a censoring threshold $d_{\text{cens}} \approx 35.6$ (so the two rightmost bars in the histogram of Fig. 3, that appear much smaller than the nearby values, are supposed to be censored). It is reasonable to assume that the censoring threshold has the same value for each sub-population, as it depends on the test methodology rather than on the tested individuals. In Fig. 3, we represent the histogram with the density for the mixture.

We observe that the separation in sub-populations and the resulting densities are very close to the ones obtained in the "naive" classical Gaussian mixture model, constructed without taking into account the detection threshold. The principal difference between the naive and censored models consists, for the later, in a larger variance that extends above the threshold. To a lesser extent, the sub-population with a median concentration can also exceed the threshold. It is worth mentioning that as expected, the probability of detection below the threshold value is sensibly the same for all three clusters (around 20%).

As a result, using the computed estimates (see Tab. III) and the model, we can calculate a theoretical false negative rate, see SI. Eq. [S3]: in this case, the value is approximately 3.8% (represented by the red area on the Fig. 3b); it mostly belongs to the third cluster. Besides being based on simulated data, hence subject to caution, this value should be treated as a lower bound, as it is possible that a fourth cluster of infected individuals exists but with very small viral load, below the detection threshold. It allows us to predict for example that at least 150 clinical tests in the series of [44] might have resulted in false negatives due to their low viral load.

To validate the censored model, we can verify that if one (i) erases the data to the right of a certain value and (ii) uses the totally censored model on the remaining data, a similar estimate should be obtained for the parameters. We display in SI. Fig. S7 the density obtained using the censored mixture estimation with $d_{\text{cens}} \approx 35.6$, 34.4 and 33.2 (removing the first two, the third, then the fourth rightmost bars in the histogram). We observe that the first and second components are globally unchanged. The mean and standard deviation of the last component are almost the same for $d_{\text{cens}} \approx 34.4$ and

$d_{\text{cens}} \approx 35.6$ (see SI. Tab. IV); only the proportions naturally decrease with the threshold. On the other hand, the mean moves slightly to the left for $d_{\text{cens}} \approx 33.2$; this is due to the fact that we loose the information of the largest bars of this component. It might also be caused by our ignorance of the exact distribution of C_t values within classes of the histogram (we recall that we assume that it is a uniform distribution).

Note that if we were to set the threshold at $d_{\text{cens}} \approx 34.4$ as threshold for the partially censored model without erasing data, the optimization procedure `nlm` would not converge. This is further indication that a detection drop happens in the neighbourhood of 35.6.

3.4 Application to other datasets

We applied a similar statistical analysis to simulated datasets of 852 infected nursing home resident and workers studied by [27]. For this dataset, two to three sub-populations are identified by our algorithm. The estimation obtained for the Gaussian fit of the C_t distribution they obtained is given in SI Tab. 3. It would be interesting to link these observed sub-populations to characteristics of the individuals (e.g. age of the patients or stage of the disease). We mention that in two other datasets of smaller size [21, 47], the statistical resolution does not allow us to distinguish between several sub-populations; we rather found that the distribution of C_t corresponds to a single Gaussian with standard deviation σ in the 5 to 6 range.

4 Group testing: application to epidemiology

We now show how the previous analysis of the tests used to measure the viral load in patients can be used to precise the epidemiological monitoring of the disease in the general population.

4.1 The number of infected individuals in pools cannot be recovered

We expect the PCR result to correspond to the sample with the highest viral load, up to a dilution-induced drift $\log_2(N)$, under the model hypothesis of Sec. 2 (cf. Fig. 2). Indeed, since the viral concentration in randomly selected infected individuals spans several order of magnitudes, we expect that

$$\log_2\left(\frac{1}{N}\sum_{i=1,\dots,j} C_i\right) \approx \log_2\left(\max_{i=1,\dots,j} C_i\right) - \log_2(N), \qquad (12)$$

for j positive samples with concentration C_j diluted in a pool of N.

We point out that the RT-qPCR viral load measure could be used to improve efficiency and cross testing of smart pooling type diagnostic methods, which are beyond the scope of this paper. We plan to investigate this aspect in future work.

Therefore, the measured value of the pooled sample viral concentration cannot be used to estimate the number of infected individual within the pool.

4.2 Group testing and the measure of viral prevalence

Here, in contrast with Sec. 1, we no longer consider that the RT-qPCR tests to be perfect. As discussed in Eq. (7), we model the concentration of the pooled sample as the average of the individual sample loads; and we assume that viral concentration becomes undetectable below a given threshold. Therefore, creating groups has the effect of
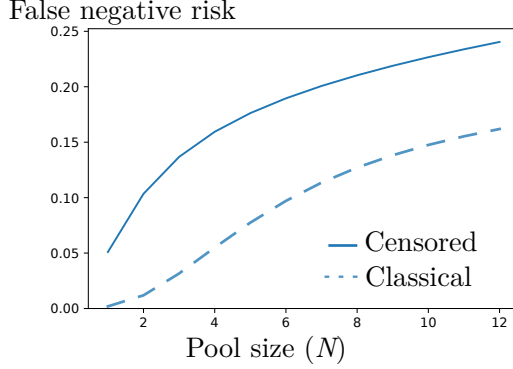
Figure 4. Estimated false negative risk probability of a sample pooling test containing a single infected individual (with a viral load distributed according to the simulated statistics presented in Sec. 3) as a function of the number of pooled individuals N, using the estimated density obtained for the classical (dashed line) and censored (solid line). The censored and classical model give different estimates of (and probably both underestimate, as we did not consider unsuccessful sampling) the false negative rate in group pooling under the model assumption of Eq. (6); such estimates will have impact on the estimation of the prevalence.

increasing the false negative rate, which has to be quantified. We then use this 346
estimation to un-bias the estimator of the prevalence in the overall population based on 347
group testing, and study its impact on the optimal choice of group sizes. 348

4.2.1 Estimation of the false negative rate induced by pooling 349

The distribution of the viral load of a single positive sample within a pool of several 350
negative samples appears as shifted towards higher C_t-values, see Fig. 2. A pooled 351
sample returns positive only if the average concentration is smaller than d_{cens}; thus 352
using the observation of Sec. 3.3, contamination will be detected in a group of N 353
individuals typically if at least one individual in the group has a viral charge larger than 354
$N2^{-d_{\text{cens}}}$. Therefore, there is a risk that low viral charge samples (that would have been 355
tested positive using individual tests) would no longer be positive in pool tests. 356
Similarly to Eq. (5), we express the increased rate of false negative due to pooling as 357
$\mathbb{P}(-\log_2(C) + \epsilon_2 \geq d_{\text{cens}} - \log_2(N))$, where $\log_2(C)$ is the viral concentration of the 358
positive individual. For simplicity we neglect the measurement error of the qPCR, i.e. 359
considering that $\rho = 0$, thus an expression for the increased rate of false negatives reads 360
$1 - \Phi(d_{\text{cens}}^{(N)})$, where 361

$$\Phi\left(d_{\text{cens}}^{(N)}\right) = \mathbb{P}\left[-\log_2(C) \leq d_{\text{cens}} - \log_2(N)\right]. \tag{13}$$

We find that, when estimated by the censored model, the false negative risk function 362
$\Phi(d_{\text{cens}}^{(N)})$ grows quicker as the pool size increases than in the uncensored model, see 363
Fig. 4. This is partly due to the fact that the censored model makes the assumption 364
that $d_{\text{cens}} \approx 35.6$, whereas $d_{\text{cens}} \approx 37.3$ in the uncensored model. Choosing a correct 365
statistical model for the distribution of C_t values has a critical impact on the estimation 366
of the false negative risk, therefore on the estimate of the prevalence, as is discussed 367
next. 368

Table 1. Table of the pool size as a function of the number of tests for a prevalence of 3% measured with a precision of 0.2% at a 95% confidence interval, for both perfect tests (with no false negatives, see Sec. 1) and imperfect tests (with false negatives; model parameters defined in Table 2); computed using Eqs. [1] and [14].

Pool size N	Perfect tests		Imperfect tests	
	Number of tests n	Sample size nN	Number of tests n	Sample size nN
1	29100	29100	29464	29464
2	14775	29550	15069	30138
3	10003	30009	10261	30783
5	6191	30955	6411	32055
10	3350	33500	3530	35300
20	1973	39460	2130	42600
30	1561	46830	1716	51480
50	1349	67450	1525	76250
100	1884	188400	2235	223500
200	10378	2075600	13105	2621000

4.2.2 Correction of the prevalence estimate by the false negative rate

Assuming a false negative rate of $1 - \Phi(d_{\text{cens}}^{(N)})$ in pool testing with groups of size N, we observe that $1 - (1 - \overline{X}_n^{(N)})^{1/N}$ (as defined using the notation of Section 1) is a consistent estimator of $p\Phi(d_{\text{cens}}^{(N)})$ (c.f. Lemma 1.1). As a result, the confidence interval constructed for the prevalence p now reads

$$
\text{CI}_{1-\alpha}(p) = \left[\frac{1 - (1 - \overline{X}_n^{(N)})^{1/N}}{\Phi(d_{\text{cens}}^{(N)})} \pm \frac{q_\alpha}{\sqrt{n}} \frac{(1 - \overline{X}_n^{(N)})^{1/N-1}\sqrt{\overline{X}_n^{(N)}\left(1 - \overline{X}_n^{(N)}\right)}}{N\Phi(d_{\text{cens}}^{(N)})} \right]. \quad (14)
$$

For the numerical applications presented in Fig. 1 and Table 1, we consider a viral load C that is distributed according to Eq. (11). As expected, due to false negatives, we find that the number of tests needed to reach a given precision on the prevalence is increased, but relatively moderately. In particular, the optimal pool size value, $N_{\text{opt}}^{(\text{imper})}$, that minimizes the number of tests needed to reach a given precision level, is close to the value $N_{\text{opt}}^{(\text{perf})}$, defined in Eq. (3).

Similarly, one can observe that using a different distribution with similar mean and variance for $-\log_2 C$ as Eq. (11) would lead to moderate changes of the values estimated in Table 1. While modelling of the viral load of an infected individual is crucial to un-bias the estimator of the prevalence via group testing, the practical implementation of such group testing strategy, i.e. the choice of the group size N and the number n of tests to use, is relatively independent of the precise statistical properties of the viral load distribution.

Based on Eq. (14), in Box 2. we propose an iterative method to estimate p, which, during a survey, allows for on-the-fly adaptations of the pool size.

4.3 Group testing and Bayesian inference of the prevalence in sub-categories of the population

The viral prevalence may vary significantly among specific categories within the overall population. In particular, a prevalence reaching 5% was measured among the health

<div style="border: 1px solid; background-color: #fce9cf; padding: 10px;">

Box 2: A protocol of prevalence determination

We propose the following procedure for the measure of prevalence via group testing:

1. Start from an a priori estimate for the prevalence ($\hat{p}_0$).

2. Based on the value of $\hat{p}_0$, estimate the number N of individuals in the pool that minimizes the total number of tests needed to achieve the estimation of the prevalence p at the targeted precision and confidence interval.

3. Construct a number of n pools containing each N individuals selected at random in the general population, with n the number of tests available for the measure.

4. Count the number of positive tests and compute the average $\overline{X}_n^{(N)}$.

5. An improved estimate of the prevalence then reads: $\hat{p}_1 = 1 - (1 - \overline{X}_n^{(N)})^{1/N}$ (cf Lemma 1.1).

Note that this method can easily be adapted into a Bayesian algorithm, with the number N of individuals tested modified at each iteration of the procedure.

</div>

care workers population in a hospital [48], which we expect to be significantly higher than the estimate prevalence within the general population.

Here we show that we do not specifically need pool samples from individuals from homogeneous categories in order to recover the distribution of prevalence within these categories.

The protocol described in Box 2 can be adapted to study different prevalences in specific sub-populations, provided that the number of individuals of each subpopulation is known for every grouped sample. In SI. Fig. S3 , we evaluate, as function of the number of tests, the credibility intervals on the prevalence within two categories of the population: one at $p_1 = 5\%$ representing 20% of the total population (a value inspired by [48]), the other being at $p_2 = 0.5\%$ (a value inspired by [49]). More information on this adaptative protocol can be found in the SI.

Remark 4.1. Note that once a difference in prevalence is noted from the epidemiological study of the general population, testing can be adapted to construct groups containing only members of one subpopulation to attain similar levels of precision for the prevalence of the sub-populations. The prevalence in the general population can then be recovered by averaging the estimators of the sub-populations. The advantage of these adaptative settings is that the existence of a difference of prevalence in populations can be tested before deployment of resources needed to measure them specifically.

4.4 Group testing for regular monitoring in communities

We now consider some applications of group testing to the early detection of an epidemic outbreak within a community, that is interconnected and reasonably closed to the outside (e.g. schools, nursing homes, detention centres).

Our results are based on a minimal working model for epidemic outbreaks within a community of A individuals. At a random date, which we choose to be time $t = 0$, the patient 0 in the population gets contaminated, and immediately starts infecting members of its community at rate λ. Each newly infected individual then contributes to spreading the disease at the same rate λ.

4.4.1 Optimization of the size of pools

We first consider the impact of group testing strategy, consisting in k group test with pools of N individuals, on the early time of the outbreak $t \ll \lambda^{-1}$. With a unique contaminated individual in the population, the detection probability reads

$$\mathbb{P}\left[+|k \text{ tests}\right] = kN\Phi_0(d_{\text{cens}}^{(N)})/A, \quad \text{with } kN \le A, \tag{15}$$

where $\Phi_0(d_{\text{cens}}^{(N)})$ is defined according to Eq. (13), with the difference that the assumed viral load of the patient 0, corresponding to that measured at early times, may need not be equal to the distribution estimated in Eq. (11) based on clinical data. For simplicity, we will assume in the following that Φ_0 is the cumulative distribution of a log-normal viral load distribution $\text{logN}(\mu_0, \sigma_0)$ of mean μ_0 and variance σ_0.

In Fig. 5, we represent the evolution of the probability to detect the patient 0 as a function of the total number of sampled individuals in a population of size $A = 120$. We observe that if μ_0 is close enough to d_{cens}, i.e. if the viral charge of the patient 0 is close to being undetectable, then there will exist an optimal size for the pools. When N becomes too large the risk of false negative overcomes the potential benefits of testing larger portions of the community (see Fig. 5a). In contrast, if the viral load of patient 0 is slightly higher, the detection probability becomes a monotonic function of the pool size N, indicating that larger pools are always beneficial. Additionally, if using multiple tests increases the detection probability when the viral load is close to the detection threshold, using multiple tests has a smaller impact when the viral load gets easier to detect.

Table 2. Table with standard parameter values (with std. the abbreviation of standard deviation).

Symbol	Meaning	Value
d_{cens}	Maximal cycle number	35.6
μ_i, σ_i, p_i	Viral load (in C_t) distribution fits	Table IV
ρ	PCR measurement error (std.)	0
ϕ	Delay before onset of symptoms	5 days
λ	Intra-community contamination rate	$0.5 \, \text{days}^{-1}$
r	Asymptomatic probability	40 %
τ	Time interval between grouped tests	$1 - 12$ days
A	Total number in the community	120 or 1000
N	Pool size	1–128
μ_0 ; σ_0	Viral load (in C_t), patient 0 (mean, std.)	$30 - 35$
ζ	Defective sampling probability	0

4.4.2 Optimization of the regularity of tests

We now consider the impact of a regular testing strategy every τ units of time, with k pools of N individuals with individuals drawn at random in the population. To compare the impact of the frequency of testing for a constant budget, we assume that the ratio k/τ is equal to a fixed constant, representing the amount of tests spent per unit of time by the community to detect infections. What we consider here is a continuous-time steady-state version of the previous paragraph 4.4.1.

We denote by r the proportion of asymptomatic individuals. Symptomatic individuals start showing signs of being contaminated a given number of days after infection, denoted by ϕ (that we assume to be constant in this simplified model). The
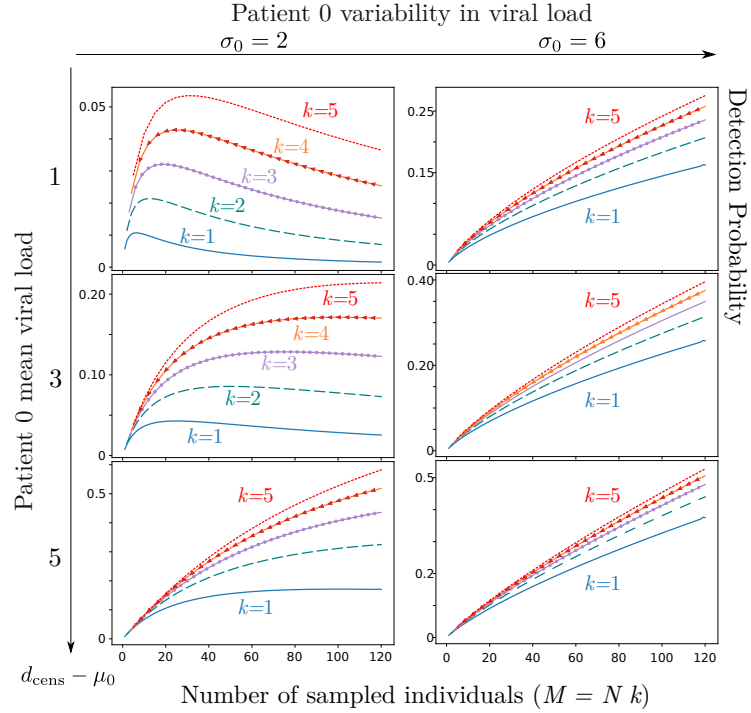
Figure 5. Detection probability of a single patient 0 with low viral load within a community of 120 as a function of the total number of sampled individuals $M = k \times N$, where k is the total number of tests used and N the number of samples pooled together in a test, with $k = 5$ (red dotted line); $k = 4$ (orange line with arrow), $k = 3$ (purple line with circles); $k = 2$ (dashed green line); $k = 1$ (solid blue line) for several values in the parameters describing the viral load of the patient 0 at the onset of contagiosity, expressed in terms of a normal distribution in C_t (the number of RT-qPCR amplification cycles) with a standard deviation σ_0 and a mean μ_0 and a threshold at a value denoted d_{cens} satisfying: $\mu_0 = d_{\text{cens}} - 1$ (top row), modelling a patient 0 with a very low viral concentration, $\mu_0 = d_{\text{cens}} - 3$ (middle row), $\mu_0 = d_{\text{cens}} - 5$ (bottom row); $\sigma_0 = 2$ (left column); $\sigma_0 = 6$ (right column).

outbreak is detected either if one of the screening tests finds a contaminated individual, or if one of the contaminated individuals shows symptoms. We denote by T_s the epidemic detection time without screening tests (as defined as the first detection of a symptomatic individual) and T_d is the one through random tests screenings time, see Fig. 6.

In this model, the number of infected individuals at the date t, denoted by $N(t)$, is distributed according to a geometric random variable with parameter $1 - e^{-\lambda t}$, as expected for such a Yule process [50]. In the absence of screening tests, we find that the average detection time reads $\langle T_s \rangle = \phi - \log(1-r)/\lambda$. The average number of contaminated individuals at that time then reads: $\langle N(T_s) \rangle = e^{\lambda \phi}/(1-r)$. In comparison, based on Eq. (15), we find that the first screening test after contamination will detect the outbreak with a probability

$$\mathbb{P}\left[+|k \text{ tests}\right] = \frac{k(e^{\lambda \tau} - 1)\Phi(d_{\text{cens}}^{(N)})}{\lambda \tau A}. \tag{16}$$

By the time of detection $t = \tau$, a number of $\langle N(\tau) \rangle = (e^{\lambda \tau} - 1)/\lambda \tau$ individuals has been contaminated.
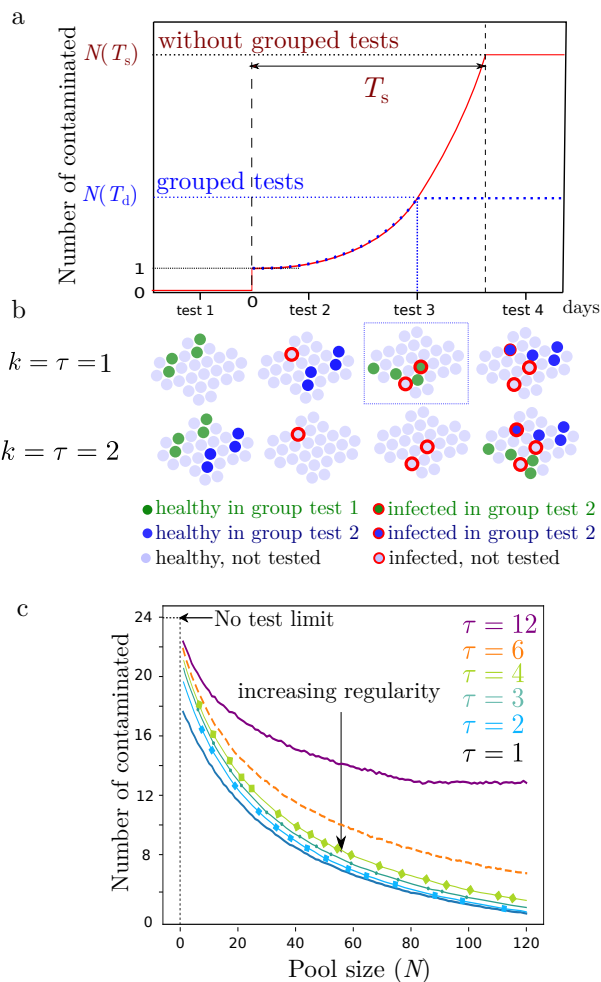
Figure 6. (a) Sketch of the time evolution of the number of contaminated individuals in a community. The patient 0 is contaminated from the outside of the community 0.8 units of time after a test date. In the absence of screening tests, the contamination is detected at the time $T = T_s$ (after appearance of the first symptoms); with grouped tests, an infected individual is detected at a time $T = T_d$. (b) Sketch of two group testing strategies, here with pools of size $N = 4$, one with a single ($k = 1$) grouped tests every day ($\tau = 1$); the other with $k = 2$ grouped tests every second day ($\tau = 2$); the second strategy (least frequent testing) fails to detect the outbreak early and results in more contamination. (c) Number of infected individuals at the detection of the outbreak as a function of the pool size, using $k = \tau$ tests performed at τ-day intervals, with $\tau = 12$ (solid purple line), $\tau = 6$ (dashed orange line), $\tau = 4$ (dark green solid line with square), $\tau = 3$ (light green solid line with circles), $\tau = 2$ (cyan line with squares) and $\tau = 1$ (solid blue line). Here we consider a large community composed of $A = 1000$ individuals. The patient 0 has a viral load concentration distributed according to a log-normal distribution with mean $\mu_0 = 30$ and standard deviation $\sigma_0 = 2 \log_2(2)$; all others parameters can be found in Table II.

Computing the average number of contaminated individuals $\langle N(T_d) \rangle$, we find that a screening strategy consisting in sampling a random subgroup of the community as frequently as possible is more efficient than the one consisting in testing larger portion of the community at less frequent time intervals. In Fig. 6, we compare different

461
462
463
464

screening scenarios for a large community composed of $A = 1000$ individuals. We vary the value of the screening time interval τ while keeping fixed (1) the average number of tests per unit of time and (2) the size of the pools on which each test is used. Our simulation range from checking N individuals every day (with one test) to checking $12 \times N$ individuals in 12 pools every 12 days.

4.4.3 Discussion

In Fig. 6, we chose to consider the case of presymptomatic patient 0 with a weak viral load ($\mu_0 \approx 30$); we find that a pooling method remains efficient for monitoring purposes even in such unfavorable settings. The recent study [51] indicates that a large fraction of presymptomatic individuals detected in a nursing homes had relatively high viral loads (in the $\mu_0 \approx 20 - 25$ range), which tends to indicate that screening methods based on pooling would be even more efficient than suggested in Figs. 5 and 6.

We also point out that, here, we have used a minimal set of parameters in order for analytical calculations to be tractable. Including more parameters (e.g. considering a time-dependent infection rate or viral charge for patients after their contamination, graph of relationship within the community) would be needed in order to obtain conclusive results to be used as healthcare guidelines. In this direction, based on stochastic simulations encompassing a large set of parameters, [52] also concludes on the efficiency of group testing in preventing epidemic outbreaks in health care structures. We hope our analytical calculations will inspire further epidemiological investigations. Without such further studies, our results should not be used literally as a guideline for clinical practice/healthcare related behaviour.

Conclusion

We consider the effect of sample dilution in RT-qPCR grouped tests and we propose a model to describe the risk of false negatives as a function of the pool size. We present a procedure to analyse experimental datasets for the viral charge of patients. Inspired by the clinical study [44], we expect the statistics of the number of amplification cycles to be well described as a mixture of 3 Gaussian variables censored at the RT-qPCR sensibility limit. Our analysis may hint at the existence of 3 sub-classes of interest that could each be interpreted in terms of medical or physiological criteria [47].

We point out that the viral load distribution that we determine here based on the clinical sampling presented in [44] is possibly biased as compared to the viral load that would be measured within a population targeted by a group testing strategy. Indeed, asymptomatic individuals may exhibit a different viral load distribution than the population that got tested. Furthermore, the measures of [44] were based on nasal swabs samples, and their statistical distribution is likely to vary according to the sampling method chosen for group testing.

Tests based on saliva samples appear to show promising results in terms of false negative risks while improving the acceptance of tests among the population, decreasing the time needed for sample collection and reducing the exposure of health care practitioners [53–57]. In this context, we are looking forward to seeing whether group testing on saliva samples would provide reliable results; statistics on the distribution of the viral load in saliva samples would then be of interest to evaluate the efficiency of a group testing strategy based on such samples.

We think group testing could provide the means for regular and massive screenings allowing the early detection of asymptomatic and pre-symptomatic individuals – a particularly crucial task to succeed in the containment and potential eradication of the epidemic [10, 58, 59].

Acknowledgments ₅₁₃

Code accessibility ₅₁₉

The codes used in this paper can be found at the following address: $_{520}$
`https://github.com/Lionning/CovidPooling` $_{521}$

References ₅₂₂

1. Salje H, Tran Kiem C, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, et al. $_{523}$ Estimating the burden of SARS-CoV-2 in France. Science. $_{524}$ 2020;3517(pasteur-02548181):3517. doi:10.1126/science.abc3517. $_{525}$

2. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. $_{526}$ Virological assessment of hospitalized patients with COVID-2019. Nature. $_{527}$ 2020;doi:10.1038/s41586-020-2196-x. $_{528}$

3. World Health Organization. Laboratory testing for 2019 novel coronavirus $_{529}$ (2019-nCoV) in suspected human cases. WHO - Interim guidance. $_{530}$ 2020;2019(January):1–7. $_{531}$

4. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. $_{532}$ Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. $_{533}$ Eurosurveillance. 2020;25(3):1–8. doi:10.2807/1560-7917.ES.2020.25.3.2000045. $_{534}$

5. Speybroeck N, Devleesschauwer B, Joseph L, Berkvens D. Misclassification errors $_{535}$ in prevalence estimation: Bayesian handling with care. International Journal of $_{536}$ Public Health. 2013;58(5):791–795. doi:10.1007/s00038-012-0439-9. $_{537}$

6. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic $_{538}$ proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond $_{539}$ Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020;25(10):1–5. $_{540}$ doi:10.2807/1560-7917.ES.2020.25.10.2000180. $_{541}$

7. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission $_{542}$ of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a $_{543}$ retrospective cohort study. The Lancet Infectious Diseases. 2020;3099(20):1–9. $_{544}$ doi:10.1016/S1473-3099(20)30287-5. $_{545}$

8. Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed Asymptomatic $_{546}$ Carrier Transmission of COVID-19. JAMA. 2020;323(14):1406. $_{547}$

9. Lavezzo E, Franchin E, Ciavarella C, Cuomo-dannenburg G, Barzon L, Sciro M, $_{548}$ et al. Suppression of COVID-19 outbreak in the municipality of Vo ', Italy. $_{549}$ medRxiv. 2020; p. 1–23. doi:10.1101/2020.04.17.20053157. $_{550}$

10. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. $_{551}$ Quantifying SARS-CoV-2 transmission suggests epidemic control with digital $_{552}$ contact tracing. Science. 2020;368(6491):6936. doi:10.1126/science.abb6936. $_{553}$

11. Sethuraman N, Jeremiah SS, Ryo A. Interpreting Diagnostic Tests for   554
    SARS-CoV-2. JAMA. 2020;.   555

12. Gollier C, Gossner O. Group Testing against Covid-19 *. Covid Economics.   556
    2020;(2):32–42.   557

13. Pouwels KB, Roope LSJ, Barnett A, Hunter DJ, Nolan TM, Clarke PM. Group   558
    Testing for SARS-CoV-2: Forward to the Past? PharmacoEconomics - Open.   559
    2020;(0123456789):2–5. doi:10.1007/s41669-020-00217-8.   560

14. Dorfman R. The Detection of Defective Members of Large Populations. The   561
    Annals of Mathematical Statistics. 1943;.   562

15. Aldridge M, Johnson O, Scarlett J. Group Testing: An Information Theory   563
    Perspective. Foundations and Trends in Communications and Information   564
    Theory. 2019;15(3-4):196–392. doi:10.1561/0100000099.   565

16. Barillot E, Lacroix B, Cohen D. Theoretical analysis of library screening using a   566
    N-dimensional pooling strategy. Nucleic Acids Research. 1991;19(22):6241–6247.   567
    doi:10.1093/nar/19.22.6241.   568

17. Thierry-Mieg N. Pooling in systems biology becomes smart. Nature Methods.   569
    2006;3(3):161–162. doi:10.1038/nmeth0306-161.   570

18. Furon T. The illusion of group testing. [Research Report] RR-9164, Inria Rennes   571
    Bretagne Atlantique. 2018; p. 1–19.   572

19. Hogan CA, Sahoo MK, Pinsky BA. Sample Pooling as a Strategy to Detect   573
    Community Transmission of SARS-CoV-2. JAMA. 2020;323(19):1967.   574
    doi:10.1001/jama.2020.5445.   575

20. Lohse S, Pfuhl T, Berkó-Göttel B, Rissland J, Geißler T, Gärtner B, et al.   576
    Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. The   577
    Lancet Infectious Diseases. 2020;3099(20).   578

21. Yelin I, Aharony N, Shaer Tamar E, Argoetti A, Messer E, Berenbaum D, et al. 579
    Evaluation of COVID-19 RT-qPCR test in multi-sample pools. Clinical Infectious 580
    Diseases. 2020;doi:10.1093/cid/ciaa531.   581

22. Ben-Ami R, Klochendler A, Seidel M, Sido T, Gurel-Gurevich O, Yassour M,   582
    et al. Pooled RNA extraction and PCR assay for efficient SARS-CoV-2 detection. 583
    medRxiv. 2020; p. 2020.04.17.20069062. doi:10.1101/2020.04.17.20069062.   584

23. Shental N, Levy S, Skorniakov S, Wuvshet V, Shemer-Avni Y, Porgador A, et al. 585
    Efficient high throughput SARS-CoV-2 testing to detect asymptomatic carriers. 586
    medRxiv. 2020; p. 2020.04.14.20064618. doi:10.1101/2020.04.14.20064618.   587

24. Schmidt M, Hoehl S, Berger A, Zeichhardt H, Hourfar K, Ciesek S, et al. FACT- 588
    Frankfurt adjusted COVID-19 testing- a novel method enables high-throughput 589
    SARS-CoV-2 screening without loss of sensitivity. medRxiv. 2020; p.   590
    2020.04.28.20074187. doi:10.1101/2020.04.28.20074187.   591

25. Mutesa L, Ndishimye P, Butera Y, Uwineza A, Rutayisire R, Musoni E, et al. A 592
    strategy for finding people infected with SARS-CoV-2: optimizing pooled testing 593
    at low prevalence. arXiv. 2020;.   594

26. Torres I, Albert E, Navarro D. Pooling of Nasopharyngeal Swab Specimens for   595
    SARS-CoV-2 detection by RT-PCR. Journal of Medical Virology. 2020; p. 25971. 596
    doi:10.1002/jmv.25971.   597

27. Cabrera JJ, Rey S, Perez S, Martinez-Lamas L, Cores-Calvo O, Torres J, et al. Pooling for SARS-CoV-2 control in care institutions. medRxiv. 2020;doi:10.1101/2020.05.30.20108597.

28. Wacharapluesadee S, Kaewpom T, Ampoot W, Ghai S, Khamhang W, Worachotsueptrakun K, et al. Evaluating the efficiency of specimen pooling for PCR-based detection of COVID-19. Journal of Medical Virology. 2020; p. 0–1. doi:10.1002/jmv.26005.

29. Martin A, Storto A, Andre B, Mallory A, Dangla R, Visseaux B, et al. High-sensitivity COVID-19 group testing by digital PCR. arXiv. 2020;.

30. Khodare A, Padhi A, Gupta E, Agarwal R, Dubey S, Sarin SK. Optimal size of sample pooling for RNA pool testing: an avant-garde for scaling up SARS CoV 2 testing. medRxiv. 2020;doi:10.1101/2020.06.11.20128793.

31. Wernike K, Keller M, Conraths FJ, Mettenleiter TC, Groschup MH, Beer M. Pitfalls in SARS-CoV-2 PCR diagnostics. bioRxiv. 2020; p. 2020.06.03.132357. doi:10.1101/2020.06.03.132357.

32. Gan Y, Du L, Faleti OD, Huang J, Xiao G, Lyu X, et al. Sample Pooling as a Strategy of SARS-COV-2 Nucleic Acid Screening Increases the False-negative Rate. medRxiv. 2020;.

33. Griesemer SB, Van Slyke G, St George K. Assessment of sample pooling for clinical SARS-CoV-2 testing. bioRxiv. 2020;.

34. Hanel R, Thurner S. Boosting test-efficiency by pooled testing strategies for SARS-CoV-2. arXiv. 2020;(1):3–5.

35. Deckert A, Bärnighausen T, Kyei N. Pooled-sample analysis strategies for COVID-19 mass testing: a simulation study. Bull World Health Organ. 2020;(April). doi:10.2471/BLT.20.257188.

36. Sinnott-Armstrong N, Klein D, Hickey B. Evaluation of Group Testing for SARS-CoV-2 RNA. medRxiv. 2020;doi:10.1101/2020.03.27.20043968.

37. Verdun CM, Fuchs T, Harar P, Elbrächter D, Fischer DS, Berner J, et al. Group testing for SARS-CoV-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies. medRxiv. 2020; p. 2020.04.30.20085290. doi:10.1101/2020.04.30.20085290.

38. Narayanan K, Frost I, Heidarzadeh A, Tseng KK, Banerjee S, John J, et al. Pooling RT-PCR or NGS samples has the potential to cost-effectively generate estimates of COVID-19 prevalence in resource limited environments. medRxiv. 2020; p. 2020.04.03.20051995. doi:10.1101/2020.04.03.20051995.

39. Bilder CR, Iwen PC, Abdalhamid B, Tebbs JM, McMahan CS. Tests in short supply? Try group testing. Significance. 2020;17(3):15–16. doi:10.1111/1740-9713.01399.

40. Jones TC, Mühlemann B, Veith T, Biele G, Zuchowski M, Hoffmann J, et al. An analysis of SARS-CoV-2 viral load by patient age. medRxiv. 2020; p. 2020.06.08.20125484. doi:10.1101/2020.06.08.20125484.

41. Thompson KH. Estimation of the Proportion of Vectors in a Natural Population of Insects. Biometrics. 1962;18(4):568. doi:10.2307/2527902.

42. Forootan A, Sjöback R, Björkman J, Sjögreen B, Linz L, Kubista M. Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). Biomolecular Detection and Quantification. 2017;12(April):1–6. doi:10.1016/j.bdq.2017.04.001.

43. Ruiz-Villalba A, van Pelt-Verkuil E, Gunst QD, Ruijter JM, van den Hoff MJ. Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). Biomolecular Detection and Quantification. 2017;14(July):7–18. doi:10.1016/j.bdq.2017.10.001.

44. Jones TC, Mühlemann B, Talitha V, Marta Z, Hofmann J, Stein A, et al. An analysis of SARS-CoV-2 viral load by patient age. Preprint Charité Hospital. 2020;.

45. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: https://www.R-project.org/.

46. Nielsen OE. Information and exponential families : in statistical theory. Chichester U.K. New York: John Wiley & Sons; 2014.

47. Liu Y, Yan LM, Wan L, Xiang TX, Le A, Liu JM, et al. Viral dynamics in mild and severe cases of COVID-19. The Lancet Infectious diseases. 2020;(20).

48. Barrett ES, Horton DB, Roy J, Gennaro ML, Brooks A, Tischfield J, et al. Prevalence of SARS-CoV-2 infection in previously undiagnosed health care workers at the onset of the U.S. COVID-19 epidemic. medRxiv. 2020; p. 2020.04.20.20072470. doi:10.1101/2020.04.20.20072470.

49. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. New England Journal of Medicine. 2020; p. NEJMoa2006100. doi:10.1056/NEJMoa2006100.

50. Meleard S. Modèles aléatoires en Ecologie et Evolution. CMAP; 2016. Available from: http://www.cmap.polytechnique.fr/IMG/pdf/LIVRE07102013.pdf.

51. Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, et al. Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. New England Journal of Medicine. 2020;382(22):2081–2090. doi:10.1056/NEJMoa2008457.

52. Smith DR, Duval A, Pouwels KB, Guillemot D, Fernandes J, Huynh BT, et al. How best to use limited tests? Improving COVID-19 surveillance in long-term care. medRxiv. 2020;.

53. Wyllie AL, Fournier J, Casanovas-Massana A, Campbell M, Tokuyama M, Vijayakumar P, et al. Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs. medRxiv. 2020;(2):2020.04.16.20067835. doi:10.1101/2020.04.16.20067835.

54. Azzi L, Carcano G, Gianfagna F, Grossi P, Gasperina DD, Genoni A, et al. Saliva is a reliable tool to detect SARS-CoV-2. Journal of Infection. 2020;(xxxx):1–6. doi:10.1016/j.jinf.2020.04.005.

55. To KKW, Tsang OTY, Leung WS, Tam AR, Wu TC, Lung DC, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. The Lancet Infectious Diseases. 2020;20(5):565–574. doi:10.1016/S1473-3099(20)30196-1.

56. Williams E, Bond K, Zhang B, Putland M, Williamson DA. Saliva as a non-invasive specimen for detection of SARS-CoV-2. Journal of Clinical Microbiology. 2020;50(April). doi:10.1128/JCM.00776-20.

57. Khurshid Z, Zohaib S, Joshi C, Moin SF, Sohail M, Speicher DJ, et al. Saliva as a non-invasive sample for the detection of SARS-CoV-2 : a systematic review. medRxiv. 2020;.

58. Treibel TA, Manisty C, Burton M, McKnight Á, Lambourne J, Augusto JB, et al. COVID-19: PCR screening of asymptomatic health-care workers at London hospital. The Lancet. 2020;6736(20):19–20. doi:10.1016/S0140-6736(20)31100-4.

59. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science. 2020;5793(April):eabb5793. doi:10.1126/science.abb5793.

Supplementary Information

Vincent Brault, Bastien Mallein, Jean-Francois Rupprecht

A   Ideal tests

We present here some of results obtained from the computations made in Sec. I, where we assumed perfect group testing and used it to measure prevalence in the population. Note that with a perfect test, the question of early detection of an outbreak in a community becomes much simpler : one just need to test everyone at regular time intervals with a single test.

A.1   Number of tests and sample size as function of the population prevalence

We trace here, for various values of the prevalence, the number of tests and total number of samples needed to archive a given precision for the confidence interval.
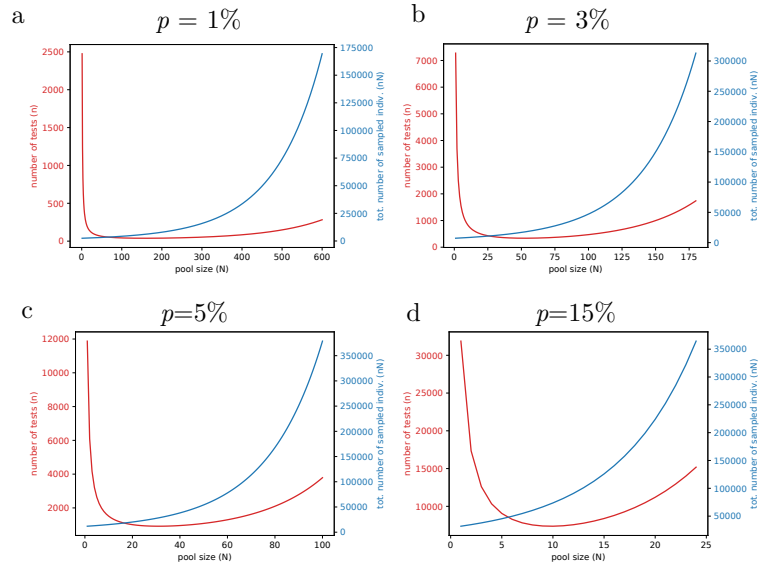

Figure S1. Total number of tests and sampled individuals so that the width of the 95% confidence interval is smaller than 0.4% as a function of the pool size N chosen for a perfect test, for a prevalence p equal to $p = 1\%$ (a), $p = 3\%$ (b), $p = 5\%$ (c), $p = 15\%$ (d).

A.2   Bayesian inference

We are now interested in a Bayesian approach to the measure of prevalence. We started with an initial prior distribution with density $f_0(p) = 6p(1-p)\mathbf{1}_{\{0 \le p \le 1\}}$ for the prevalence, and for each new test j we do the following:

1. take the the mean value $\bar{p}_{j-1} = \int_0^1 p f_{j-1}(p)\mathrm{d}p$ of the prior;

2. choose the size N_j of the pool of the jth test computed as (cf. Eq. (3)):

$$N_j = \left\lfloor -\frac{c_\star}{\log(1 - \bar{p}_{j-1})} \right\rfloor; \tag{S1}$$
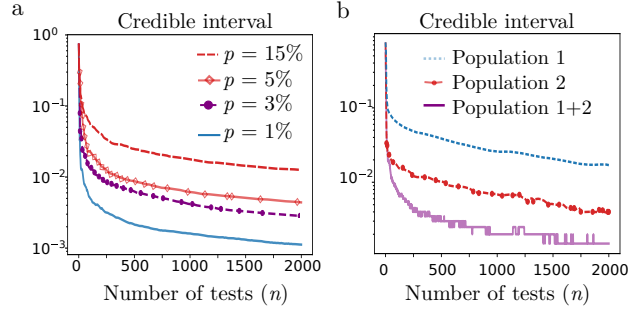
Figure S2. (a) Width of the 95% credible interval on the prevalance p with adaptative Bayesian sampling as a function of the number of tests n for a set of values in the prevalence ranging from $p = 15\%$ (top, magenta dashed line) to $p = 1\%$ (bottom, blue solid line). (b) Width of the credible intervals in a two-category mixed population for the prevalence either: in the general population (magenta solid line); for the less exposed population 1 with a prevalence of 0.5%, representing 80% of the general population (blue dashed line); for the more at-risk population 2 with a prevalence of 5% representing 20% of the general population (red dotted line with circles).

3. choose N_j individuals at random and test them in a group:

   - if the test is positive, then $f_j(p) = C_j^+(1 - (1 - p)^{N_j})f_{j-1}(p)$;

   - if the test is negative, then $f_j(p) = C_j^-(1 - p)^{N_j}f_{j-1}(p)$;

   with $C_j^\pm$ normalizing constants, chosen such that $\int_0^1 f_j(p)\mathrm{d}p = 1$.

We trace in Fig. S2 the result in blue of this experiment, the 95% credible interval being $[a_j, b_j]$, with a_j being the 2.5%th quantile of f_j and b_j its 97.5% quantile.

Simultaneously to this statistical experiment, one can follow the prevalence in sub-populations of interest. For example, if we assume the population consists of two sub-populations 1 and 2 with different prevalences p_1 and p_2. Starting with a prior distribution $f_j(p_1, p_2)\mathrm{d}p_1\mathrm{d}p_2$ for these prevalences, if a group consisting of a individuals of the first sub-population and b individuals of the second population is sampled positive, then Bayes rules gives $C_{j+1}^+(1 - (1 - p_1)^a(1 - p_2)^b)$ for the updated law of (p_1, p_2). A similar update is made if the test is negative. As a result, we get estimates for the prevalence in each sub-population at the same time as we are measuring the prevalence in the overall population.

We test the above statistical experiment on a population which is composed of two sub-populations, one large subpopulation of sparsely exposed individuals (prevalence 0.5%, representing 4/5th of the whole population), and a smaller subpopulation of very exposed individuals (prevalence 5%). At each step, we choose the size of the pool according to the available estimate for the prevalence in the complete population. The composition of the pool in terms of individuals of each sub-population is chosen at random (at the jth step, there are $\mathrm{Ber}(N_j, 0.8)$ individuals of the first sub-population). We also update our estimation of the prevalences (p_1, p_2) in each of the two sub-populations.

The results are traced in Fig. S2 in orange and green curves. One can see that the width of the credibility intervals of the sub-populations decay much slower than for the whole population. The reason is that the size of the groups are optimized to measure as precisely as possible the mean value p.

However, observe that even with a naive group construction (without segregating individuals according to their sub-population), one can extract information on the
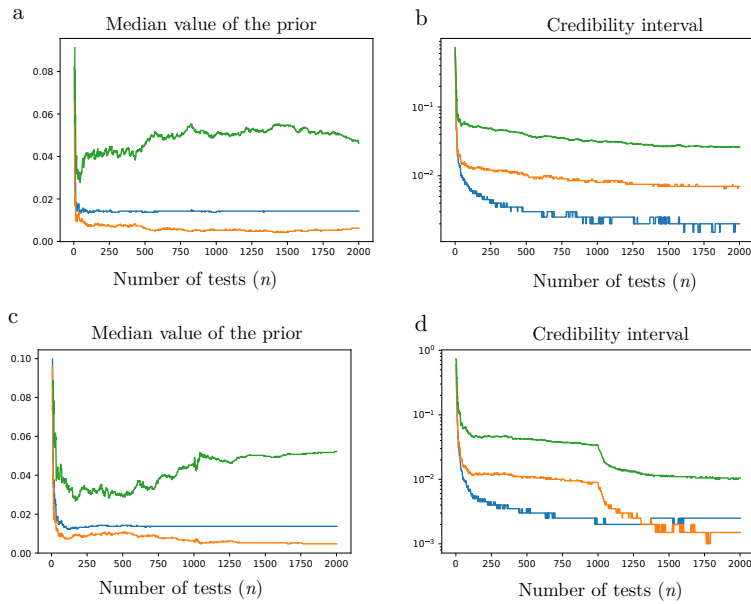
Figure S3. (a-b) Bayesian estimation of the parameters of a mixed population, consisting of 80% individuals of type 1 with a prevalence of 0.5% and 20% individuals of type 2 with a prevalence of 5%. Pooled samples are constituted by sampling randomly individuals from the two sub-populations, with a size optimized for the speed of convergence of the overall prevalence of 1.4%. (a-c) Median value of the priors, overall population in blue, first resp second population in orange resp green. (b-d) Width of the 95% credible intervals. In (c,d), the first 1000 tests are made on groups whose size is optimized to estimate the prevalence of the overall population, the next 1000 tests are divided into two groups that are used on homogeneous sets of the sub-populations, in groups optimized to estimate the prevalences within these sub-populations. This has the effect of drastically improving the speed of convergence of the estimator of the prevalence in the subpopulations.

prevalence of the sub-populations of interest. Therefore, a design for the measure of the $\qquad$ 745
prevalence in a stratified population could be the following: in a first time, pool testing $\qquad$ 746
is implemented on randomly constructed group of individuals from the general $\qquad$ 747
population. Data is then analysed to detect sub-populations with different prevalences $\qquad$ 748
(e.g. according to geography, age, occupation, ...). In a second time, once $\qquad$ 749
sub-populations of interest are identified, pool testing is applied to each of the $\qquad$ 750
sub-populations independently. We implemented this method if Fig. S3, with the same $\qquad$ 751
number of tests a much more detailed estimate of the prevalence is obtained. $\qquad$ 752

B   Censored Gaussians

In this section, we present the simple mixing models of Sec. 3, as well as some complementary graphs and the estimations obtained for the parameters of this models.
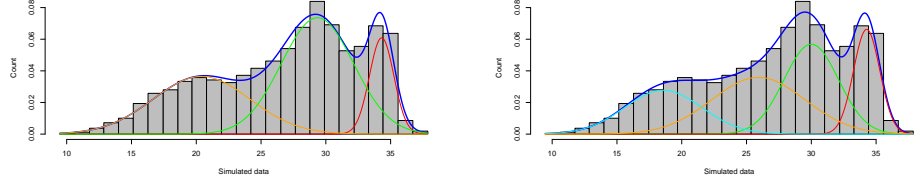
B.1   Naive method based on mixing models

In this section, we trace the density estimated by a simple mixture of Gaussian variable presented in Sec. 3.1. An estimation of the parameters of this mixture are given in Table 3.

Figure S4. Representation of the histogram with the densities estimated with 3 classes (on the left) and 4 classes (on the right): the color lines (other than blue) represent the density of each component and the blue line the density of the mixture.

As we do not have access to raw data, we performed simulations to generate a reconstructed datasets with consistent histograms to Fig. 1 from [44], with randomized position of the points within each class. We applied the above procedure to 100 independently reconstructed data, in order to limit the influence of the random part. Among these 100 simulations, we obtain 95 times 3 clusters and 5 times 4 clusters. When there are 3 clusters, the estimation of the parameters is very stable (standard deviation less than 0.03 for each) but there is a little more variability in the case of 4 clusters in particular for the two classes with the largest averages (but the standard deviation does not exceed 0.25).

760
761
762
763
764
765
766
767
768

B.2   Proof of the Theorem 3.3

To prove the lemma 3.3, we observe that for every $x \in \mathbb{R}$ we have the following decomposition of the density f_X if $q > 0$:

$$f_X(x) = b(\eta) \exp\left[\langle \eta, T(x) \rangle\right], \tag{S2}$$

with $< \cdot, \cdot >$ is the scalar product, $\eta = \left(\frac{\mu}{\sigma^2}, \quad -\frac{1}{2\sigma^2}, \quad \ln q\right)$ the natural parameters,

$$T(x) = \left(x, \quad x^2, \quad \mathbb{1}_{\{x > d_{\mathrm{cens}}\}}\right),$$

the sufficient statistics and

$$b(\eta) = \frac{1}{q + (1-q)F_{\mu,\sigma}(s)} \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right).$$

For the totally censured model, we have the same decomposition with the third parameters and taking $q = 0$. Thanks the decomposition in Eq. (S2), the (partially) censured model belongs to the family of exponential laws and the maximum likelihood estimators are strongly consistent and asymptotically normal.

B.3 Simulations

To study the quality of the estimators defined in Sec. 3.2, we simulated 10^4 samples of size $n \in \{10^2, 10^3, 10^4, 10^5\}$ of variables following the model $\mathcal{CN}_{d_{\text{cens}}}(0, 1, p)$ with $d_{\text{cens}} \in \{-2, -1, 0, 1, 2, 3\}$ and $q \in \{0, 0.1, 0.5, 0.9\}$. We provide boxplots estimations of the parameters in Fig. S5 and a zoom on significant part in Fig. S6. Note that these parameters ($\mu = 0, |d_{\text{cens}}| \leq 3$) are very different from the ones expected for C_t values, but the model can be straightforwardly adapted by an affine transformation to measured parameters of interest.

Observe from Fig. S5 that the estimations are generally close to the parameters but we can sometimes have very large deviations. We find that the more n increases, the better the estimator. The threshold seems to have a weak influence on the estimation of the partially censored model but, for the fully censored model, we see that the more d_{cens} increases and the more the quality of the estimators increases; especially when $d_{\text{cens}} = -2$ which represents approximately the 2.3% quantile. Note that we observe large deviations in the partially censored model when d_{cens} is equal to 2; this may seem counter-intuitive since we have access to around 97.7% of uncensored Gaussian information. However, this leaves few observations for the estimation of p (which we observe on the graphs of the last line) and this weakens in this case the model because censorship no longer really has any reason to be. We therefore recommend using the model only when the number of observations after censorship is sufficient to estimate the parameter p.

B.4 Censored mixture model

In this section, we present the complementary graphs of Sec. 3. The statistical model presented here has the following density defined for all $x \in \mathbb{R}$ by:

$$f(x) = \sum_{k=1}^{3} \pi_k \frac{f_{\mu_k, \sigma_k}(x)}{q_k + (1 - q_k) F_{\mu_k, \sigma_k}(d_{\text{cens}})} \left[1 + (q_k - 1) \mathbb{1}_{\{x > d_{\text{cens}}\}} \right]. \tag{S3}$$

The completely censored mixture model has the same density than the Eq. (S3) with $q_k = 0$.

With the model in Eq. (S3), we can estimate the theoretical false negative rate by the following formula:

$$\mathbb{P}\left(\text{false negative}\right) = \sum_{k=1}^{3} \pi_k \left[1 - F_{\mu_k, \sigma_k}(d_{\text{cens}}) \right] (1 - q_k). \tag{S4}$$
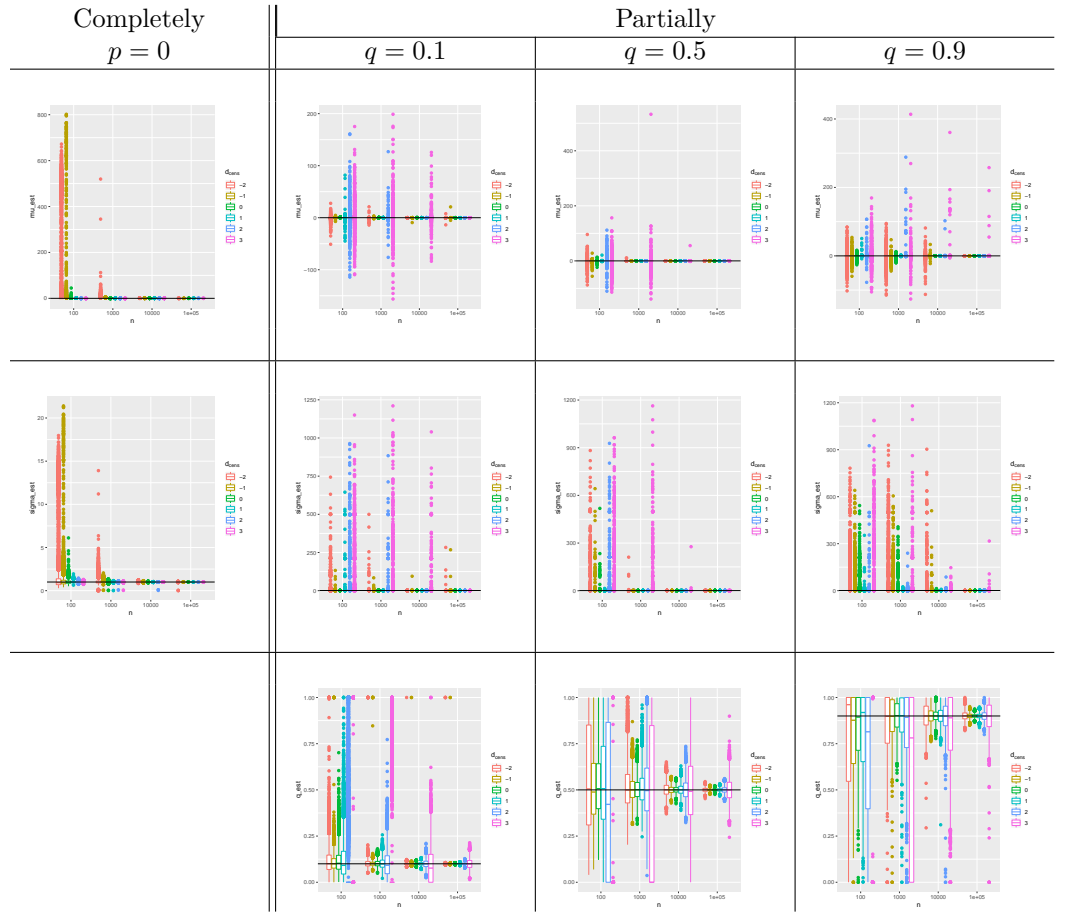
Figure S5. Boxplots of the estimations of μ (first row), σ (second row) and p (last row ; only for partially censored model) in function of model (columns), the size n of sample (x-axis) and the value of the threshold s (color). The true value is symbolised by the horizontal black line.

Table 3. Estimated parameters for the naive Gaussian mixture fit and the censored Gaussian mixture fits defined in Eq. (S3), for the datasets available in [44] and [27]. Note the consistency of the estimations, in particular in the partially and completely censored models.

[44]

Model	$q_{i\,i=1..3}$	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
Naive		20.41	3.74	0.34	29.43	2.81	0.52	34.32	0.89	0.14
Partially	0.2	20.14	3.60	0.32	29.35	2.96	0.53	34.78	1.32	0.14
Completely		20.13	3.60	0.33	29.41	3.02	0.54	34.81	1.31	0.13

[27]

Model	$q_{i\,i=1..3}$	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
Naive		19.75	2.05	0.20	25.61	2.99	0.39	34.28	2.36	0.40
Partially	0.4	20.16	2.19	0.26	26.03	2.58	0.43	34.54	2.66	0.41
Completely		20.55	3.45	0.31	26.33	2.11	0.24	34.41	2.98	0.43
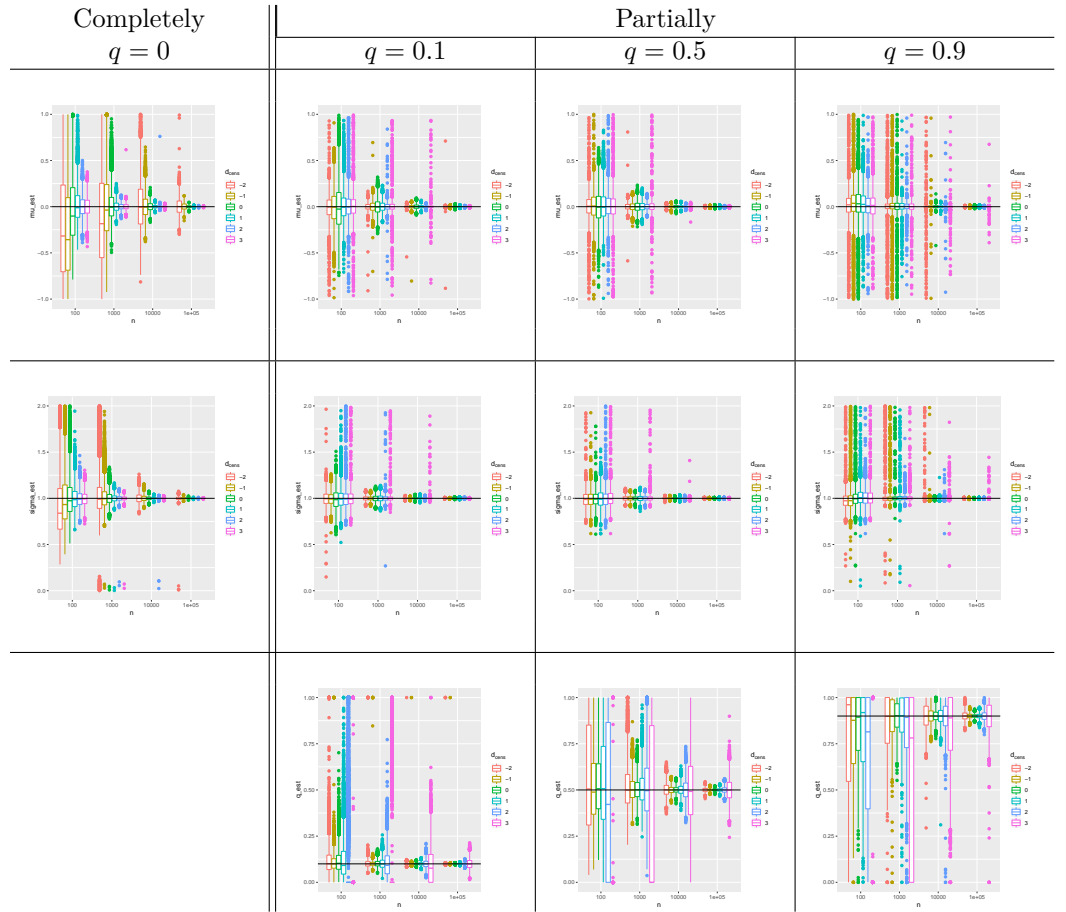
Figure S6. Zoom on boxplots of the estimations of μ (first row), σ (second row) and q (last row ; only for partially censored model) in function of model (columns), the size n of sample (x-axis) and the value of the threshold s (color). The true value is symbolised by the horizontal black line.
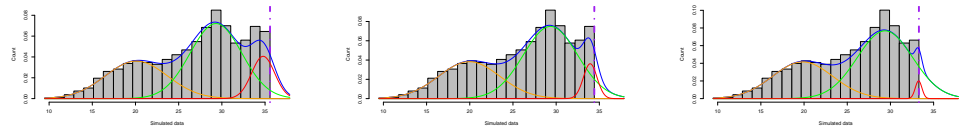
Figure S7. Density of the fits of the censored model with three components (obtained when erasing data to the right of the threshold) with a threshold at 35.6 (left), 34.4 (middle) and 33.2 (right): the orange, green and red lines represent the density of each component and the blue line the density of the mixture. The histogram correspond to the one presented in [44].

Table 4. Estimated parameters for the censored Gaussian mixture fit define in Eq. (S3) for different values of the threshold d_{cens}, applied to reconstructed data data with same distribution as in [44] erased above d_{cens}.

d_{cens}	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
35.6	20.13	3.60	0.33	29.41	3.02	0.54	34.81	1.31	0.13
34.4	20.13	3.61	0.35	29.35	2.99	0.57	34.21	1.03	0.08
33.2	19.97	3.56	0.03	29.40	3.14	0.59	33.21	1.16	0.48

C Estimation of the false-negative risk in the presence of multiple positive individuals in the pool

We treat here the case of a pool of N samples that contains $k > 1$ positive individuals. It is particularly relevant to consider these case when pooling correlated samples, (such has individuals living in the same household, or workers sharing the same area). Indeed, in this case, knowing that one individual is contaminated increases the probability that more individuals in the pool are as well, which should make the detection easier.

For the sake of completeness, here we also consider the risk of defective sampling (e.g. that the swabs fails to collect viral load in an infected individual), which we denote ζ. The probability of having a negative pool result given that there is k positive samples within the pool reads, according to the model presented in Eq. (7):

$$\mathbb{P}\left[-|k+\right] = \sum_{j=1}^{k} \binom{k}{j} \zeta^{k-j} (1-\zeta)^j \mathbb{P}\left[\log_2\left(\sum_{i=1,\ldots,j} C_i/N\right) > d_{\text{cens}}\right]. \quad (S5)$$

Under the two assumptions that:

1. the viral load distribution spans several order of magnitudes (e.g. log-normal distributed), so that, following Eq. (12):

$$\mathbb{P}\left[\log_2\left(\sum_{i=1,\ldots,j} C_i/N\right) > d_{\max}\right] = \mathbb{P}\left[\min_{i=1,\ldots,j}(\log_2(C_i)) > d_{\max}^{(N)}\right], \quad (S6)$$

with $d_{\max}^{(N)} = d_{\text{cens}} - \log_2(N)$.

2. the viral loads (not the infection status) between the k infected individuals are independent, in which case:

$$\mathbb{P}\left[\min_{i=1,\ldots,j}(\log_2(C_i)) > d_{\max}^{(N)}\right] = \mathbb{P}\left[\log_2(C_1) > d_{\max}^{(N)}\right]^j, \quad (S7)$$

we find that Eq. (S5) takes the simple expression:

$$\mathbb{P}\left[-|k+\right] = \left(\zeta + (1-\zeta)(1 - \mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right])\right)^k, \quad (S8)$$

$$= \left(1 - (1-\zeta)\mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right]\right)^k. \quad (S9)$$

In Fig. S8, in the case of correlated samples, we find that the false negative risk in pooling is greatly reduced if there is more than one positive sample in the pool. The origin of such false-negative reduction is the large variability in viral load and the fact that the amplification technique is particularly sensitive to the highest viral load in the sample. Such false-negative reduction is robust to the presence of a finite risk of defective sampling $\zeta = 5\%$.

In addition, one may expect the number of positive k to be distributed according to a binomial distribution with a parameter p corresponding to the prevalence of the disease. Conditioned on the probability that there is at least one individual that is infected within the pool, the conditional probability that $k \geq 1$ is the number of infected individiduals reads

$$\mathbb{P}\left[k+|+\right] = \frac{1}{1-(1-p)^N}\binom{N}{k}p^k(1-p)^{N-k}, \quad (S10)$$
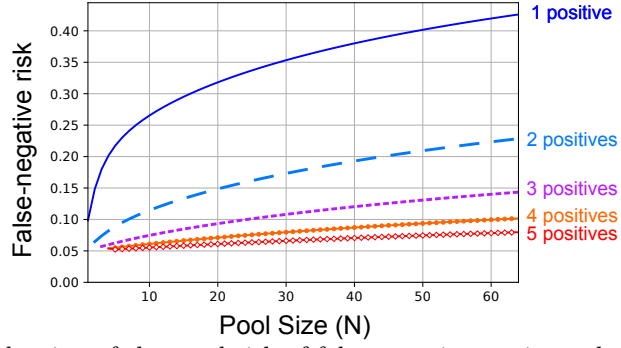
Figure S8. Evaluation of the total risk of false negatives estimated according to Eq. (S5) as a function of the pool size N for several values of the number of positive samples in the pool $k = 1$ (blue solid line); $k = 2$ (cyan dashed line); $k = 3$ (magenta line); $k = 4$ (diamond orange line); $k = 5$ (circle red line). We consider a risk that the sample is defective $\zeta = 0.05$.
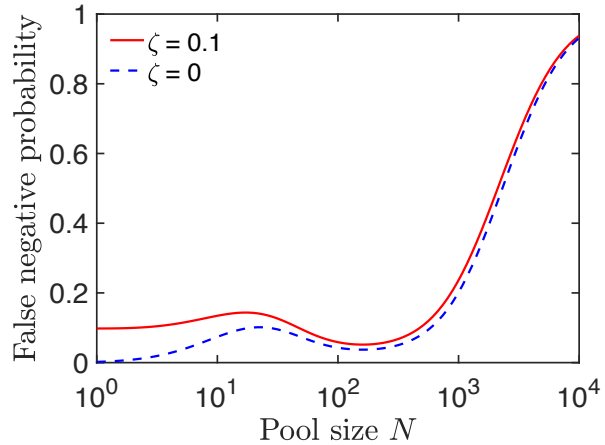
which leads to the following expression for the averaged probability that the pool test turns negative although there is at least one positive individuals in the community

$$\mathbb{P}\left[-|+\right] = \frac{1}{1 - (1-p)^N} \sum_{k=1}^{N} \binom{N}{k} p^k (1-p)^{N-k} \left(\zeta + (1-\zeta)(1 - \mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right]) \right)^k,$$

(S11)

$$= \frac{1}{1 - (1-p)^N} \left\{ \left(p \left(\zeta + (1-\zeta)(1 - \mathbb{P}\left[\log_2(C_1) < d_{\max}^{(N)}\right]) \right) + 1 - p \right)^N - (1-p)^N \right\}.$$

(S12)

As shown in Fig. S9, the averaged false-negative probability risk is not necessarily a monotonous function.

Figure S9. Example of a counter-intuitive evolution in averaged community false negative risk of false negatives, as defined through Eq. (S12), as a function of the pool size N (solid red curve) with a defective sampling probability $\zeta = 0.10$. (dashed blue curve) without defective sampling. $\mu = 27$ and $\sigma = 2$.