Using viral load and epidemic dynamics to optimize pooled testing in resource constrained settings

Authors: Brian Cleary[1,*,†], James A. Hay[2,3,*,†], Brendan Blumenstiel[1], Maegan Harden[1], Michelle Cipicchio[1], Jon Bezney[1], Brooke Simonton[1], David Hong[4], Madikay Senghore[2], Abdul K. Sesay[5], Stacey Gabriel[1], Aviv Regev[6,7,8,9.*], Michael J. Mina[1,2,3,10,*]

Affiliations:

[1] Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA

[2] Centre for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

[3] Department of Immunology and Infectious Diseases, Harvard School of Public Health

[4] Wharton Statistics, University of Pennsylvania, Philadelphia, PA, USA

[5] Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine, PO Box 273, Banjul, The Gambia

[6] Klarman Cell Observatory, Broad Institute of MIT and Harvard

[7] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142 USA

[8] Howard Hughes Medical Institute, Chevy Chase, MD, USA

[9] Current address: Genentech, 1 DNA Way, South San Francisco, CA, USA

[10] Department of Pathology, Brigham and Women's Hospital, Harvard Medical School

[†]These authors contributed equally to this work

[*]Correspondence should be addressed to: bcleary@broadinstitute.org (B.C.), jhay@hsph.harvard.edu (J.H.), aregev@broadinstitute.org (A.R.), mmina@hsph.harvard.edu (M.M.)

Abstract

Extensive virological testing is central to SARS-CoV-2 containment, but many settings face severe limitations on testing. Group testing offers a way to increase throughput by testing pools of combined samples; however, most proposed designs have not yet addressed key concerns over sensitivity loss and implementation feasibility. Here, we combine a mathematical model of epidemic spread and empirically derived viral kinetics for SARS-CoV-2 infections to identify pooling designs that are robust to changes in prevalence, and to ratify losses in sensitivity against the time course of individual infections. Using this framework, we show that prevalence can be accurately estimated across four orders of magnitude using only a few dozen pooled tests without the need for individual identification. We then exhaustively evaluate the ability of different pooling designs to maximize the number of detected infections under various resource constraints, finding that simple pooling can identify up to 20 times as many positives compared to individual testing with a given budget. We illustrate how pooling affects sensitivity and overall detection capacity during an epidemic and on each day post infection, finding that sensitivity loss is mainly attributed to individuals sampled at the end of infection. Crucially, we confirm that our theoretical results can be accurately translated into practice using pooled human nasopharyngeal specimens. Our results show that accounting for variation in sampled viral loads provides a nuanced picture of how pooling affects sensitivity to detect epidemiologically relevant infections. Using simple, practical group testing designs can vastly increase surveillance capabilities in resource-limited settings.

Introduction

The ongoing pandemic of SARS-CoV-2, a novel coronavirus, has caused over 24 million reported cases of coronavirus disease 2019 (COVID-19) and 800,000 reported deaths between December 2019 and August 2020. (1) Although wide-spread virological testing is essential to inform disease status and where outbreak mitigation measures should be targeted or lifted, sufficient testing of populations with meaningful coverage has proven difficult. (2–7) Disruptions in the global supply chains for testing reagents and supplies, as well as on-the-ground limitations in testing throughput and financial support, restrict the usefulness of testing– both for identifying infected individuals and to measure community prevalence and epidemic trajectory. While these issues have been at the fore in even the highest-income countries, the situation is even more dire in low income regions of the world. Cost barriers alone mean it is often simply not practical to prioritize community testing in any useful way, with the limited testing that exists necessarily reserved for the healthcare setting. These limitations urge new, more efficient, approaches to testing to be developed and adopted both for individual diagnostics and to enable public health epidemic control and containment efforts.

Group or pooled testing offers a way to increase efficiency by combining samples into a group or pool and testing a small number of pools rather than all samples individually. (8–10) For classifying individual samples, including for diagnostic testing, the principle is simple: if a pool tests negative, then all of the constituent samples are assumed negative. If a pool tests positive, then the constituent samples are putatively positive and must be tested again individually (**Fig. 1A**). Further efficiency gains are possible through combinatorial pooling, where, instead of testing every sample in every positive pool, each sample can instead be represented across multiple pools and potential positives are identified based on the pattern of pooled results (**Fig. 1B**). (9,10)

Simple pooling designs can also be used to assess prevalence without individual specimen identification (**Fig. 1C**). It has already been shown that the frequency of positive pools can allow estimation of the overall prevalence. (11) Crucially however, we show here that prevalence estimates can be greatly honed by considering quantitative viral loads measured in each positive pool, rather than simply using binary results (positive / negative). In short, the viral (RNA) load measurement from a pool is proportional to the sum of the (diluted) viral loads from each positive sample in the pool. Thus, here we show how evaluating the viral loads greatly improves potential efficiency gains in prevalence estimates by providing crucial information on the estimated number of positive samples in the pool – when the expected distribution of viral loads across specimens is known, which is easily measured empirically in a given lab. (12,13) Although this approach requires more complex statistical methods, the efficiency gains for public health surveillance can be large, and simplifying templates can be produced to improve ease of use and access to these types of analyses. The outcome is a highly efficient method for estimating population prevalence and enabling robust public health surveillance where it was previously out of reach.

Whilst the literature on theoretically optimized pooling designs for COVID-19 testing has grown rapidly, formal incorporation of biological variation (*i.e.,* viral loads) and incorporation of general position along the epidemic curve, has received little attention. (14–17) Test sensitivity for example is not a fixed value, but depends on viral load, which can vary by many orders of magnitudes across individuals and over the course of an infection. (18–20) This large variation within a single infection affects sensitivity to detect infections at different points in the disease course, which has implications for appropriate intervention and the interpretation of a viral load measurement from a sample pool.

Here, we comprehensively evaluate designs for pooled testing of SARS-CoV-2 whilst accounting for epidemic dynamics and variation in viral loads arising from viral kinetics and extraneous features such as sampling variation. We demonstrate efficient, logistically feasible

pooling designs for individual identification (*i.e.,* diagnostics) and prevalence estimation (*i.e.,* population surveillance). To do this, we use realistic simulated viral load data at the individual level over time, representing the entire time course of an epidemic to generate synthetic data that reflects the true distribution of viral loads in the population at any given time of the epidemic. We then used these data to derive optimal pooling strategies for different use cases and resource constraints *in-silico*. Finally, we demonstrate the approach using discarded de-identified human nasopharyngeal swabs initially collected for diagnostic and surveillance purposes.

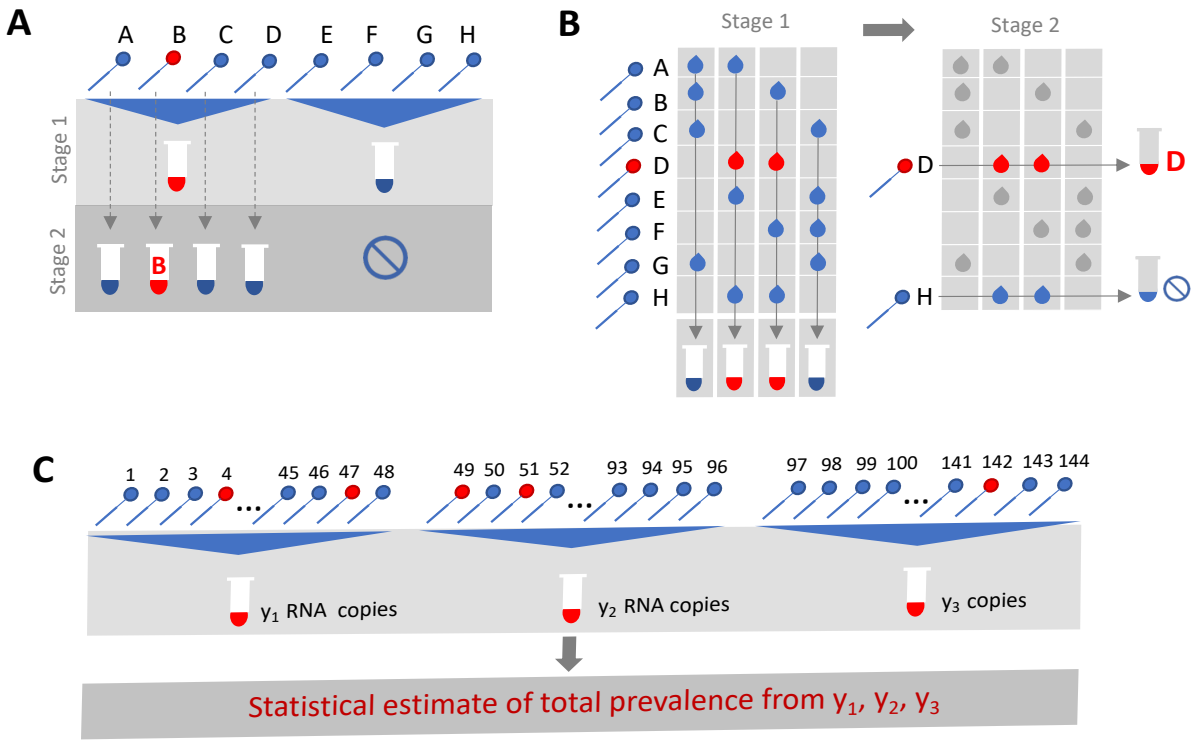


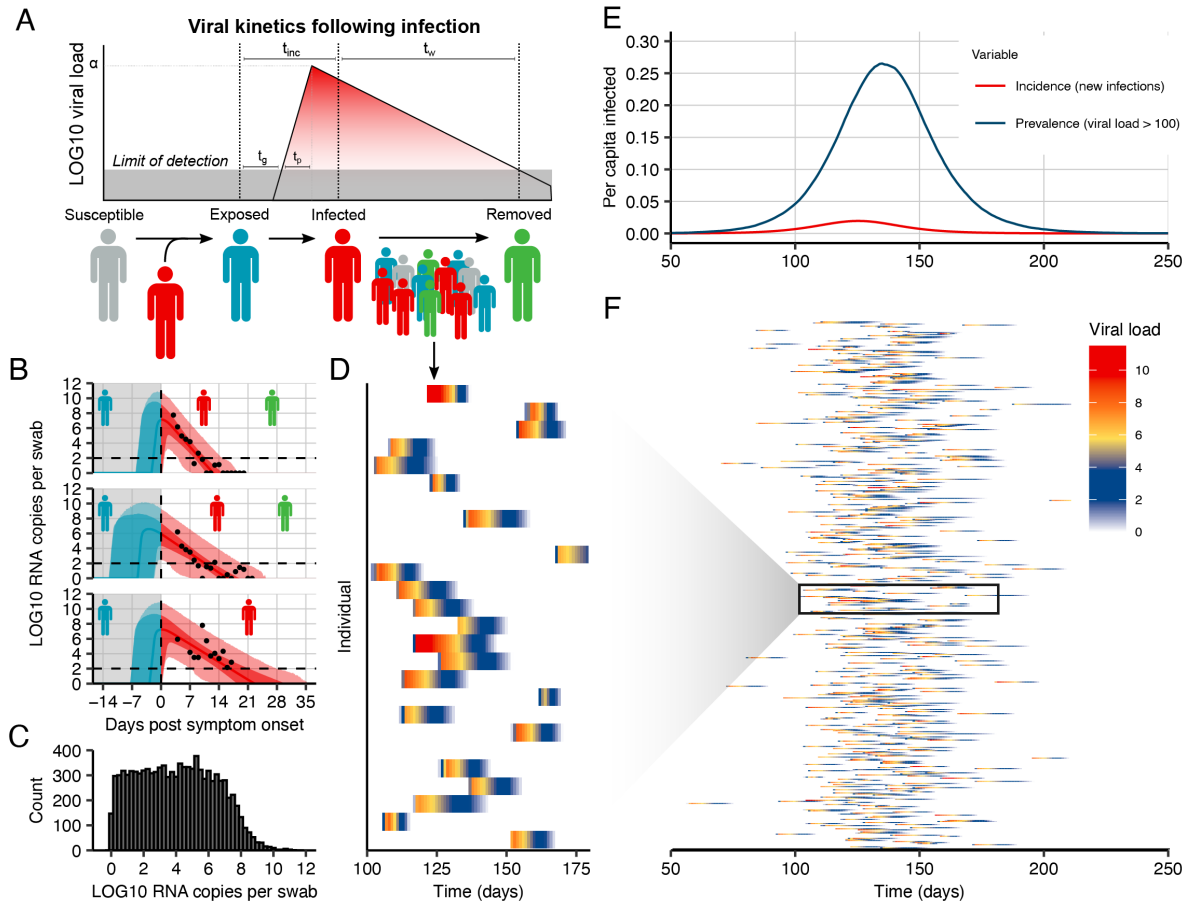
Fig. 1. Group testing designs for sample identification or prevalence estimation. In group testing, multiple samples are pooled and tests are run on one or more pools. The results of these tests can be used for identification of positive samples **(A, B)** or to estimate prevalence **(C)**. **(A)** In the simplest design for sample identification, samples are partitioned into non-overlapping pools. In stage 1 of testing, a negative result (Pool 2) indicates each sample in that pool was negative, while a positive result (Pool 1) indicates at least one sample in the pool was positive. These putatively positive samples are subsequently individually tested in stage 2 to identify positive results. **(B)** In a combinatorial design, samples are included in multiple pools as shown in stage 1. All samples that were included in negative pools are identified as negative, and the remaining putatively positive samples that were not included in any negative test are tested individually in stage 2. **(C)** In prevalence estimation, samples are partitioned into pools. The pool measurement will depend on the number and viral load of positive samples, and the dilution factor. The (quantitative) results from each pool can be used

115 to estimate the fraction of samples that would have tested positive, had they been tested
116 individually.

Results

Modelling a synthetic population to assess pooling designs

To identify optimal pooling strategies for distinct scenarios, we required realistic estimates of viral loads across epidemic trajectories. We developed a population-level mathematical model of SARS-CoV-2 transmission that incorporates empirically measured within-host virus kinetics, and used these simulations to generate population-level viral load distributions representing real data sampled from population surveillance, either using nasopharyngeal swab or sputum samples (**Fig. 2**). Full details are provided in **Materials and Methods** and **Supplementary Material 1, sections 1-4**. These simulations generated a synthetic, realistic epidemic with a peak daily per incidence of 19.5 per 1000 people, and peak daily prevalence of RNA positivity (viral load greater than 100 virus RNA copies per ml) of 265 per 1000 (**Fig. 2E**). We used these simulation data to evaluate optimal group testing strategies at different points along the epidemic curve for diagnostic as well as public health surveillance, where the true viral loads in the population is known fully.



131

132 **Fig. 2: Viral kinetics model fits, simulated infection dynamics and population-wide viral**
133 **load kinetics. (A)** Schematic of the viral kinetics and infection model. Individuals begin
134 susceptible with no viral load, acquire the virus from another infectious individual (exposed),
135 experience an increase in viral load and possibly develop symptoms (infected), and finally
136 either recover following viral waning or die (removed). This process is simulated for many
137 individuals. **(B)** Model fits to time-varying viral loads in swab samples. The black dots show
138 observed *log10* RNA copies per swab; solid lines show posterior median estimates; dark
139 shaded regions show 95% credible intervals (CI) on model-predicted latent viral loads; light
140 shaded regions show 95% CI on simulated viral loads with added observation noise. The blue
141 region shows viral loads before symptom onset and red region shows time after symptom
142 onset. The horizontal dashed line shows the limit of detection. **(C)** Distribution of positive viral
143 loads from 10,000 individuals sampled at day 140. **(D)** 25 simulated viral loads over time. The

heatmap shows the viral load in each individual over time. **(E)** Simulated infection incidence and prevalence of virologically positive individuals from the SEIR model. Incidence was defined as the number of new infections per day divided by the population size. Prevalence was defined as the number of individuals with viral load > 100 (*log10* viral load > 2) in the population divided by the population size on a given day. **(F)** As in **(D)**, but for 500 individuals. The distribution of viral loads reflects the increase and subsequent decline of prevalence. We simulated from inferred distributions for the viral load parameters, thereby propagating substantial individual-level variability through the simulations.

Improved testing efficiency for estimating prevalence

We developed a statistical method to estimate prevalence of SARS-CoV-2 based on cycle threshold (Ct) values measured from pooled samples (**Materials and Methods**), potentially using far fewer tests than would be required to assess prevalence based on number of positive samples identified. We used our synthetic viral load data to assess inferential accuracy under a range of sample availabilities and pooling designs. Across the spectrum of simulated pools and tests we found that simple pooling allows accurate estimates of prevalence across at least four orders of magnitude, ranging from 0.02% to 20%, with up to 400-times efficiency gains (*i.e.,* 400 times fewer tests) than would be needed without pooling (**Fig. 3**). For example, in a population prevalence study that collects ~2,000 samples, we accurately estimated infection prevalences as low as 0.05% by using only 24 total qPCR tests (*i.e.*, 24 pools of 96 samples each; **Fig. 3A**; **Fig. S1**). Importantly, because the distribution of Ct values may differ depending on the sample type (sputum vs. swab), the instrument, and the phase of the epidemic (growth vs. decline, **Fig. S2**), in practice, the method should be calibrated to viral load data (*i.e.,* Ct values) specific to the laboratory and instrument (which can differ from one laboratory to the next) and the population under investigation.

Estimation error arises in two stages: sample collection effects, and as part of the inference method (**Fig. 3B**). Error from sampling collection became less important with increasing numbers of positive samples, which occurred with increasing population prevalence or by increasing the total number of tested samples (**Fig. 3B**; **Fig. S2**). At very low prevalence, small sample sizes (*N*) risk missing positives altogether or becoming biased by false positives. We found that accuracy in prevalence estimation is greatest when population prevalence is greater than 1/N and that when this condition was met, partitioning samples into more pools always improves accuracy (**Fig. S2**). In summary, very accurate estimates of

177 prevalence can be attained using only a small fraction of the tests that would be needed in

178 the absence of pooling.

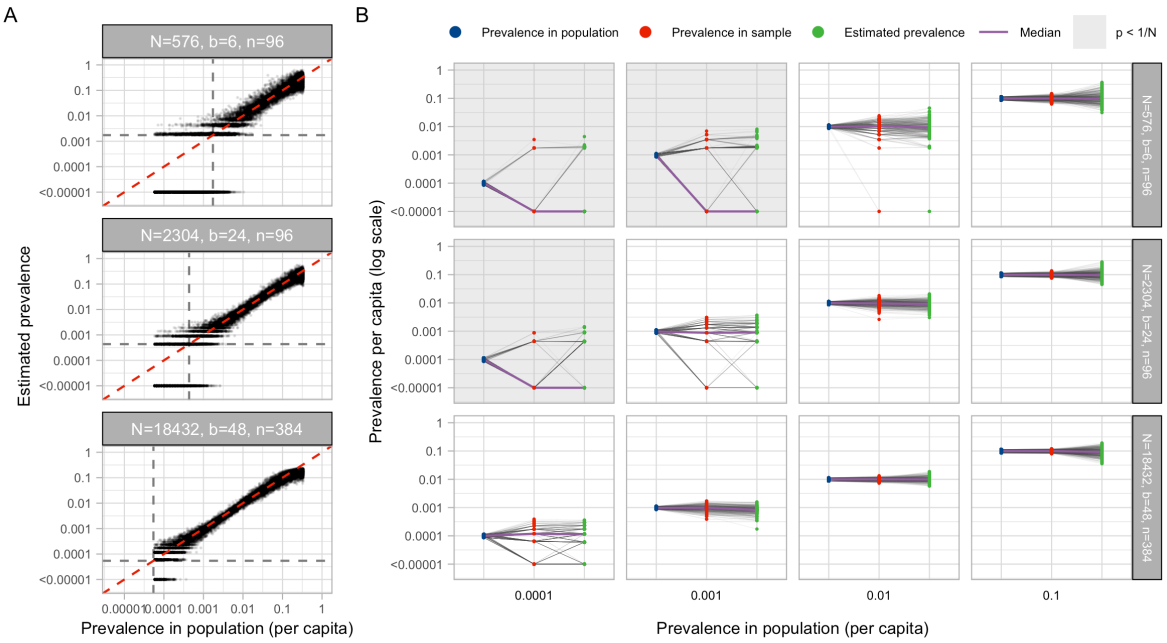


Fig. 3: Estimating prevalence from a small number of pooled tests. In prevalence estimation, a total of N individuals are sampled and partitioned into *b* pools (with *n=N/b* samples per pool). The true prevalence in the entire population (x-axis in A) varies over time with epidemic spread. Population prevalences shown here are during the epidemic growth phase. **(A)** Estimated prevalence against true population prevalence using 100 independent trials sampling *N* individuals at each day of the epidemic. Each facet shows a different pooling design (more pooling designs shown in **Fig. S1**). Dashed grey lines show one divided by the sample size, *N*. **(B)** For a given true prevalence (x-axis, blue points), estimation error is introduced both through binomial sampling of positive samples (red points) and inference on the sampled viral loads (green points). Sampling variation is a bigger contributor at low prevalence and low sample sizes. When prevalence is less than one divided by *N* (grey boxes), inference is less accurate due to the high probability of sampling only negative individuals or inclusion of false positives.

Pooled testing for individual identification

We next analyzed effectiveness of group testing for identifying individual sample results at different points along the epidemic curve with the aim of identifying simple, efficient pooling strategies that are robust to a range of prevalences (**Fig. 1A&B**). Using the simulated viral load data described in **Materials and Methods**, we evaluated a large array of pooling designs *in silico* (**Table S1**). Based on our models of viral kinetics and given a PCR limit of detection of 100 viral copies per ml, we first estimated a baseline sensitivity of conventional (non-pooled) PCR testing of 85% during the epidemic growth phase (*i.e.*, 15% of the time we sample an infected individual with a viral load greater than 1 but below the LOD of 100 viral copies per mL, **Fig. 4A**), which largely agrees with reported estimates. (21,22) This reflects sampling during the latent period of the virus (after infection but prior to significant viral growth) or in the relatively long duration of low viral titers during viral clearance.

Sensitivity of pooled tests, relative to individual testing, is affected by the dilution factor of pooling and by the population prevalence – with lower prevalence resulting in generally lower sensitivity as positives are diluted into many negatives (**Fig. 4A**). The decrease in sensitivity is roughly linear with the log of the dilution factor employed, which largely depends on the number and size of the pools and, for combinatorial pooling, the number of pools that each sample is placed into (**Fig. S3A-C**).

There is a less intuitive relationship between sensitivity and prevalence as it changes over the course of the epidemic. Early in an epidemic there is an initial dip in sensitivity for both individual and pooled testing (**Fig. 4A**). Early during exponential growth of an outbreak, a random sample of infected individuals will be sampled closer to their peak viral load, while later on there is an increasing mixture of newly infected with individuals with lower viral loads at the tail end of their infection. We found that this means at peak prevalence, sensitivity of pooled testing increases as samples with lower viral loads, which would otherwise be missed

due to dilution, are more likely to be 'rescued' by coexisting in the same pool with high viral load samples and thus get individually retested (at their undiluted concentration) during the validation stage. During epidemic decline, fewer new infections arise over time and therefore a randomly selected infected individual is more likely to be sampled during the recovery phase of their infection, when viral loads are lower (**Fig. S4D**). Overall sensitivity is therefore lower during epidemic decline, as more infected individuals have viral loads below the limit of detection; during epidemic growth (up to day 108), overall sensitivity of RT-PCR for individual testing is 85%, whereas during epidemic decline (from day 168 onward) it is 60% (**Fig. S5A**). Sensitivity of RT-PCR for individual testing was ~75% across the whole epidemic. We note that in practice, sensitivity is likely higher than estimated here, because individuals are not sampled entirely at random. Together, these results describe how sensitivity is affected by the combination of epidemic dynamics, viral kinetics, and pooling design when individuals are sampled randomly from the population.

We find that on average the majority of false negatives arise from individuals sampled seven days or more after their peak viral loads, or around seven days after what is normally considered symptom onset (~75% in swab samples during epidemic growth; ~96% in swab samples during epidemic decline; ~68% in sputum during epidemic growth). Importantly, only ~3% of false negative swab samples arose from individuals tested during the first week following peak viral load during epidemic growth, and only ~1% during epidemic decline – (peak titers usually coincide with symptom onset) – and thus most false negatives are from individuals with the least risk of onward transmission (**Fig. S3D&E**).

As mentioned above, the lower sensitivity of pooled testing is counterbalanced by gains in efficiency. When prevalence is low, efficiency is roughly the number of samples divided by the number of pools, since there are rarely putative positives to test individually. However, the number of validation tests required will increase as prevalence increases, and designs that are initially more efficient will lose efficiency (**Fig. 4B**). In general, we find that at very

low population prevalence the use of fewer pools each with larger numbers of specimens offers relative efficiency gains compared to larger numbers of pools, as the majority of pools will test negative. However, as prevalence increases, testing a greater number of smaller pools pays off as more validations will be performed on fewer samples overall (**Fig. 4B**). For combinatorial designs with a given number of total samples and pools, splitting each sample across fewer pools results in a modest efficiency gains (dashed versus solid lines in **Fig. 4B**).

To address realistic resource constraints, we integrated our analyses of sensitivity and efficiency with limits on daily sample collection and testing capacity to maximize the number of positive individuals identified (see **Materials and Methods**). We analyzed the total number of samples screened and the fold increase in the number of positive samples identified relative to individual testing for a wide array of pooling designs evaluated over a period of 50 days during epidemic spread (days 40-90 where point prevalence reaches ~2.5%; **Fig. 4C&D**). Because prevalence changes over time, the number of validation tests may vary each day despite constant pooling strategies. Thus, tests saved on days requiring fewer validation tests can be stored for days where more validation tests are required.

Across all resource constraints considered, we found that effectiveness ranged from one (when testing every sample individually is optimal) to 20 (*i.e.*, identifying 20x more positive samples on a daily basis compared with individual testing within the same budget; **Fig. 4D**). As expected, when capacity to collect samples exceeds capacity to test, group testing becomes increasingly effective. Simple pooling designs are most effective when samples are in slight excess of testing capacity (2-8x), whereas we find that increasingly complex combinatorial designs become the most effective when the number of samples to be tested greatly exceeds testing capacity. Additionally, when prevalence is higher (*i.e.,* sample prevalence from 1.03% to 9.90%), the optimal pooling designs shift towards combinatorial pooling, and the overall effectiveness decreases – but still remains up to 4x more effective

than individual testing (**Fig. S6**). Our results were qualitatively unchanged when evaluating the effectiveness of pooling sputum samples, and the optimal pooling designs under each set of sample constraints were either the same or very similar (**Fig. S7**). Furthermore, we evaluated the same strategies during a 50-day window of epidemic decline (days 190-250) and found that similar pooling strategies were optimally effective, despite lower overall sensitivity as described above (**Fig. S5**).

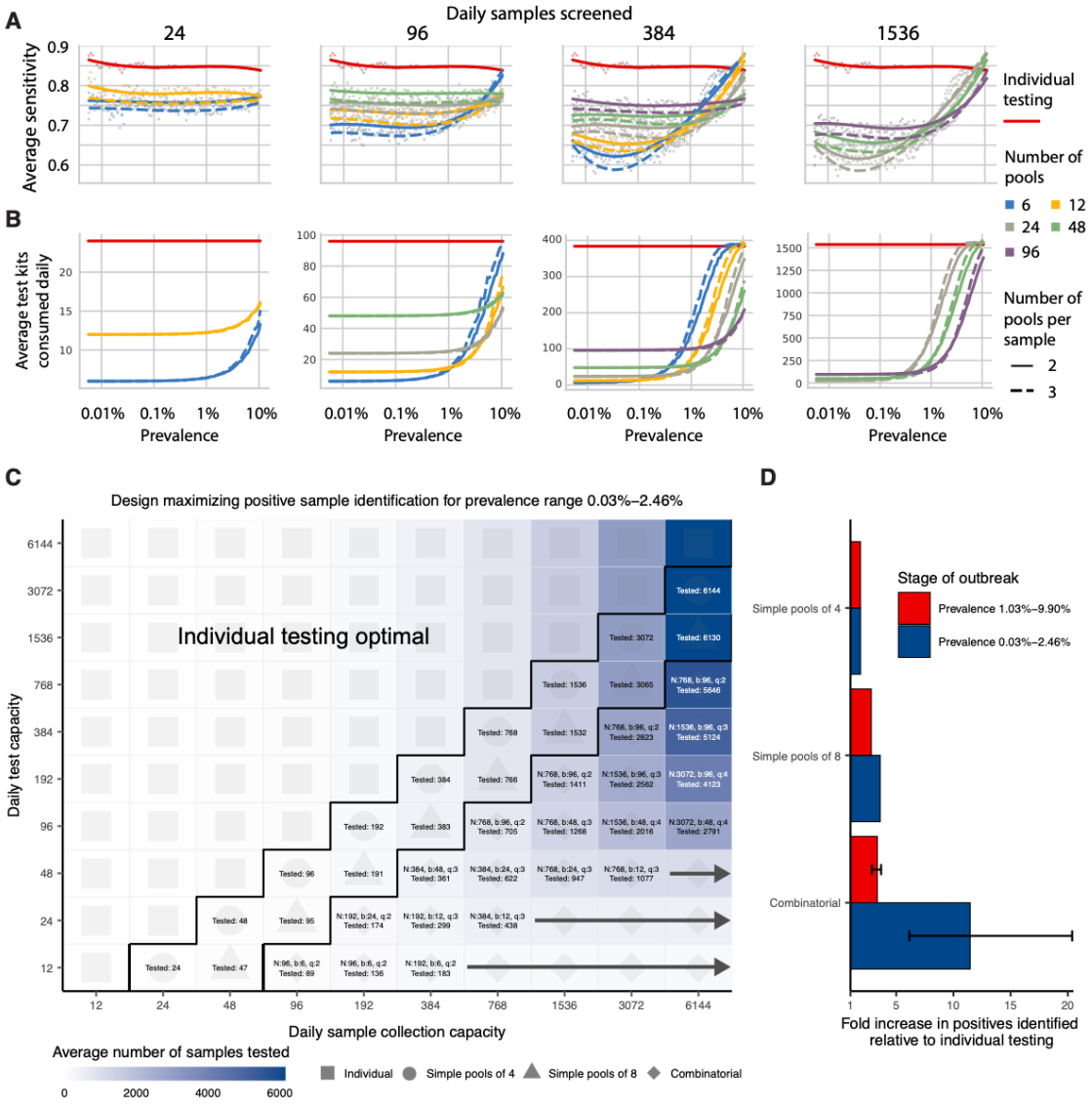


Fig. 4: Group testing for sample identification. We evaluated a variety of group testing designs for sample identification (**Table S1**) on the basis of sensitivity **(A)**, efficiency **(B)**, total number of positive samples identified **(C)** and the fold increase in positive samples identified relative to individual testing **(D)**. (**A** and **B**) The average sensitivity (**A**, y-axis, individual points and spline) and average number of tests needed to identify individual positive samples (**B**, y-axis) using different pooling designs (individual lines) were measured over days 20-110 in our simulated population, with results plotted against prevalence (x-axis, log-scale). Results show the average of 200,000 trials, with individuals selected at random on each day in each trial.

Pooling designs are separated by the number of samples tested on a daily basis (individual panels); the number of pools (color); and the number of pools into which each sample is split (dashed versus solid line). Solid red line indicates results for individual testing. **(C)** Every design was evaluated under constraints on the maximum number of samples collected (columns) and average number of reactions that can be run on a daily basis (rows) over days 40-90. Text in each box indicates the optimal design for a given set of constraints (number of samples per batch (N), number of pools (b), number of pools into which each sample is split (q), average number of total samples screened per day). Color indicates the average number of samples screened on a daily basis using the optimal design. Arrows indicate that the same pooling design is optimal at higher sample collection capacities due to testing constraints. **(D)** Fold increase in the number of positive samples identified relative to individual testing with the same resource constraints. Error bar shows range amongst optimal designs.

Pilot and validation experiments

We validated our pooling strategies using anonymized clinical nasopharyngeal swab specimens. To evaluate simple pooling across a range of inputs, we diluted 5 nasopharyngeal clinical swab samples with viral loads of 89000, 12300, 1280, 140 and 11 viral copies per ml, respectively, into 23 negative nasopharyngeal swab samples (pools of 24). Further details are provided in **Materials and Methods,** and **Supplementary Material 1, sections 7&8**. The results matched the simulated sampling results: the first three pools were all positive, the fourth was inconclusive (negative on N1, positive on N2), and the remaining pool was negative (**Fig. 5A**, **Table S2**). These results are as expected because the EUA approved assay used has a limit of detection of ~100 virus copies per ml, such that the last two specimens fall below the limit of detection given a dilution factor of 24 (i.e. 0.46 and 5.8 virus copies per ml once pooled).

We next tested combinatorial pooling, first using only a modest pooling design. We split 48 samples, including 1 positive, into 6 pools with each sample spread across three different pools. The method correctly identified the three pools containing the positive specimen (**Fig. 5B**, **Table S2**). One negative sample was included in the same 3 pools as the positive sample; thus, 8 total tests (6 pools + 2 validations) were needed to accurately identify the status of all 48 samples, a 6x efficiency gain, which matched our expectations from the simulations.

We next performed two larger validation studies (**Materials and Methods,** and **Supplementary Material 1, section 8)**. To validate combinatorial pooling, we used anonymized samples representing 930 negative and 30 distinct positive specimens (3.1% prevalence), split across 10 batches of 96 specimens each (**Table S3**). For each batch of 96, we split the specimens into 6 pools and each specimen was spread across 2 pools (**Fig. 5C**, **Table S4**). For this combinatorial pooling design and prevalence, our simulations suggest

that we would expect to identify 26 out of 30 known positives (87%) and would see a 2.81x efficiency gain – using only 35% of the number of tests compared to no pooling. We identified 24 of the 30 known positives (80%) and, indeed, required 35% fewer tests (341 vs 960, a 2.8x efficiency gain).

To further validate our methods for prevalence estimation, we created a large study representing 2,304 samples with a (true) positive prevalence of 1%. We aimed to determine how well our methods would work to estimate the true prevalence using $1/48^{th}$ the number of tests compared to testing samples individually. To do this, we randomly assigned 24 distinct positive samples into 48 pools, with each pool containing 48 samples (**Table S3**; to create the full set of pools, we treated some known negatives as distinct samples across separate pools). By using the measured viral loads detected in each of the pools, our methods estimated a prevalence of 0.87% (compared to the true prevalence of 1%) with a bootstrapped 95% confidence interval of 0.52% - 1.37% (**Fig. 5D**), and did so using 48x fewer tests than without pooling. This level of accuracy is in line with our expectations from our simulations. Notably, the inference algorithm applied to these data used viral load distributions calibrated from our simulated epidemic, which in turn had viral kinetics calibrated to samples collected and tested on another continent, demonstrating robustness of the training procedure.

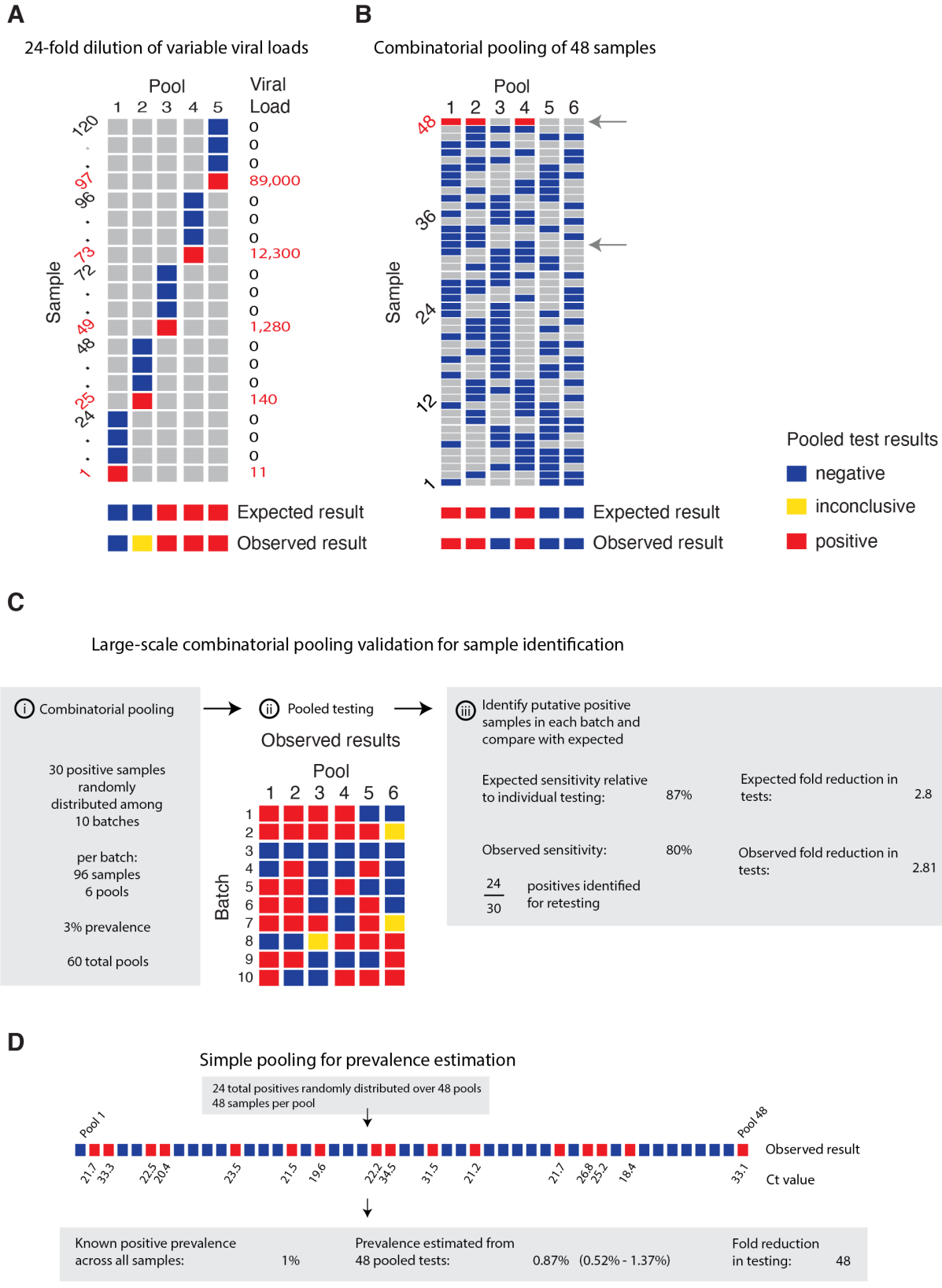


Fig. 5: Validation of simple and combinatorial pooling. Pooled testing of samples was validated experimentally. **(A)** Five pools (columns of matrix), each consisting of 24 nasopharyngeal

swab samples (rows of matrix; 23 negative samples per pool and 1 positive, with viral load indicated in red on right) were tested by viral extraction and RT+qPCR. Pooled results indicated as: negative (blue), inconclusive (yellow), or positive (red). **(B)** Six combinatorial pools (columns) of 48 samples (rows; 47 negative and 1 positive with viral load of 12,300) were tested as above. Pools 1, 2, and 4 tested positive. Arrows indicate two samples that were in pools 1, 2, and 4: sample 32 (negative), and sample 48 (positive). **(C)** Previously tested de-identified samples were pooled using a combinatorial design with 96 samples, 6 pools, and 2 pools per sample. 30 positive samples were randomly distributed across 10 batches of the design. Viral RNA extraction and RT-qPCR were performed on each pool, with the results used to identify potentially positive samples. **(D)** Samples were pooled according to a simple design (48 pools with 48 samples per pool). 24 positive samples were randomly distributed among the pools (establishing a 1% prevalence). The pooled test results were used with an MLE procedure to estimate prevalence (0.87%), and bootstrapping was used to estimate a 95% confidence interval (0.52% - 1.37%).

Discussion

Our results show that group testing for SARS-CoV-2 can be a highly effective tool to increase surveillance coverage and capacity when resources are constrained. For prevalence testing, we find that fewer than 40 tests can be used to accurately infer prevalences across four orders of magnitude, providing large savings on tests required. For individual identification, we determined an array of designs that optimize the rate at which infected individuals are identified under constraints on sample collection and daily test capacities. These results provide pooling designs that maximize the number of positive individuals identified on a daily basis, while accounting for epidemic dynamics, viral kinetics, viral loads measured from nasopharyngeal swabs or sputum, and practical considerations of laboratory capacity.

While our experiments suggest that pooling designs may be beneficial for SARS-CoV-2 surveillance and identification of individual specimens, there are substantial logistical challenges to implementing theoretically optimized pooling designs. Large-scale testing without the use of pooling already requires managing thousands of specimens per day, largely in series. Pooling adds complexity because samples must be tracked across multiple pools and stored for potential re-testing. These complexities can be overcome with proper tracking software (including simple spreadsheets) and standard operating procedures in place before pooling begins. Such procedures can mitigate the risk of handling error or specimen mix-up. In addition, expecting laboratories to regularly adapt their workflow and optimize pool sizes based on prevalence may not be feasible in some settings. (8,16) A potential solution is to follow a simple, fixed protocol that is robust to a range of prevalences. **Supplementary Material 2** provides an example spreadsheet guiding a technician receiving 96 labeled samples to create 6 pools, enter the result of each pool and be provided a list of putative positives to be retested.

Enhancing efficiency at the expense of sensitivity must be considered depending on the purpose of testing. For prevalence testing, accurate estimates can be obtained using very few tests if individual identification is not the aim. For individual testing, although identifying all positive samples that are tested is of course the objective, increasing the number of specimens tested when sacrificing sensitivity may be a crucially important tradeoff. This tradeoff is particularly pertinent because the specimens most likely to be lost due to dilution are those samples with the lowest viral loads already near the limit of detection (**Fig. S3D&E**). Although there is a chance that the low viral load samples missed are on the upswing of an infection – when identifying the individual would be maximally beneficial – the asymmetric course of viral titers over the full duration of positivity means that most false negatives would arise from failure to detect late-stage, low-titer individuals who are less likely to be infectious. (20) Optimal strategies and expectations of sensitivity should also be considered alongside the phase of the epidemic and how samples are collected, as this will dictate the distribution of sampled viral loads. For example, if individuals are under a regular testing regimen or are tested due to recent exposure or symptom onset, then viral loads at the time of sampling will typically be higher, leading to higher sensitivity in spite of dilution effects.

Testing throughput and staffing resources should also be considered. If a testing facility can only run a limited number of tests per day, it may be preferable to process more samples at a slight cost to sensitivity. Back-logs of individual testing can result in substantial delays in returning individual test results, which can ultimately defeat the purpose of identifying individuals for isolation - potentially further justifying some sensitivity losses. (23,24) Choosing a pooling strategy will therefore depend on target population and availability of resources. For testing in the community or in existing sentinel surveillance populations (*e.g.,* antenatal clinics), point prevalence is likely to be low (<0.1 - 3%), which may favor strategies with fewer pools. (6,25–27) Conversely, secondary attack rates in contacts of index cases may vary from

<1% to 17% depending on the setting (*e.g.*, casual vs. household contacts), (28–30) and may be even higher in some instances, (31,32) favoring more pools. These high point prevalence sub-populations may represent less efficient use cases for pooled testing, as our results suggest that pooling for individual identification is inefficient once prevalence reaches 10%. However, group testing may still be useful if testing capacity is severely limited – for example, samples from all members of a household could be tested as a pool and quarantined if tested positive, enabling faster turnaround than testing individuals. This approach may be even more efficient if samples can be pooled at the point of collection, requiring no change to laboratory protocols.

Our modelling results have a number of limitations and may be updated as more data become available. First, our simulation results depend on the generalizability of the simulated Ct values, which were based on viral load data from symptomatic patients. Although some features of viral trajectories, such as viral waning, differ between symptomatic and asymptomatic individuals, population-wide data suggest that the range of Ct values do not differ based on symptom status. (20,33) Furthermore, we have assumed a simple hinge function to describe viral kinetics. Different shapes for the viral kinetics trajectory may become apparent as more data become available. Nonetheless, our simulated population distribution of Ct values is comparable to existing data and we propagated substantial uncertainty in viral kinetics parameters to generate a wide range of viral trajectories. For prevalence estimation, the MLE framework requires training on a distribution of Ct values. Such data can be available based on past tests from a given laboratory, but care should be taken to use a distribution appropriate for the population under consideration. For example, training the virus kinetics model on data skewed towards lower viral loads (as would be observed during the tail end of an epidemic curve) may be inappropriate when the true viral load distribution is skewed higher (as might be the case during the growth phase of an epidemic curve). Nevertheless, we used our simulated

distribution of Ct, which were fit to virus kinetics in published reports from distinct labs from across the world and obtained highly accurate results throughout. Thus, despite the limitations just mentioned, this shows that the virus kinetics models are quite robust and may not, in practice, require new fitting to the individual laboratory or population. In addition, while we assume that individuals are sampled from the population at random in our analysis, in practice samples that are processed together are also typically collected together, which may bias the distribution of positive samples among pools.

We have shown that simple designs that are straightforward to implement have the potential to greatly improve testing throughput across the time course of the pandemic. These principles likely also hold for pooling of sera for antibody testing, which remains an avenue for future work. There are logistical challenges and additional costs associated with pooling that we do not consider deeply here, and it will therefore be up to laboratories and policy makers to decide where these designs are feasible. Substantial coordination will therefore be necessary to make group testing practical but investing in these efforts could enable community screening where it is currently infeasible and provide epidemiological insights that are urgently needed.

Materials and Methods

Simulation model of infection dynamics and viral load kinetics

We developed a population-level mathematical model of SARS-CoV-2 transmission that incorporates realistic within-host virus kinetics. Full details are provided in **Supplementary Material 1, sections 1-4,** but we provide an overview here. First, we fit a viral kinetics model to published longitudinally collected viral load data from nasopharyngeal swab and sputum samples using a Bayesian hierarchical model that captures the variation of peak viral loads, delays from infection to peak and virus decline rates across infected individuals (**Fig. 5A&B**; **Fig. S8**). (19) By incorporating estimated biological variation in virus kinetics, this model allows random draws each representing distinct within-host virus trajectories. We then simulated infection prevalence during a SARS-CoV-2 outbreak using a deterministic Susceptible-Exposed-Infected-Removed (SEIR) model with parameters reflecting the natural history of SARS-CoV-2 (**Fig. 5D**). For each simulated infection, we generated longitudinal virus titers over time by drawing from the distribution of fitted virus kinetic curves, using distributions derived using either nasopharyngeal swab or sputum data (**Fig. S4**). All estimated and assumed model parameters are shown in **Table S5**, with model fits shown in **Fig. S8.** Posterior estimates and Markov chain Monte Carlo trace plots are shown in **Fig. S9** and **Fig. S10.** We accounted for measurement variation by: i) transforming viral loads into Ct values under a range of Ct calibration curves, ii) simulating false positives with 1% probability, and, importantly, iii) simulating sampling variation. We assumed a limit of detection (LOD) of 100 RNA copies / ml.

Estimating prevalence from pooled test results

We adapted a statistical (maximum likelihood) framework initially developed to estimate HIV prevalence with pooled antibody tests to estimate prevalence of SARS-CoV-2 using pooled samples. (12,13) The framework accounts for the distribution of viral loads (and uncertainty

around them) measured in pools containing a mixture of negative and potentially positive samples. By measuring viral loads from multiple such pools, it is possible to estimate the prevalence of positive samples without individual testing. See **Supplementary Material 1, section 5** for full details.

We evaluated prevalence estimation under a range of sample availabilities (*N total samples; N*=288 to ~18,000) and pooling designs. We varied the pool size of combined specimens (*n samples per pool*; *n*=48, 96, 192, or 384) and the number of pools (*b*=6, 12, 24, or 48). For each combination, we estimated the point prevalence from pooled tests on random samples of individuals drawn during epidemic growth (days 20-120) and decline (days 155-300). Because the data is realistic but simulated, we used ground truth prevalence in the population and, separately, in the specific set of samples collected from the overall population to assess accuracy of our estimates (see for example **Fig. 3B**). We calculated estimates for 100 entirely distinct epidemic simulations.

Pooled tests for individual sample identification

Using the same simulated population, we evaluated a range of simple and combinatorial pooling strategies for individual positive sample identification (**Supplementary Material 1, section 6**). In simple pooling designs, each sample is placed in one pool, and each pool consists of some pre-specified number of samples. If a pool tests positive, all samples that were placed in that pool are retested individually (**Fig. 1A**). For combinatorial pooling, each sample is split into multiple, partially overlapping pools (**Fig. 1B**). (9,10) Every sample that was placed in any pool that tested negative is inferred to be negative, and the remaining samples are identified as potential positives. Here, we consider a very simple form of combinatorial testing, where identified potential positive samples are individually tested in a validation stage.

A given pooling design is defined by three parameters: the total number of individuals to be tested (*N*); the total number of pools to test (*b*); and the number of pools a given sample is included in (*q*). For instance, if we have 50 individuals (*N*) to test, we might split the 50 samples into four pools (*b*) of 25 samples each, where each sample is included in two pools. Note that, by definition, in simple pooling designs each sample is placed in one pool (*q*=1).

To identify optimal testing designs under different resource constraints, we systematically analyzed a large array of pooling designs under various sample and test kit availabilities. We evaluated different combinations of between 12 and ~6,000 available samples/tests per day. The daily testing capacity shown is the daily average, though we assume that there is some flexibility to use fewer or more tests day to day (*i.e.,* there is a budget for period of time under evaluation).

For each set of resource constraints, we evaluated designs that split *N* samples between 1 to 96 distinct pools, and with samples included in *q*=1 (simple pooling), 2, 3, or 4 (combinatorial pooling) pools (**Table S1**). To ensure robust estimates (especially at low prevalences of less than 1 in 10,000), we repeated each simulated pooling protocol at each time point in the epidemic up to 200,000 times.

In each scenario, we calculated: i) the sensitivity to detect positive samples when they existed in the pool; ii) the efficiency, defined as the total number of samples tested divided by the total number of tests used; iii) the total number of identified true positives (total recall); and iv) the effectiveness, defined as the total recall relative to individual testing.

Pilot experiments

For validation experiments of our simulation-based approach, we used fully de-identified, discarded human nasopharyngeal specimens obtained from the Broad Institute of MIT and Harvard. In each experiment, sample aliquots were pooled before RNA extraction and qPCR

and pooled specimens were tested using the Emergency Use Authorization (EUA) approved SARS-CoV-2 assay performed by the Broad Institute CLIA laboratory. The protocol is described in full detail in **Supplementary Material 1, section 7**. Specifics of each pooling approach is detailed alongside the relevant results and in **Supplementary Material 1, section 8**.

References

1. Coronavirus disease (COVID-19) Weekly Epidemiological Update Global epidemiological situation. [cited 2020 Aug 28]. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200824-weekly-epi-update.pdf?sfvrsn=806986d1_4

2. World Health Organization (WHO). Critical preparedness, readiness and response actions for COVID-19. World Heal Organ. 2020;

3. Cheng MP, Papenburg J, Desjardins M, Kanjilal S, Quach C, Libman M, et al. Diagnostic Testing for Severe Acute Respiratory Syndrome–Related Coronavirus-2. Ann Intern Med. 2020;

4. Yoo JH, Chung MS, Kim JY, Ko JH, Kim Y, Kim YJ, Kim JM, Chung YS, Kim HM, Han MG KSY. Report on the epidemiological features of coronavirus disease 2019 (covid-19) outbreak in the republic of korea from january 19 to march 2, 2020. J Korean Med Sci. 2020;

5. Lee VJ, Chiew CJ, Khong WX. Interrupting transmission of COVID-19: lessons from containment efforts in Singapore. J Travel Med. 2020;

6. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med. 2020;

7. Grassly NC, Pons-Salort M, Parker EPK, White PJ, Ferguson NM, Ainslie K, et al. Comparison of molecular testing strategies for COVID-19 control: a mathematical modelling study. Lancet Infect Dis. 2020 Aug [cited 2020 Aug 20];0(0). Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309920306307

8. Dorfman R. The Detection of Defective Members of Large Populations. Ann Math Stat. 1943;

9. Du D, Hwang FK, Hwang F. Combinatorial group testing and its applications. Vol. 12. World Scientific; 2000.

10. Porat E, Rothschild A. Explicit non-adaptive combinatorial group testing schemes. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2008.

11. Centre for Disease Prevention E. Methodology for estimating point prevalence of SARS-CoV-2 infection by pooled RT-PCR testing. 2020 [cited 2020 Aug 28]. Available

from: https://www.ecdc.europa.eu/en/publications-data/methodology-estimating-point-prevalence-sars-cov-2-infection-pooled-rt-pcr

12. Zenios SA, Wein LM. Pooled testing for HIV prevalence estimation: Exploiting the dilution effect. Vol. 17, Statistics in Medicine. John Wiley & Sons, Ltd; 1998 [cited 2020 Aug 14]. p. 1447–67. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/%28SICI%291097-0258%2819980715%2917%3A13%3C1447%3A%3AAID-SIM862%3E3.0.CO%3B2-K

13. Wein LM, Zenios SA. Pooled Testing for HIV Screening: Capturing the Dilution Effect. Vol. 44, Operations Research. INFORMS; [cited 2020 Aug 14]. p. 543–69. Available from: https://www.jstor.org/stable/171999

14. Hogan CA, Sahoo MK, Pinsky BA. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. JAMA. 2020 May 19 [cited 2020 Jul 3];323(19):1967. Available from: https://jamanetwork.com/journals/jama/fullarticle/2764364

15. Bilder CR, Iwen PC, Abdalhamid B, Tebbs JM, McMahan CS. Tests in short supply? Try group testing. Significance (Oxford, England). 2020 Jun [cited 2020 Jul 8];17(3):15–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32536952

16. Ben-Ami R, Klochendler A, Seidel M, Sido T, Gurel-Gurevich O, Yassour M, et al. Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection. Clin Microbiol Infect. 2020 Sep 1;26(9):1248–53.

17. Shental N, Levy S, Wuvshet V, Skorniakov S, Shalem B, Ottolenghi A, et al. Efficient high-throughput SARS-CoV-2 testing to detect asymptomatic carriers. Sci Adv. 2020 Aug 21 [cited 2020 Aug 28];eabc5961. Available from: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abc5961

18. Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. New England Journal of Medicine. 2020.

19. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020;

20. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2 viral load dynamics, duration of viral shedding and infectiousness: a living systematic review and meta-analysis. medRxiv. 2020 Jul 29 [cited 2020 Aug 28];pre-print:2020.07.25.20162107. Available from:

http://medrxiv.org/content/early/2020/07/28/2020.07.25.20162107.abstract

21. Williams TC, Wastnedge E, McAllister G, Bhatia R, Cuschieri K, Kefala K, et al. Sensitivity of RT-PCR testing of upper respiratory tract samples for SARS-CoV-2 in hospitalised patients: a retrospective cohort study. medRxiv. 2020 Jun 20 [cited 2020 Aug 28];2020.06.19.20135756. Available from: http://medrxiv.org/content/early/2020/06/20/2020.06.19.20135756.abstract

22. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, Zambrano-Achig P, del Campo R, Ciapponi A, et al. FALSE-NEGATIVE RESULTS OF INITIAL RT-PCR ASSAYS FOR COVID-19: A SYSTEMATIC REVIEW. medRxiv. 2020 Aug 13 [cited 2020 Sep 25];2020.04.16.20066787. Available from: https://doi.org/10.1101/2020.04.16.20066787

23. Larremore DB, Wilder B, Lester E, Shehata S, Burke JM, Hay JA, et al. Test sensitivity is secondary to frequency and turnaround time for COVID-19 surveillance. medRxiv. 2020 Jun 27 [cited 2020 Aug 28];2020.06.22.20136309. Available from: https://www.medrxiv.org/content/10.1101/2020.06.22.20136309v2

24. Paltiel AD, Zheng A, Walensky RP. Assessment of SARS-CoV-2 Screening Strategies to Permit the Safe Reopening of College Campuses in the United States. JAMA Netw open. 2020 Jul 1 [cited 2020 Aug 28];3(7):e2016818. Available from: https://jamanetwork.com/

25. Riley S, Ainslie K, Eales O, Jeffrey B, Walters C, Atchison C, et al. Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study. medRxiv. 2020 Jul 11 [cited 2020 Aug 28];2020.07.10.20150524. Available from: https://doi.org/10.1101/2020.07.10.20150524

26. Pouwels KB, House T, Robotham J V, Birrell P, Gelman AB, Bowers N, et al. Community prevalence of SARS-CoV-2 in England: Results from the ONS Coronavirus Infection Survey Pilot. medRxiv. 2020 Jul 7 [cited 2020 Jul 8];2020.07.06.20147348. Available from: https://www.medrxiv.org/content/10.1101/2020.07.06.20147348v1

27. Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Barzon L, Del Vecchio C, et al. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature. 2020 Jun 30 [cited 2020 Aug 28];584(7821):425. Available from: https://doi.org/10.1038/s41586-020-2488-1

28. Jing Q-L, Liu M-J, Zhang Z-B, Fang L-Q, Yuan J, Zhang A-R, et al. Household

secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study. Lancet Infect Dis. 2020 Jun [cited 2020 Aug 28];0(0). Available from: https://doi.org/10.1016/S1473-3099

29. Li W, Zhang B, Lu J, Liu S, Chang Z, Cao P, et al. The characteristics of household transmission of COVID-19. Clin Infect Dis. 2020 Apr 17 [cited 2020 Aug 28]; Available from: /pmc/articles/PMC7184465/?report=abstract

30. Huang Y-T, Tu Y-K, Lai P-C. Estimation of the secondary attack rate of COVID-19 using proportional meta-analysis of nationwide contact tracing data in Taiwan. J Microbiol Immunol Infect. 2020 Jun 11 [cited 2020 Jul 8]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/32553448

31. Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. Vol. 395, The Lancet. Lancet Publishing Group; 2020 [cited 2020 Aug 28]. p. e47. Available from: http://ees.elsevier.com/thelancet/www.thelancet.com

32. Hamner L, Dubbel P, Capron I, Ross A, Jordan A, Lee J, et al. High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice - Skagit County, Washington, March 2020. MMWR Morb Mortal Wkly Rep. 2020 May 15 [cited 2020 Jul 8];69(19):606–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32407303

33. Lennon NJ, Bhattacharyya RP, Mina MJ, Rehm HL, Hung DT, Smole S, et al. Comparison of viral levels in individuals with or without symptoms at time of COVID-19 testing among 32,480 residents and staff of nursing homes and assisted living facilities in Massachusetts. medRxiv. 2020 Jul 26 [cited 2020 Aug 14];2020.07.20.20157792. Available from: https://doi.org/10.1101/2020.07.20.20157792

34. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020 May 28 [cited 2020 Aug 14];581(7809):465–9. Available from: https://doi.org/10.1038/s41586-020-2196-x

35. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020 Mar 10 [cited 2020 Apr 3]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/32150748

36. lazymcmc. [cited 2018 Jun 21]. Available from: https://github.com/jameshay218/lazymcmc

37. Poon LLM, Chan KH, Wong OK, Cheung TKW, Ng I, Zheng B, et al. Detection of SARS Coronavirus in Patients with Severe Acute Respiratory Syndrome by Conventional and Real-Time Quantitative Reverse Transcription-PCR Assays. Clin Chem. 2004 Jan [cited 2020 Aug 14];50(1):67–72. Available from: https://pubmed.ncbi.nlm.nih.gov/14709637/

Acknowledgements: We thank Benedicte Gnangnon, Rounak Dey, Xihong Lin, Edgar Doriban, Rene Niehus and Heather Shakerchi for useful discussions. Work was supported by the Merkin Institute Fellowship at the Broad Institute of MIT and Harvard (B.C.), by the National Institute of General Medical Sciences (#U54GM088558; J.H. and M.M.); and by an NIH DP5 grant (M.M.), and the Dean's Fund for Postdoctoral Research of the Wharton School and NSF BIGDATA grant IIS 1837992 (D.H.).

Author contributions: B.C., J.H., A.R. and M.M. conceived the study. B.C. and J.H. performed the simulations and modelling. B.B., M.H, M.C, J.B, and B.S performed the pooled testing. B.C., J.H, A.R, and M.M. analyzed results and wrote the manuscript with input from all authors.

Competing interests: A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and until July 31, 2020, was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Asimov, and Neogene Therapeutics. From August 1, 2020, A.R. is an employee of Genentech.

Code availability: The SEIR model, viral kinetics model and MCMC were implemented in R version 3.6.2. The remainder of the work was performed in Python version 3.7. The code used for the MCMC framework is available at: https://github.com/jameshay218/lazymcmc. All other code and data used are available at: https://github.com/cleary-lab/covid19-group-tests.

Data availability: Raw data generated in this study are available in **Table S2** and **Table S6.** Simulated data can be regenerated using the accompanying code.

Supplementary Material 1: additional materials and methods

1. Model of infection dynamics

We implemented a deterministic compartmental model to describe the increase and subsequent decline of SARS-CoV-2 infection incidence and prevalence. The model captured the following progression: individuals begin fully susceptible to infection (S); become exposed but not yet infectious (E); become infectious (I); and are finally removed (R). Our aim was to capture changes in incidence over time (increasing incidence rate before the peak and declining incidence thereafter) and the resulting population-level distribution of viral loads, and we therefore did not model additional complexity such as deaths, pre-symptomatic transmission or age-stratified outcomes. Although incorporating these mechanisms may alter the simulated epidemic dynamics for a given set of transmission parameters, we do not expect them to impact inference from the pooled testing analyses.

The model was defined by the following set of ordinary differential equations:

$$\frac{dS}{dt} = \frac{-\beta SI}{N}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E$$

$$\frac{dI}{dt} = \sigma E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

where β is the transmission rate, scaled to give a basic reproductive number R_0 of 2.5; $1/\sigma$ is the incubation period, assumed to have a mean of 6.4 days; $1/\gamma$ is the infectious period assumed to have a mean of 7 days; and N is the population size, set to 12,500,000 to generate at least 10,000,000 infected individuals. All model parameters and assumed values are

described in **Table S5**. Using this model, we simulated per capita daily infection probability (i.e., the daily incidence) and prevalence.

2. Viral load kinetics

Following infection (t_{inc}) and after a latent period (t_g), viral load was assumed to increase exponentially, peaking a short time (t_p) after the end of the latent period and before the onset of symptoms (t_o). Viral load then decreased monotonically to undetectable levels, defined by a time parameter (t_w) giving the number of days post symptom onset before crossing the limit of detection. These assumptions are equivalent to linear growth and decay on a log scale. We chose to model viral load waning with respect to symptom onset rather than peak viral load, as almost all available time-series viral load data are presented with respect to symptom onset time. The log_{10} viral load over time, $v(t)$, was given by:

$$v(t) = \begin{cases} 0, & t_{inf} < t \leq t_g \\ \frac{\alpha}{t_p} t, & t_g < t \leq t_p \\ \alpha - \frac{\alpha}{t_{inc} - t_p + t_w}(t - t_p), & t_g < t \leq t_p \end{cases}$$

where t_{inf} is the time of infection; t_{inc} is the incubation period for symptom onset; α is the peak viral load; t_g is the latent period before viral growth; t_p is the time taken to reach peak viral load post; and t_w is the number of days from symptom onset to becoming undetectable. Note that all individuals are assigned a symptom onset time regardless of whether they show symptoms or not, as the parameter is used here to describe viral kinetics.

3. Fitting the viral kinetics model

Time-series viral load data were obtained from a case series of 9 hospitalized COVID-19 patients from a single hospital in Munich, Germany. (34) These data provide regular measurements of log_{10} RNA copies per nasopharyngeal swab and per ml of sputum. Data

were extracted using a web plot digitizer (https://automeris.io/WebPlotDigitizer/). We used a random-effects model to infer individual-level viral kinetics parameters alongside population-level distributions. Under this model, each individual had their own parameter values for t_{inc}, α, t_g, t_p and t_w. α and t_w drawn from a multivariate normal distributed with estimated means ($\bar{\alpha}, \overline{t_w}$), standard deviations ($\sigma_\alpha, \sigma_{t_w}$) and correlation ($\rho_{\alpha,t_w}$) (**Table S5**).

Although the viral load measurements were taken after the onset of symptoms, we are interested in modelling the entire time course from the time of infection, including uncertainty in the incubation period. We therefore estimated t_{inc} for each individual by placing a strongly informative log-normal prior on the incubation period based on previous estimates. (35) Because no observations were made before symptom onset, the posterior distribution for this parameter matched the prior.

The model was fit separately to the swab and sputum data using a Markov chain Monte Carlo (MCMC) framework, using a likelihood that assumes normally distributed observation error with estimated standard deviation σ_{obs} (**Fig. S8**). The likelihood function accounted for truncation at the lower limit of detection (LOD) by integrating over the left-hand tail of the normal cumulative density function for model-predicted viral loads less than the LOD. Furthermore, Wölfel et al. present data with a limit of detection of 0 log_{10} copies / swab or / ml, but a limit of quantification of 2 log_{10} copies. We therefore considered that observed viral loads between 0 and 2 had unknown between 0 and 2. We placed uniform priors on $\bar{\alpha}$ and $\overline{t_w}$, half Cauchy priors on σ_α and σ_{t_w} with scale parameters of 1, a uniform prior on ρ_{α,t_w} between -1 and 1 and informative uniform priors on t_g, and t_p.

Code to regenerate the MCMC fitting is available in the accompanying GitHub repository. (36) Briefly, an adaptive Metropolis-Hastings algorithm was run for 3,000,000 iterations with the first 1,000,000 iterations discarded as burn in to obtain posterior estimates for the model parameters (**Fig. S9**). Convergence was assessed visually based on trace plots of 3

748 independent chains (**Fig. S10**). All estimated parameters had an effective sample size (ESS)
749 greater than 200 and $\hat{R} < 1.1$.

750 **4. Simulation framework**

751 We combined the estimates from the transmission model and viral kinetics model above to
752 simulate viral loads for a population of 12,500,000 individuals over time. We performed
753 separate simulations for swab and sputum data. For each individual, we simulated a time of
754 infection (or no infection) from the time-varying probability of infection, $p(t)$. For each infected
755 individual, an incubation period was drawn from a log normal distribution with mean (of the
756 natural log) = 1.621 and standard deviation (of the natural log) = 0.418 as estimated previously.
757 (35) This incubation period was used as a reference point for t_w, but we did not explicitly model
758 symptomatic vs. asymptomatic individuals. Next, we drew viral kinetics parameters for each
759 individual from the joint posterior distribution. Finally, we simulated the time-varying viral load
760 of each infected individual, starting at the time of infection. Because of the high variance of
761 the drawn kinetics parameters, some simulated viral loads reached very high values, and we
762 therefore truncated all simulated viral loads at 16 log_{10} virus copies per ml. These simulations
763 provide a population of viral loads incorporating realistic variability in viral load kinetics and
764 between-individual parameters.

765 **5. Estimating prevalence from pooled test results**

766 In this section, we adapt a previously described maximum likelihood estimate (MLE)
767 framework to estimate prevalence of positive samples, *p*, amongst *N* samples using *b* pooled
768 tests, with each pooling containing $n = \frac{N}{b}$. (12,13) Briefly, the MLE is determined by computing
769 the conditional probability of an observed RT-qPCR result given that there were *k* positive
770 samples in the pool, and integrating over all values of *k*. To compute these conditional
771 distributions, we estimate an empirical distribution of cycle threshold (Ct) values that follows
772 from the distribution of viral loads in infected individuals derived from a subset of the simulated

viral load measurements described above. This is conceptually similar to using the empirical distribution of all real measurements from a population taken to date.

Let $f_{k,n} = P(Y|\, k,n)$ represent the conditional probability of the observed Ct values in each pool, $Y = [y_1,\ldots,y_b]$, given that there were *k* positive samples in the pool and *n-k* negative samples. Computing these distributions will be the central task of generating prevalence estimates.

To begin, we model Ct values as a function of the underlying viral loads:

$$y(v) = y_0 - log_2(10)\, log_{10}(v),$$

where *v* is the viral load (in copies per ml). To capture variance in the LOD between labs or across batches within a lab, we let the intercept, y_0, be normally distributed with unknown mean and standard deviation. We assume uniform priors over each y_0 ($\mu \sim Uniform(37{,}39)$, $\sigma \sim Uniform(0.1{,}1)$). These ranges and the log-linear relationship are based on a calibration of Ct values to a dilution curve of positive control SARS-CoV-2 RNA samples (data not shown) and are consistent with prior measurements calibrated to SARS coronavirus. (37)

$f_{k,n}$ depends not only on the distribution of Ct values arising from a given viral load, but also on the distribution of viral loads in infected and uninfected individuals. Here we assume that the viral load in uninfected individuals is always zero (we do, however, allow for false positive PCR results, as discussed below). Given access to the true distribution of viral loads, $f_{k,n}$ can be generated by the convolution of *k* density functions. Since we do not have access to the true distributions, we approximate $f_{k,n}$ through a kernel density estimate (KDE) of empirical convolutions. Specifically, for a given value of *n* and each value of *k* from 1 to *n*, we generate 10,000 random combinations of *k* viral loads,

$$V_{k,n}: v_i = \sum_{j=1}^{k} z_j \quad for\ i\ from\ 1\ to\ 10{,}000,$$

where z_j are the randomly sampled viral loads. For each *i*, we sample $\mu \sim Uniform(37,39)$, $\sigma \sim Uniform(0.1,1)$ and $y_0 \sim N(\mu, \sigma)$ and then generate Ct values:

$$X_{k,n}: x_i = y(v_i) \ \ for\ i\ from\ 1\ to\ 10{,}000.$$

We then approximate $f_{k,n} = KDE(X_{k,n})$ with a bandwidth parameter of 0.1.

Several adjustments to these density approximations are needed to account for false positives and undefined Ct values (*i.e.*, undetected viral RNA). To allow for Ct values that might arise from false positives, we define $f_{0,n} = f_{1,n}\, r$, where *r* is an assumed false positive rate of PCR (here, we set $r = 0.2\%$ and allow this to be misspecified when simulating false positives below). When viral RNA is undetected and $k > 1$, we let $f_{k,n}(undetected) = \int_{40}^{\infty} f_{k,n}(x)\ dx$, where a Ct value of 40 is used as a typical limit of detectable RNA. When $k = 0$, $f_{0,n}(undetected) = 1 - r$.

Following the model of Zenios and Wein, (12) the MLE of prevalence, *p*, is defined as:

$$p_{estimated} = \frac{1}{nb} \sum_{i=1}^{b} \sum_{k=0}^{n} kP(k \mid y_i, p),$$

where *n* is the number of samples per pool, *b*, is the number of pools, and y_i is the Ct value observed in pool *i*. We calculate $P(k \mid y_i, p)$ using the conditional densities above:

$$P(k \mid y_i, p) = \frac{\binom{n}{k} p^k (1-p)^{n-k} f_{k,n}(y_i - log_2(10)\, log_{10}(n))}{\sum_{j=0}^{n} \binom{n}{k} p^j (1-p)^{n-j} f_{j,n}(y_i - log_2(10)\, log_{10}(n))},$$

substituting the appropriate values of f when the Ct value is undefined. Note that each of the observed Ct values is adjusted to account for *n*-fold dilution. Finally, we find $p_{estimated}$ through an iterative expectation-maximization algorithm, where at each iteration

$$p_{t+1} = \frac{1}{nb}\sum_{i=1}^{b}\sum_{k=0}^{n} kP(k \mid y_i, p_t).$$

Using this model, we evaluated prevalence estimation across time in our simulated populations. We first partition the simulated population into training and testing sets. Observed viral loads in 20% of individuals (selected at random) are used to train the KDEs above. Furthermore, training data were divided into growth phase (days 20 to 120) or decline phase (days 155 to 300) data, allowing us to ascertain the effect of training and estimating on data from consistent or inconsistent phases of the epidemic. All remaining prevalence estimation and analyses are done on the testing population. We simulate viral loads in *b* pools of samples as $V: v_i = \sum_{j=1}^{n} Poisson(\frac{z_j}{n})$ for *i* from *1* to *b*, where Poisson sampling is used to model the sampling of a small volume from each swab. From these we simulate Ct values, *Y*. At each time point, we randomly select *N* individuals, partition them into *b* pools, and simulate the pooled viral loads and Ct values, *Y*, with $y_0 \sim N(38.5,1)$ sampled randomly in each trial and applying the transformation described previously. To capture false positive PCR results at a rate of 1%, with 1% probability in each pool we add a viral load of $Poisson(\frac{z_j}{n})$, with z_j selected at random from all positive samples with $\frac{z}{n} > LOD$ on a given date. We then compute the true prevalence in the population, p_{true}, the prevalence in the population of *N* samples, $p_{sampled}$, and the prevalence estimated from *b* tests, $p_{estimated}$. To quantify accuracy we calculate the mean and standard deviation of these values at each time point across 100 random trials and summarize differences between $p_{estimated}$ and either p_{true} or $p_{sampled}$ with the normalized root mean squared error (NRMSE; either $\frac{RMSE_{true}}{p_{true}}$ or $\frac{RMSE_{sampled}}{p_{sampled}}$; **Fig. S2**).

To demonstrate the importance of calibrating the correct empirical distribution of Ct values, **Fig. S2** shows the accuracy of the framework in recovering the true population prevalence during the growth and decline phases of the epidemic when the model is calibrated to Ct

values measured during the growth and decline phases. Estimation accuracy is highest when the calibrated Ct values were measured during the same phase as is being measured.

6. Group testing for sample identification

We evaluated an array of pooling designs for individual identification as described in the main text. For any given design we can simulate pooled testing results on the simulated population described above. For each time point, we randomly select *N* individuals from the population. Each individual is then randomly assigned to *q* out of *b* pools. For each pool, we then calculate the net sampled viral load, $v_i = \sum_{j \in pool_i}^{d} Poisson(\frac{z_j}{d})$ using the viral loads z_j for each individual included in $pool_i$, where Poisson sampling is used to model a small volume sampled from each swab. If $v_i > LOD$ the pool is assigned a positive value, otherwise the pool is negative (here, with *LOD*=100). To allow for false positive PCR results at a rate of 1%, with probability 1% we set the result of the pool to be positive, regardless of the viral load.

Simulations were run by repeating the random pooling, pooled testing, and decoding procedures described in the main text. At each point in time, 200,000 trials are run selecting *N* individuals at random in each trial. We then record the number of validation tests, *s*, in each trial, corresponding to the number of putative positive samples. Average efficiency at a point in time, *t*, is then calculated as $Efficiency_t = \frac{N}{b+\underline{s}}$. Average sensitivity is calculated as $Sensitivity_t = \frac{N_{putative}}{N_{positive}}$ where $N_{positive}$ is the total number of infected individuals (viral load > 1) that were sampled across all trials and $N_{putative}$ is the number of such individuals who were identified as putative positives. From these values we calculate $Total\ Recall_t = Efficiency_t\ Sensitivity_t$.

Based on these results we evaluated a large array of group testing designs (**Table S1**) under a series of constraints on number of samples collected and number of reactions run per day (**Fig. 4C**). For a given design, D: (N, b, q) we calculate the average number of tests run, $s_{total} =$

$b + \underline{s}$, and the total number of times that *D* could be run on a daily basis, $r_D = min(\frac{samples\ collected}{N}, \frac{testing\ kits}{s_{total}})$. We then calculate the effective testing capacity on a daily basis, $c_D = N\, r_D\, Sensitivity_D$, and average this value over days 40 to 90 (**Fig. 4C**) or 80 to 108 (**Fig. S6**) of the simulated epidemic. Notably, when sample collection is limiting, $c_D = samples\ collected * Sensitivity_{(samples\ collected, samples\ collected, 1)}$, corresponding to individual testing. Finally, the optimal design for a given combination of constraints is the design with the greatest average effective testing capacity. In **Fig. 4C** and **Fig. S6** we show the relative effectiveness, corresponding to effective capacity relative to individual testing. **Fig. S7** shows the same results as **Fig. 4** and **Fig. S6**, but when samples were simulated based on viral trajectories using the sputum data. **Fig. S5** shows the same results again, but for the decline phase of the epidemic.

7. Broad Institute CRSP SARS-CoV-2 Real-time Reverse Transcriptase (RT)-PCR Diagnostic Assay

RNA is isolated from nasopharyngeal and oropharygneal specimens in 50ul VTM/UTM using MagMAX™-96 Viral RNA Isolation Kits (Thermo Fisher Scientific, AMB18365), performed on an Agilent Bravo Liquid Handling Platform running VWorks (Build 11.4.0.1233), and is reverse transcribed to cDNA and subsequently amplified in the Applied Biosystems® ViiA7 Real-Time PCR Instrument with QuantStudio version 1.3 software. In the process, the probe anneals to a specific target sequence located between the forward and reverse primers. During the extension phase of the PCR cycle, the 5' nuclease activity of Taq polymerase degrades the probe, causing the reporter dye to separate from the quencher dye, generating a fluorescent signal. With each cycle, additional reporter dye molecules are cleaved from their respective probes, increasing the fluorescence intensity. Fluorescence intensity is monitored at each PCR cycle by Applied Biosystems® ViiA7 Real-Time PCR System with QuantStudio version

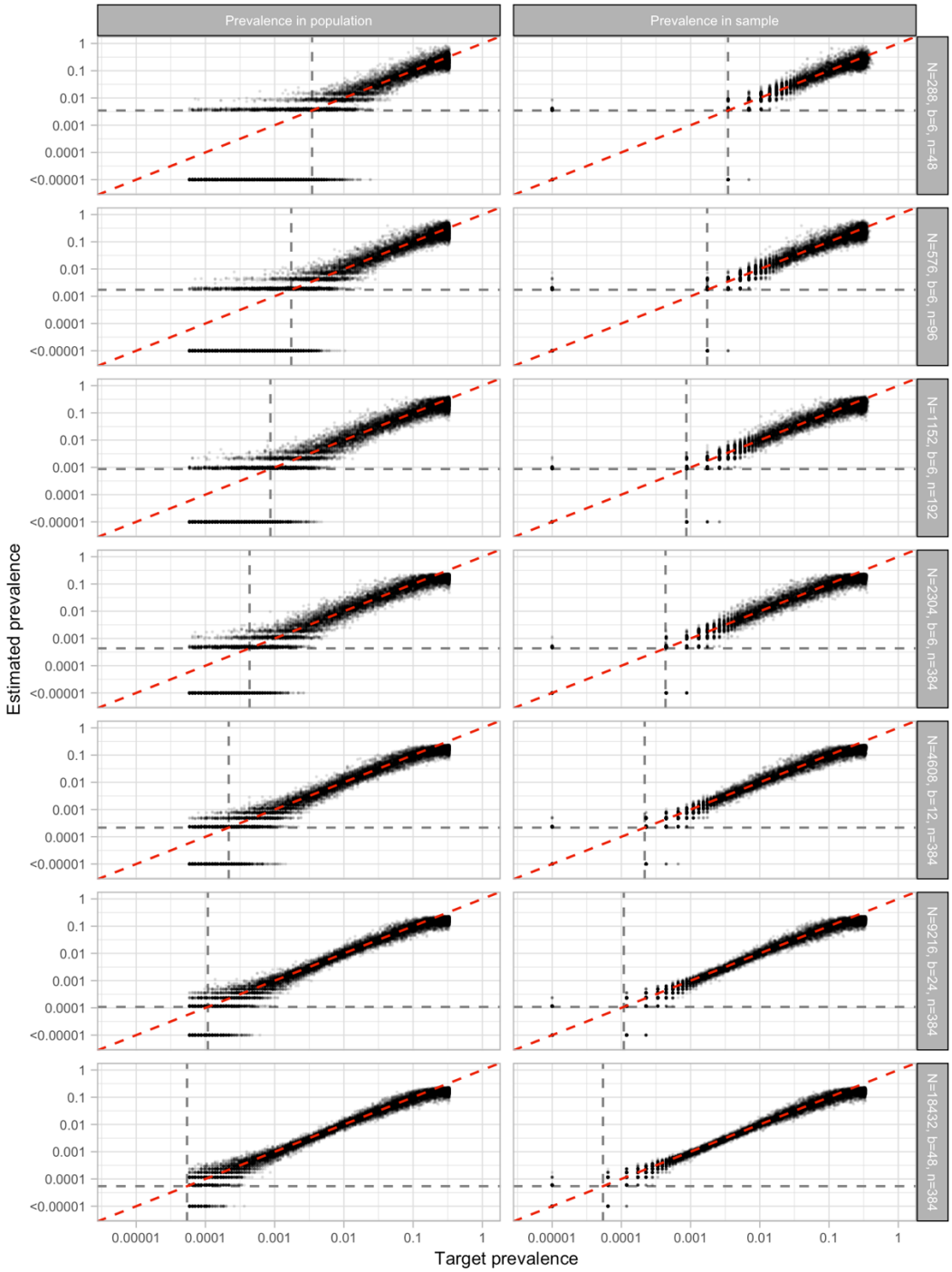
1.3 software. The software allows the fluorescence intensity to be monitored at each PCR cycle to allow for the qualitative detection of nucleic acid from SARS-CoV-2.

To create the final pools for our larger validation studies (10 batches of 96 for sample identification, and 48 pools of 48 samples for prevalence estimation), we treated a total of 96 known negative specimens as distinct samples across batches (in the case of sampled identification study) or pools (for the prevalence study). Each positive sample in a batch or pool was a distinct sample, and only used in one batch or pool. For the prevalence study, only pools with one or more positive samples were tested using the assay above (the remaining pools were assigned an "undefined" Ct value). All pools in all batches were tested in the identification study, regardless of whether they contained a positive sample.

8. Selecting pool compositions for large-scale sample identification validation

To form pools, we put each of the $\mathrm{N} = 96$ individuals into $q = 2$ out of $b = 6$ pools (A-F) by cycling through the following ordered list of pool pairs: AB, CD, EF, BC, DF, AE, BD, AF, CE, BE, CF, AD, BF, DE, AC. Namely, individual 1 is put in pools A and B, individual 2 in C and D, and so on; after individual 15, we return to the beginning of the list and cycle through again. Finally, since each pair of pools is assigned to multiple individuals, we reordered the individuals by what pair of pools they were put in. This final reordering simplifies the presentation to one more conveniently arranged for manual pipetting.

A strength of the procedure is that it is simple and flexible; it can easily be carried out for any number of individuals. Moreover, the resulting design here has pools with 32 individuals each, so the final pool sizes are balanced. In addition, the first 6 pool pairs are each assigned to 7 individuals and the remaining 9 pairs are each assigned to 6 individuals, so the pairs of pools are also approximately balanced. Indeed, the ordered list of pool pairs above was chosen to achieve this balance; the list uses each pair of pools once and the five consecutive sets of three pairs all use each pool once.



911

912 **Fig. S1:** Population prevalence (left) or prevalence in sample (right) against maximum
913 likelihood prevalence estimates. Population prevalence shown here are during the epidemic
914 growth phase. Results shown are from 100 independent trials at each day of the epidemic.
915 Each facet shows the pooling design with the fewest pools (tests used) for each sample size.
916 Dashed grey lines show one divided by the sample size, N.

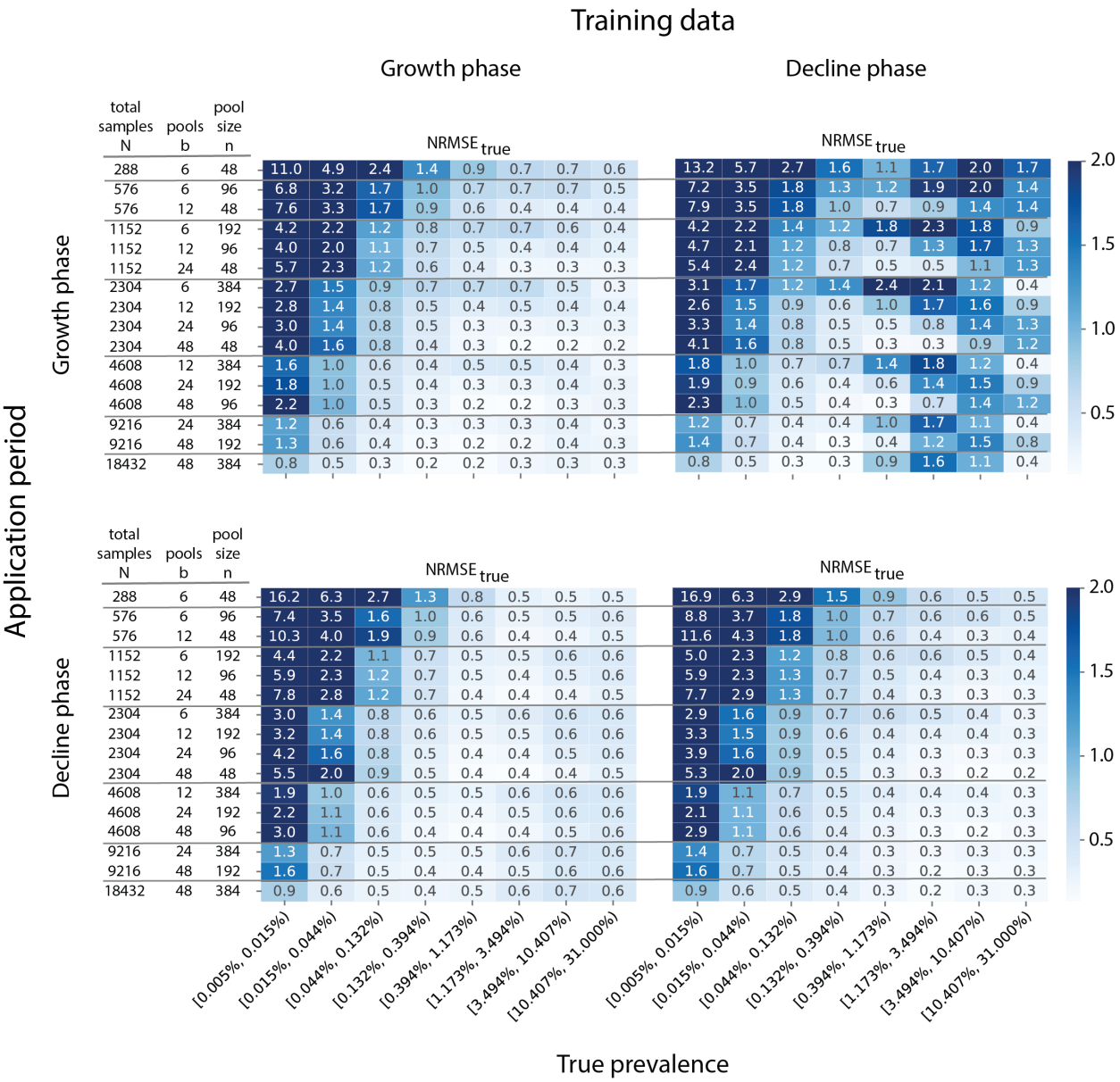
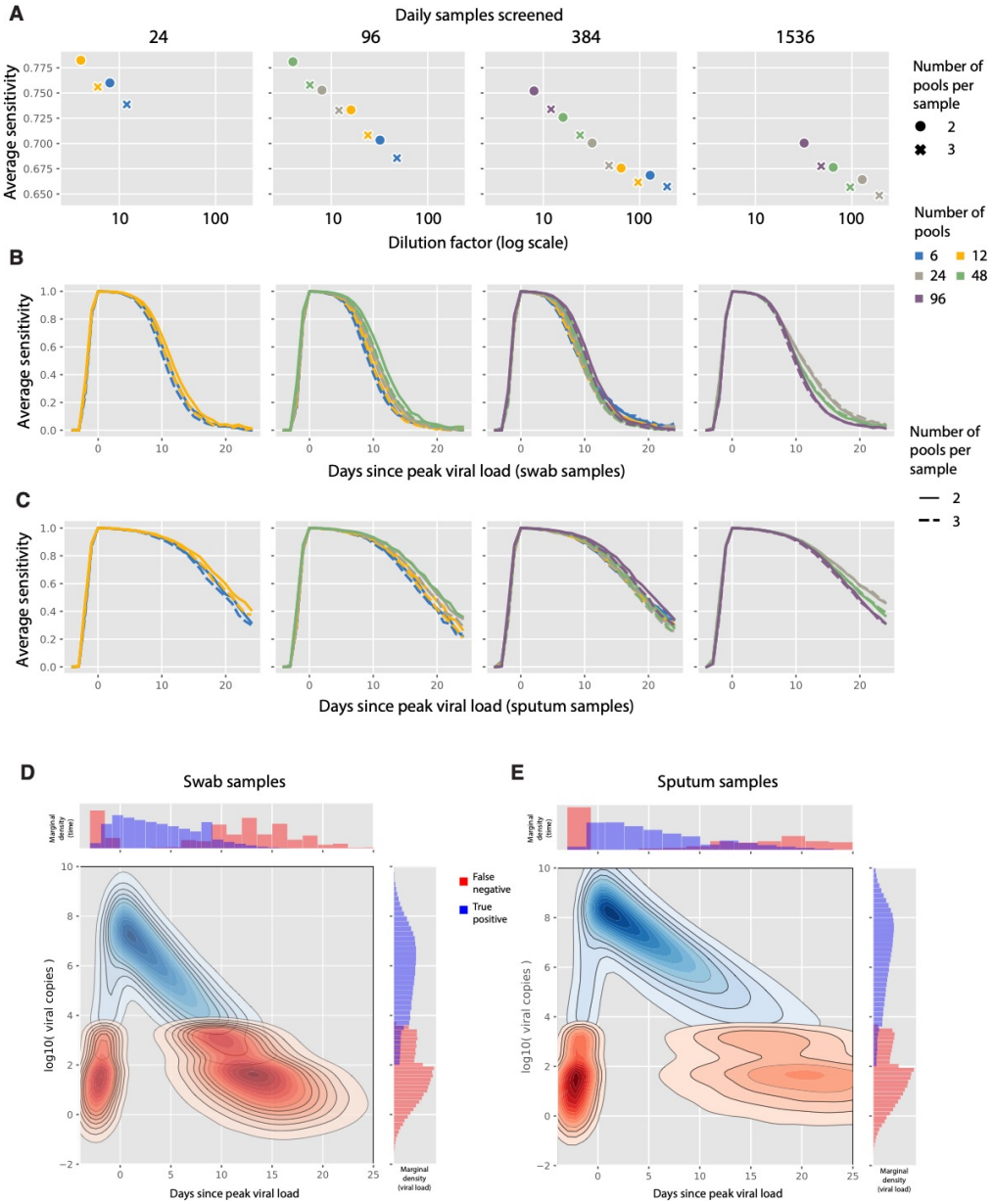


Fig. S2: Prevalence estimation can depend on training and application period. Prevalence estimation uses past observations to learn conditional distributions of pooled Ct values. When the inference algorithm is applied to Ct values from new pooling data, these training distributions are used to generate a maximum likelihood estimate. The error in these estimates (here, NRMSE, the root mean squared error normalized to the true prevalence) depends on the pooling design (individual rows) and true prevalence (columns, binned prevalence windows). In addition, error can depend on a (mis)match between the training

925 (two columns of panels) and application (two rows of panels) period. Colorbar and annotated

926 values within boxes indicate NRMSE.



927

928 **Fig. S3:** Sensitivity of sample identification relative to dilution factor and time since peak viral
929 load. **(A)** The sensitivity (y-axis) of different designs (individual points) is plotted against the
930 dilution factor of each design (x-axis, log scale). Pooling designs are separated by the
931 number of swab samples tested on a daily basis (individual panels); the number of pools
932 (color); and the number of pools into which each sample is split (circle versus cross). (**B** and
933 **C**) As in **(A)**, with sensitivity plotted against days since peak viral load (x-axis) for swab **(B)**

and sputum **(C)** samples. (**D** and **E**) The density of false negative (red) and true positive (blue) results is shown as a function of days since peak viral load (x-axis) and viral load (y-axis, *log10* scale) at the time of sample collection and pooled testing for swab **(D)** and sputum **(E)** samples. Contour plots depict 2-d density, and histograms show marginal density over time (top) and viral load (right).

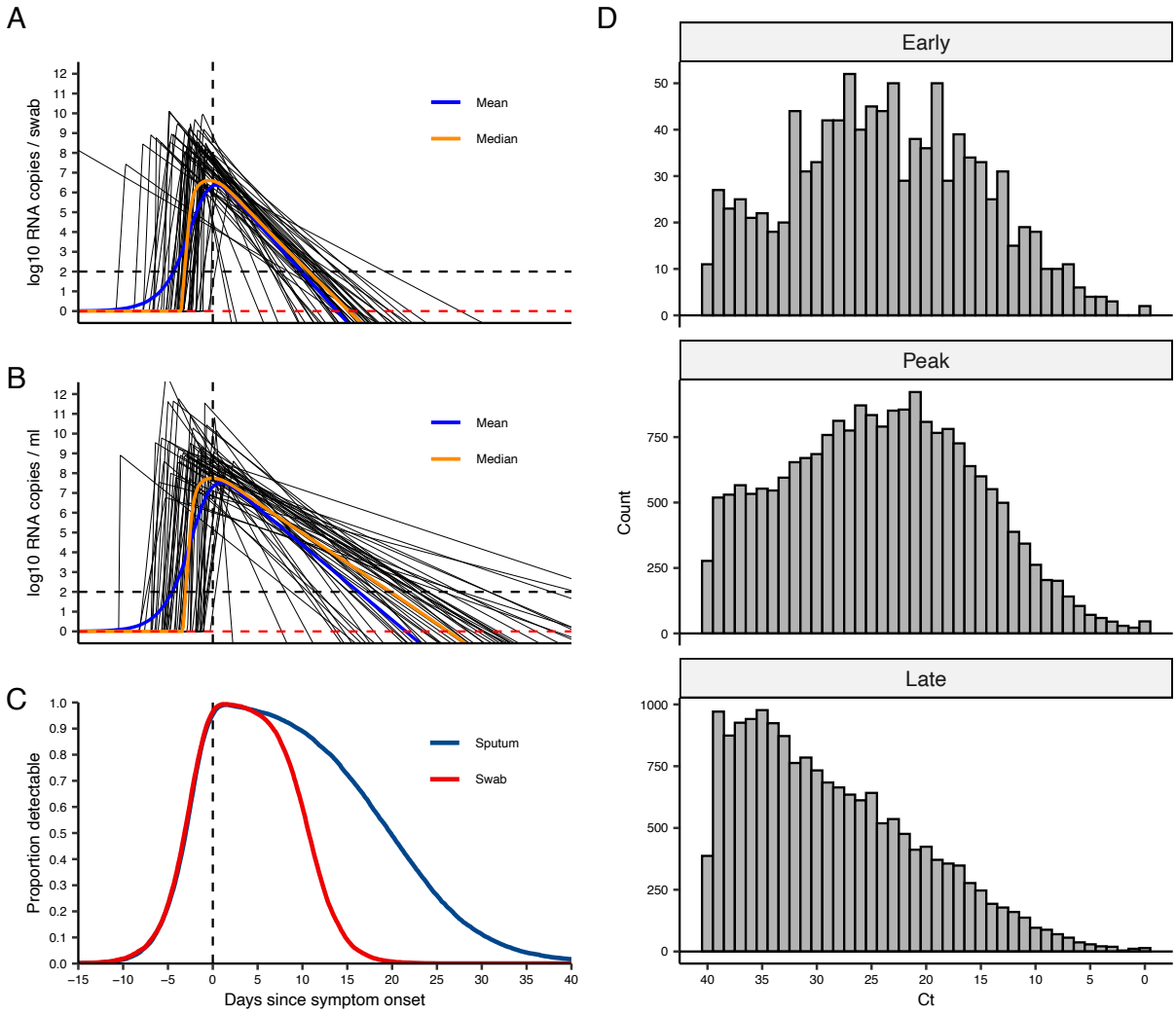


Fig. S4: Simulated viral loads. **(A)** Black lines show 50 randomly drawn viral load trajectories based on parameters estimated from fitting to swab data. Vertical dashed line shows time of symptom onset. Horizontal dashed lines show limit of detection assumed in subsequent analyses (black line) and limit of detection reported by Wölfel et al (red line). Blue and orange line shows median viral load on each day with respect to symptom onset. **(B)** as in **(A)**, but fitted to sputum data. **(C)** Proportion of true viral loads above the limit of detection (LOD = 2 *log10* RNA copies / swab or ml) on each day with respect to symptom onset for simulated swab (red) and sputum (blue) data. Note that the distribution for observed viral loads will differ slightly due to the addition of observation and sampling error. **(D)** Distribution of simulated cycle threshold values from swab samples at different stages of the epidemic.

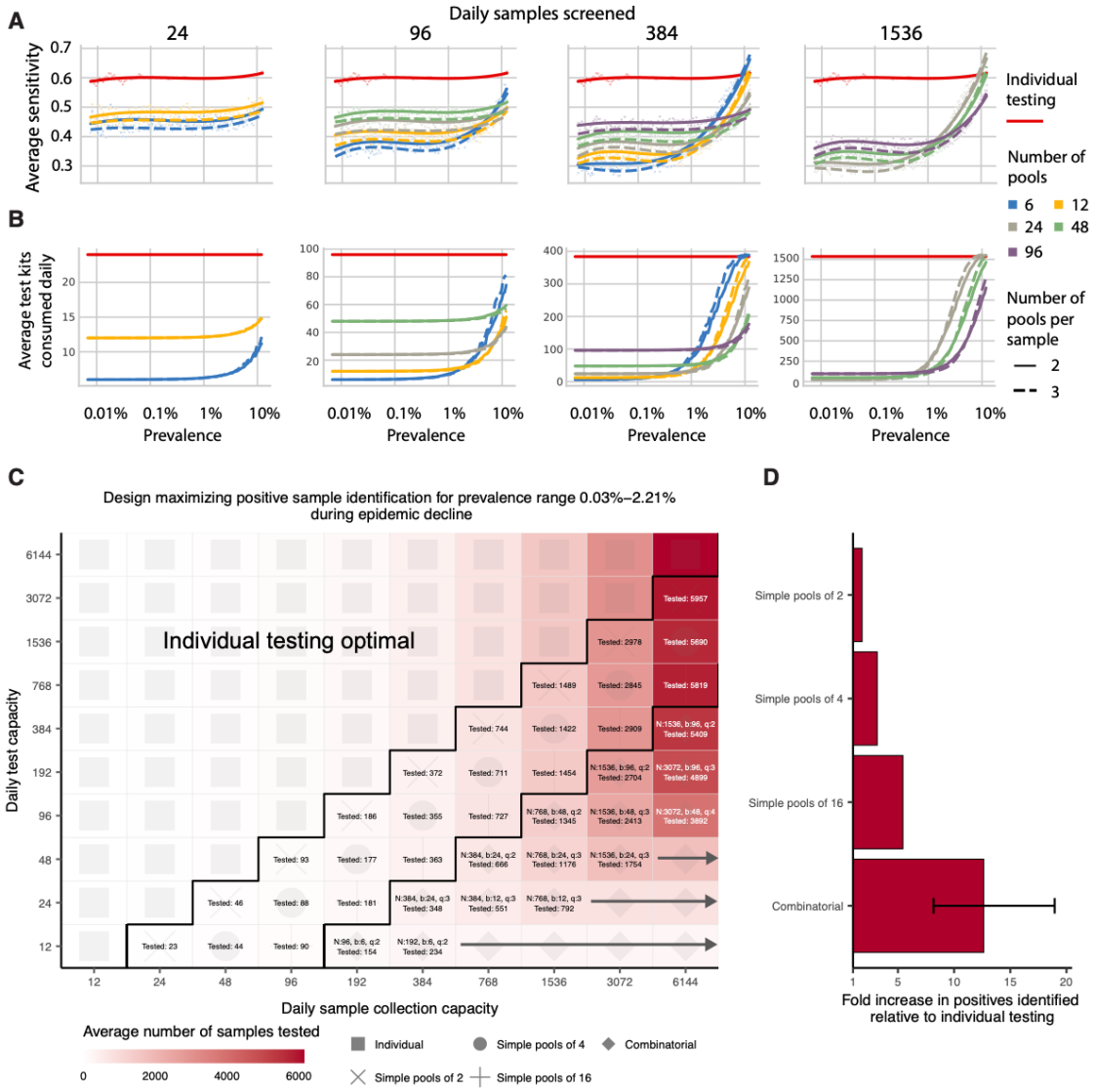


Fig. S5: Group testing for sample identification during epidemic decline. Results shown are based on the same testing strategies as in **Fig. 4**. We evaluated a variety of group testing designs for sample identification (**Table S1**) on the basis of sensitivity **(A)** shows sensitivity and **(B)** shows efficiency. total number of positive samples identified **(C)** and the fold increase in positive samples identified relative to individual testing **(D)**. (**A** and **B**) The average sensitivity (**A**, y-axis, individual points and spline) and average number of tests needed to identify individual positive samples (**B**, y-axis) using different pooling designs (individual lines) were measured over days 20-110 in our simulated population, with results plotted against

prevalence (x-axis, log-scale). Results show the average of 200,000 trials, with individuals selected at random on each day in each trial. Pooling designs are separated by the number of samples tested on a daily basis (individual panels); the number of pools (color); and the number of pools into which each sample is split (dashed versus solid line). Solid red line indicates results for individual testing. **(C)** Every design was evaluated under constraints on the maximum number of samples collected (columns) and average number of reactions that can be run on a daily basis (rows) over days 40-90. Text in each box indicates the optimal design for a given set of constraints (number of samples per batch (N), number of pools (b), number of pools into which each sample is split (q), average number of total samples screened per day). Color indicates the average number of samples screened on a daily basis using the optimal design. Arrows indicate that the same pooling design is optimal at higher sample collection capacities due to testing constraints. **(D)** Fold increase in the number of positive samples identified relative to individual testing with the same resource constraints. Error bar shows range amongst optimal designs.

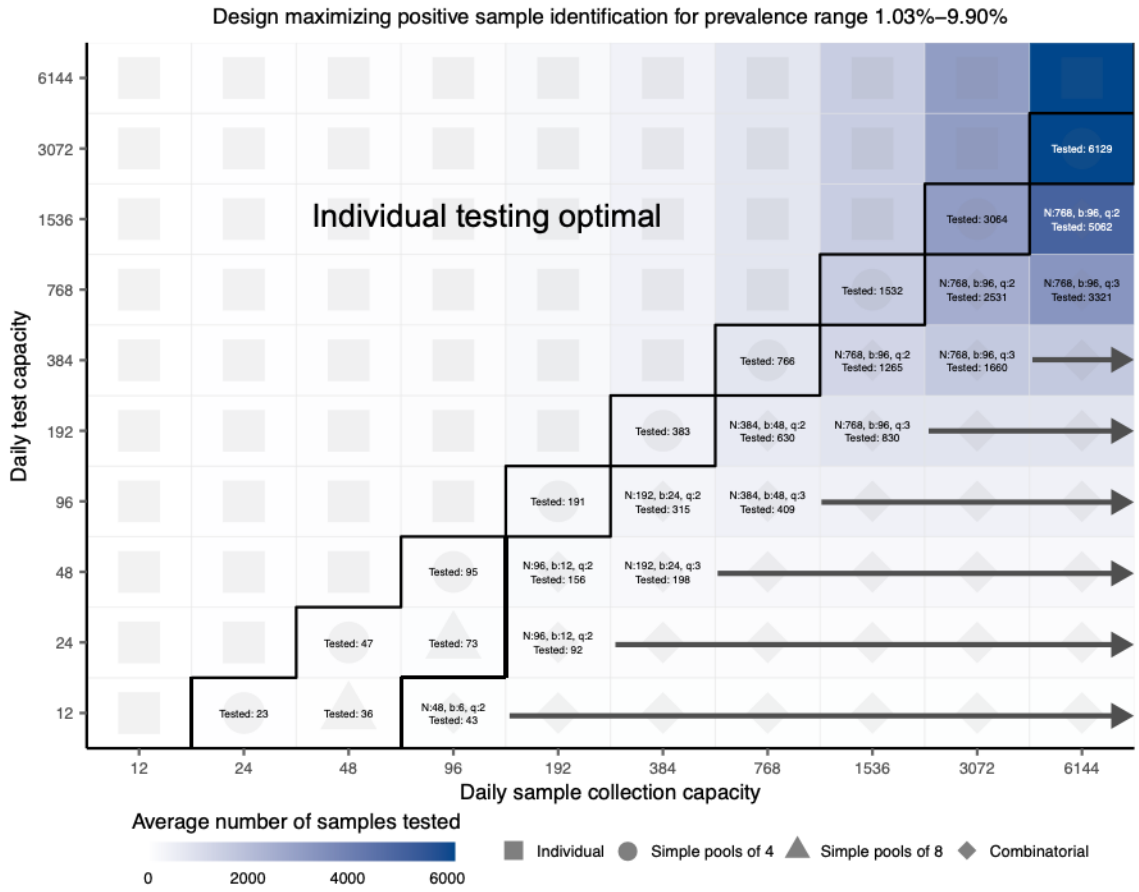
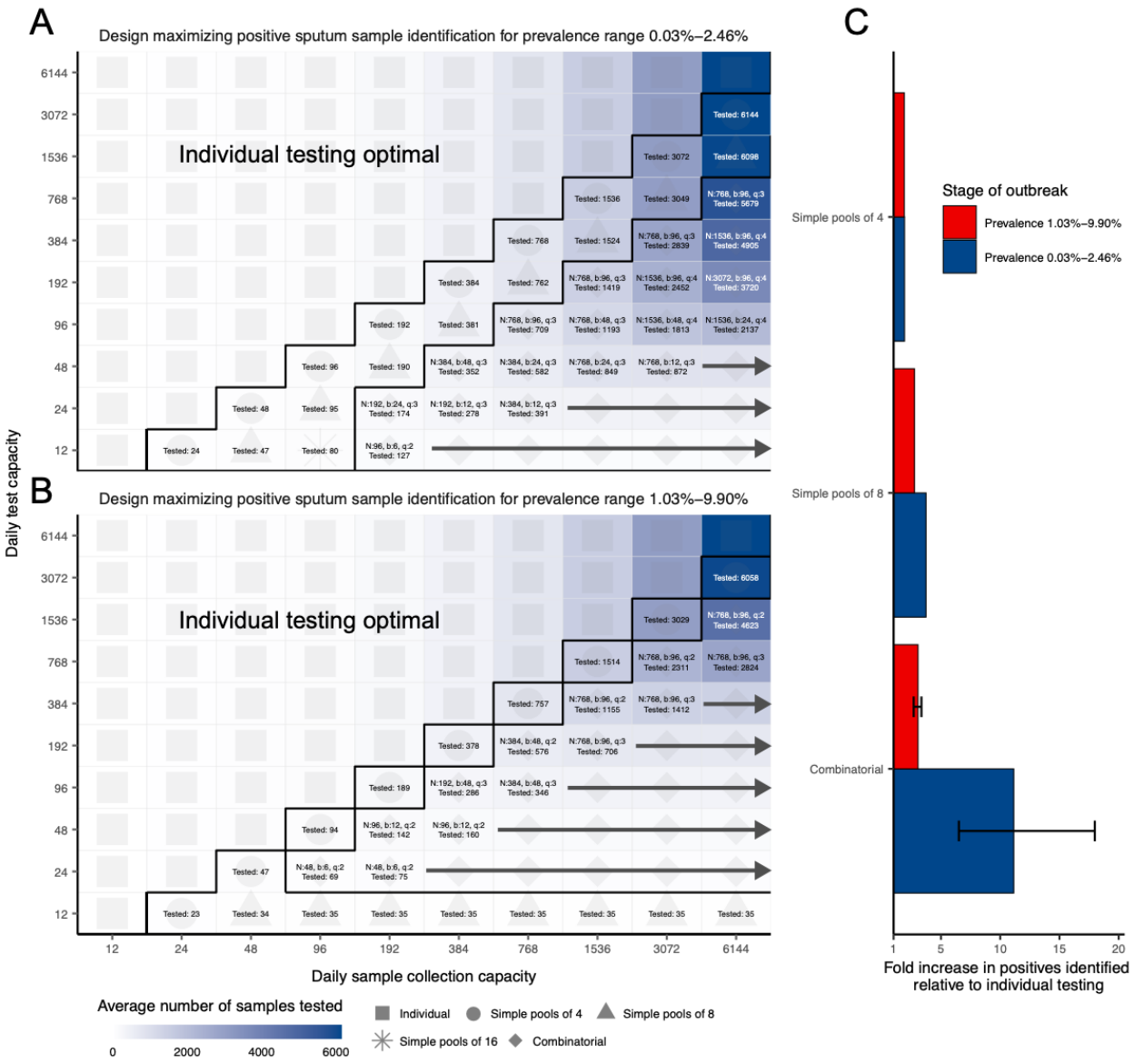


Fig. S6: Effectiveness of optimal testing design under resource constraints at high prevalence. As in **Fig. 4**, every design was evaluated under constraints on the maximum number of samples collected (columns) and average number of reactions that can be run on a daily basis (rows), here from days 80-108. Text in each box indicates the optimal design for a given set of constraints (number of samples per batch (N), number of pools (b), number of pools into which each sample is split (n), average number of total samples screened per day). Color indicates the average number of samples screened on a daily basis using the optimal design. Arrows indicate that the same pooling design is optimal at higher sample collection capacities.



983

984 **Fig. S7:** Effectiveness of optimal testing design under resource constraints using sputum data.
985 **(A)** and **(B)** show optimal designs at lower and higher prevalence ranges respectively. Text in
986 each box indicates the optimal design for a given set of constraints (number of samples per
987 batch (N), number of pools (b), number of pools into which each sample is split (n), average
988 number of total samples screened per day). Color indicates the average number of samples
989 screened on a daily basis using the optimal design. Arrows indicate that the same pooling
990 design is optimal at higher sample collection capacities. **(C)** Fold increase in the number of

991 positive samples identified relative to individual testing with the same resource constraints.

992 Error bar shows range amongst optimal designs.

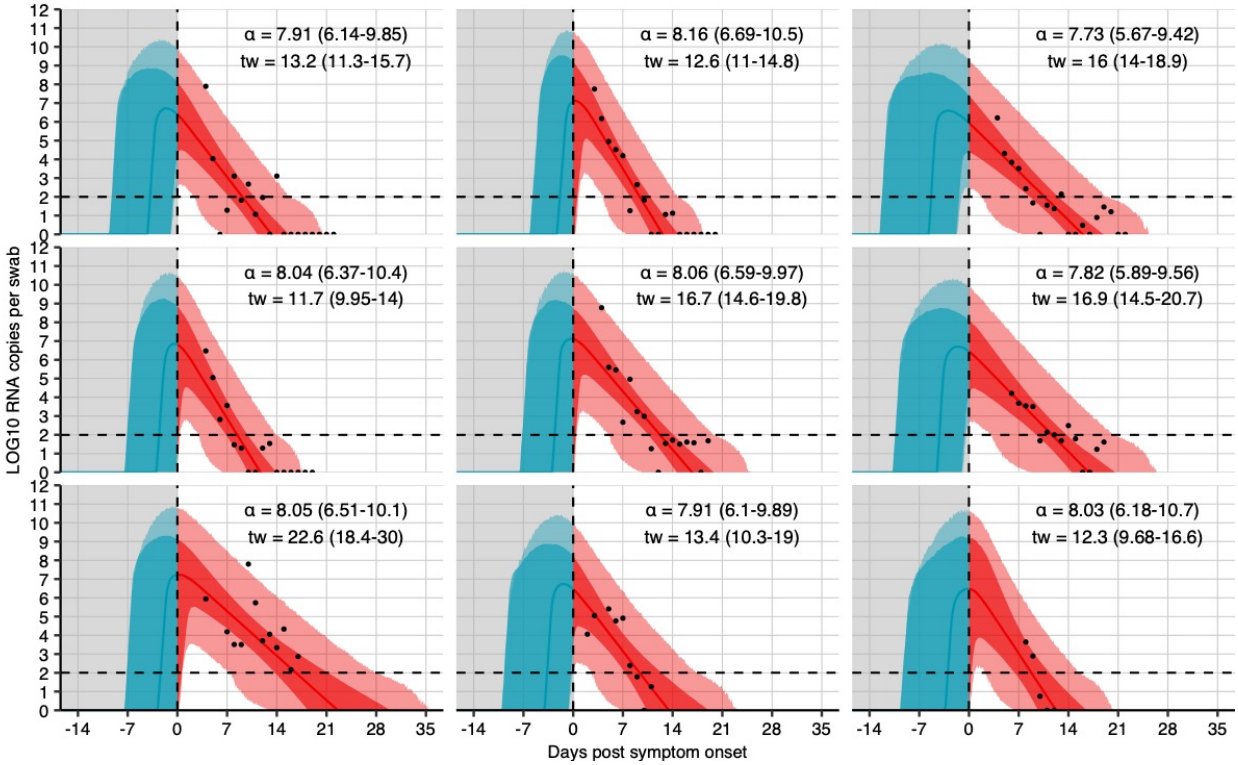
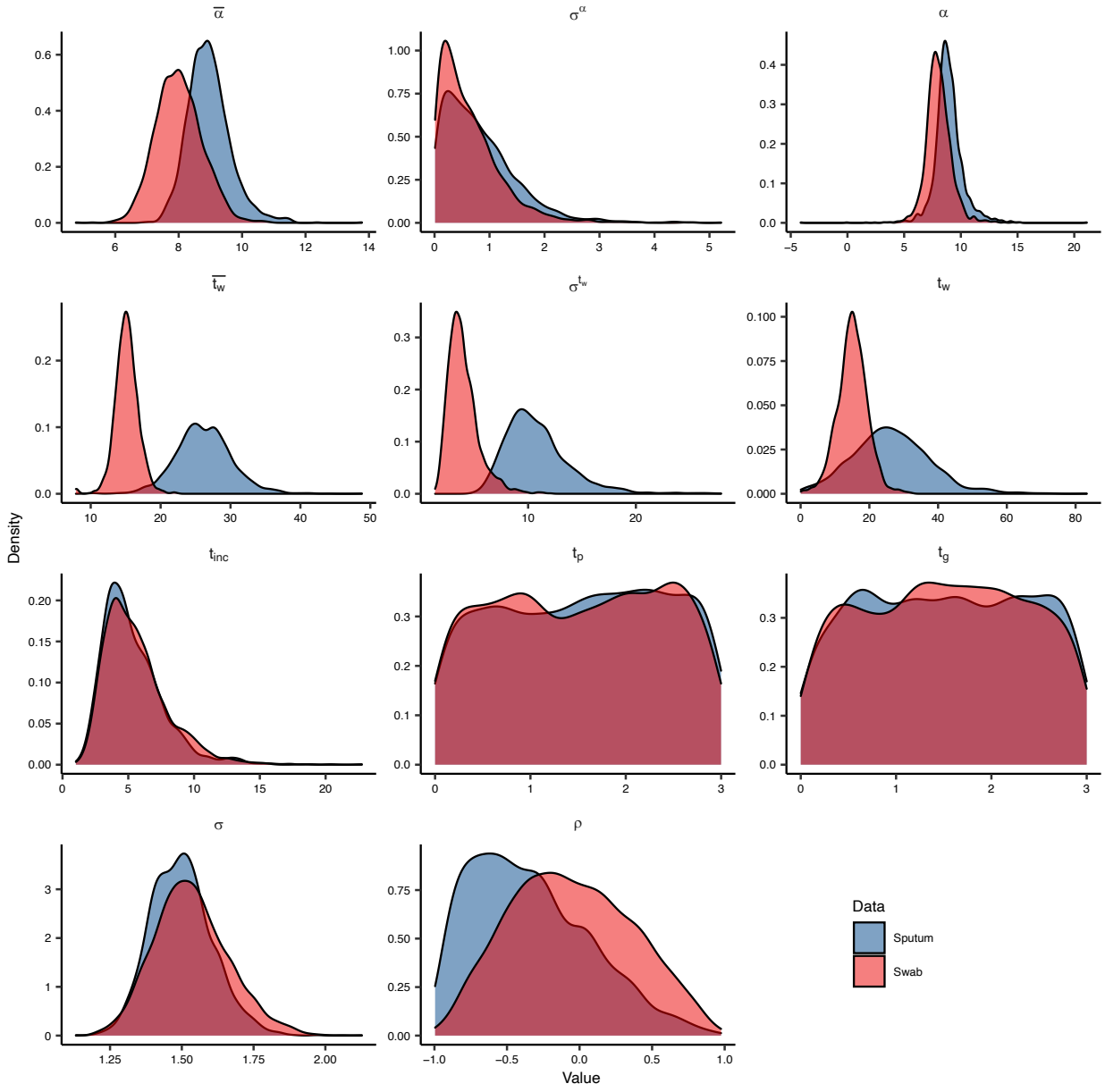


Fig. S8: Model fits to swab viral loads. Data were extracted from Wölfel et al. 2020. (34) The black dots show observed *log10* RNA copies/swab; solid lines show posterior median estimates; dark shaded regions show 95% credible intervals (CI) on model-predicted latent viral loads; light shaded regions show 95% CI on simulated viral loads with added observation noise. The vertical dashed line shows the time of symptom onset. The horizontal dashed line shows the limit of quantification and the y-axis shows limit of detection reported by Wölfel et al. Inset text shows posterior median and 95% credible intervals of the fitted parameters.



1002

1003 **Fig. S9:** Posterior distributions of estimated parameters fitted to swab and sputum data.

1004 Parameters are described in **Table S5**.

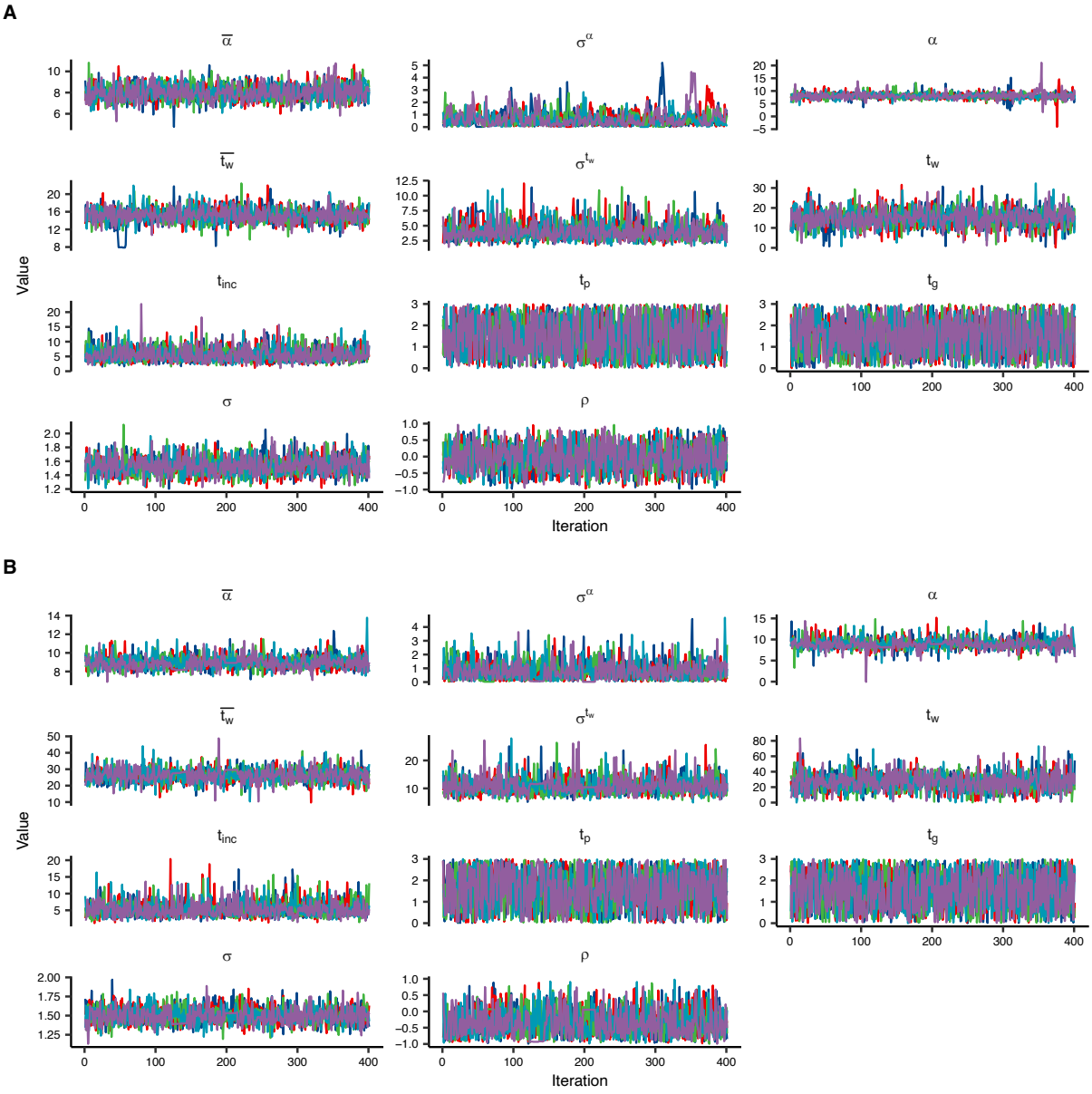


Fig. S10: Markov chain Monte Carlo trace plots from fitting to swab and sputum data. Each color shows an independent chain. Trace plots showing the path of the Markov chain Monte Carlo sampler, demonstrating convergence to the same stationary distribution for all chains. Parameters are described in **Table S5**. **(A)** Swab data. **(B)** Sputum data.

1010 **Table S1.** List of all group test designs for sample identification.

Number of pools	Number of samples	Number of pools per sample
1	2	1
1	4	1
1	8	1
1	16	1
1	32	1
1	64	1
6	12	2
6	12	3
6	24	2
6	24	3
6	48	2
6	48	3
6	96	2
6	96	3
6	192	2
6	192	3
6	384	2
6	384	3
12	24	2
12	24	3
12	24	4
12	48	2
12	48	3
12	48	4
12	96	2
12	96	3
12	96	4
12	192	2
12	192	3
12	192	4
12	384	2
12	384	3
12	384	4
12	768	2
12	768	3
12	768	4
24	48	2

24	48	3
24	48	4
24	96	2
24	96	3
24	96	4
24	192	2
24	192	3
24	192	4
24	384	2
24	384	3
24	384	4
24	768	2
24	768	3
24	768	4
24	1536	2
24	1536	3
24	1536	4
48	96	2
48	96	3
48	96	4
48	192	2
48	192	3
48	192	4
48	384	2
48	384	3
48	384	4
48	768	2
48	768	3
48	768	4
48	1536	2
48	1536	3
48	1536	4
48	3072	2
48	3072	3
48	3072	4
96	192	2
96	192	3
96	192	4
96	384	2
96	384	3
96	384	4
96	768	2

96	768	3
96	768	4
96	1536	2
96	1536	3
96	1536	4
96	3072	2
96	3072	3
96	3072	4
96	6144	2
96	6144	3
96	6144	4

1011

Table S2: Cycle threshold values from qPCR on pooled samples with variable viral load. Five positive samples with variable viral loads (89,000, 12,300, 1,280, 140, and 11, for samples 1-5, respectively) were each pooled with 23 negative samples ("Dil_1" to "Dil_5"). Values in bottom table show the Ct values for each of five pools for three primer pairs (N1, N2, and RP), along with a final call (positive, negative, or inconclusive).

Sample Name	N1 CT	N2 CT	RP CT	Call
24_Dil_1	28.93	29.58	28.62	POS
24_Dil_2	31.07	31.73	28.71	POS
24_Dil_3	33.29	33.91	28.37	POS
24_Dil_4	Undetermined	35.9	29.17	Inconclusive
24_Dil_5	Undetermined	Undetermined	28.62	Negative

Pool_1	30.49	31.18	28.3	POS
Pool_2	31.49	31.93	28.75	POS
Pool_3	Undetermined	Undetermined	29.62	Negative
Pool_4	31.71	31.66	29.34	POS
Pool_5	Undetermined	Undetermined	30.01	Negative
Pool_6	Undetermined	Undetermined	29.99	Negative

Table S3: Positive sample distribution within validation pools. 30 positive samples were randomly distributed across batches of combinatorial pooling (B1-B10) and randomly assigned sample numbers within batches. 24 separate positive samples were randomly assigned to 1 of 48 simple pools ("Prevalence", pool number 2...48).

Positive sample id	Batch	Sample number within batch
1	B5	46
2	B1	77
3	B5	21
4	B10	35
5	B2	10
6	B8	89
7	B1	5
8	B7	17
9	B6	6
10	B1	8
11	B7	91
12	B8	79
13	B2	20
14	B10	53
15	B8	30
16	B6	15
17	B5	50
18	B5	67
19	B9	7
20	B1	47
21	B9	6
22	B4	66
23	B2	46
24	B5	78
25	B2	96
26	B8	72

27	B3	83
28	B10	39
29	B8	20
30	B7	24
Positive sample id	**Batch**	**Pool number**
31	Prevalence	2
32	Prevalence	3
33	Prevalence	6
34	Prevalence	7
35	Prevalence	12
36	Prevalence	12
37	Prevalence	16
38	Prevalence	16
39	Prevalence	18
40	Prevalence	22
41	Prevalence	23
42	Prevalence	25
43	Prevalence	26
44	Prevalence	26
45	Prevalence	28
46	Prevalence	29
47	Prevalence	29
48	Prevalence	30
49	Prevalence	35
50	Prevalence	37
51	Prevalence	38
52	Prevalence	38
53	Prevalence	40
54	Prevalence	48

1022

Table S4: Pool design for combinatorial test with 96 samples. Each of 96 samples was split in equal volume into 2 out of 6 pools.

Sample	Pools	
1	A	B
2	A	B
3	A	B
4	A	B
5	A	B
6	A	B
7	A	B
8	C	D
9	C	D
10	C	D
11	C	D
12	C	D
13	C	D
14	C	D
15	E	F
16	E	F
17	E	F
18	E	F
19	E	F
20	E	F
21	E	F
22	B	C
23	B	C
24	B	C
25	B	C
26	B	C
27	B	C
28	B	C
29	D	F
30	D	F
31	D	F
32	D	F
33	D	F
34	D	F
35	D	F
36	A	E
37	A	E

38	A	E
39	A	E
40	A	E
41	A	E
42	A	E
43	B	D
44	B	D
45	B	D
46	B	D
47	B	D
48	B	D
49	A	F
50	A	F
51	A	F
52	A	F
53	A	F
54	A	F
55	C	E
56	C	E
57	C	E
58	C	E
59	C	E
60	C	E
61	B	E
62	B	E
63	B	E
64	B	E
65	B	E
66	B	E
67	C	F
68	C	F
69	C	F
70	C	F
71	C	F
72	C	F
73	A	D
74	A	D
75	A	D
76	A	D
77	A	D
78	A	D
79	B	F

80	B	F
81	B	F
82	B	F
83	B	F
84	B	F
85	D	E
86	D	E
87	D	E
88	D	E
89	D	E
90	D	E
91	A	C
92	A	C
93	A	C
94	A	C
95	A	C
96	A	C

1025

Table S5: Description of all parameters used in the viral kinetics and transmission models.

Values shown are posterior median estimates and 95% credible intervals or fixed value.

Parameter	Description	Posterior mean (95% CI)	Prior (parameters)
Viral kinetics model (swab)			
$\overline{\alpha}$	Mean of peak viral load parameters	7.98 (6.66-9.47)	Uniform(-100,100)
σ^{α}	Standard deviation of peak viral load parameters	0.498 (0.0254-2.13)	Half-Cauchy (1)
α_i	Individual peak viral load parameters	Varied	Multivariate normal
$\overline{t_w}$	Mean time from peak viral load to undetectable	15.1 (12-18.4)	Uniform(-100,100)
σ^{tw}	Standard deviation of viral waning parameters	3.72 (2.02-7.55)	Half-Cauchy (1)
t_w	Individual time from peak viral load to undetectable	Varied	Multivariate normal
t_{inc}	Time from infection to symptom onset	5.52 (2.23-11.5)	Log normal (1.621, 0.418)
t_g	Duration of pre-viral growth latent period	Varied	Uniform(0,3)
t_p	Time from growth initiation to peak viral load	Varied	Uniform(0,3)
σ	Standard deviation of observation process	1.52 (1.31-1.82)	Uniform(0,10)
ρ	Correlation between $\overline{\alpha}$ and $\overline{t_w}$	-0.0808 (-0.79-0.765)	LKJ correlation matrix (η=2)
Viral kinetics model (sputum)			
$\overline{\alpha}$	Mean of peak viral load parameters	8.91 (7.8-10.5)	Uniform(-100,100)
σ^{α}	Standard deviation of peak viral load parameters	0.674 (0.0417-2.32)	Half-Cauchy (1)
α_i	Individual peak viral load parameters	Varied	Multivariate normal
$\overline{t_w}$	Mean time from peak viral load to undetectable	26.0 (18.2-34.4)	Uniform(-100,100)
σ^{tw}	Standard deviation of viral waning parameters	10.4 (6.51-18.2)	Half-Cauchy (1)
t_w	Individual time from peak viral load to undetectable	Varied	Multivariate normal
t_{inc}	Time from infection to symptom onset	5.52 (2.23-11.5)	Log normal (1.621, 0.418)
t_g	Duration of pre-viral growth latent period	Varied	Uniform(0,3)
t_p	Time from growth initiation to peak viral load	Varied	Uniform(0,3)
σ	Standard deviation of observation process	1.49 (1.31-1.73)	Uniform(0,10)
ρ	Correlation between $\overline{\alpha}$ and $\overline{t_w}$	-0.398 (-0.925-0.579)	LKJ correlation matrix (η=2)
SEIR Model			
R_0	Basic reproductive number	2.5 (fixed)	NA
$1/\gamma$	Infectious period	7 days (fixed)	NA
$1/\sigma$	Incubation period	6.4 days (fixed)	NA
I_0	Initial number of infected individuals	100 (fixed)	NA
N	Total population size	12500000 (fixed)	NA

Table S6: RT-qPCR results for pooling validations. 10 batches of combinatorial pools (B1-B10, pools A-F in each batch) and 19 simple pools ("Prevalence", pools P2...P48) were tested by RT-qPCR. Test results (POS, NEG, INCONCLUSIVE, or INVALID) and Ct values for N1, N2, and RP primers are shown for each pool.

Batch	PoolID	Result	N1 Ct	N2 Ct	RP Ct
B1	A	POS	29.39	30.37	31.36
B1	B	POS	21.32	23.31	30.48
B1	C	POS	32.76	35.59	32.12
B1	D	POS	25.39	27.54	31.34
B1	E	NEG	Undetermined	Undetermined	32.47
B1	F	NEG	Undetermined	Undetermined	32.55
B2	A	POS	20.37	21.34	30.26
B2	B	POS	17.19	18.52	31.24
B2	C	POS	27.22	27.15	30.95
B2	D	POS	22.61	23.93	31.7
B2	E	POS	26.63	28.73	32.57
B2	F	INCONCLUSIVE	Undetermined	39.46	32.5
B3	A	NEG	Undetermined	Undetermined	32.45
B3	B	NEG	Undetermined	Undetermined	32.51
B3	C	NEG	Undetermined	Undetermined	32.37
B3	D	NEG	Undetermined	Undetermined	32.88
B3	E	NEG	Undetermined	Undetermined	31.97
B3	F	NEG	Undetermined	Undetermined	33.08
B4	A	NEG	Undetermined	Undetermined	31.31
B4	B	POS	23.46	25	30.38
B4	C	NEG	Undetermined	Undetermined	32.72
B4	D	NEG	Undetermined	Undetermined	32.26
B4	E	POS	23.52	24.98	30.63
B4	F	NEG	Undetermined	Undetermined	32.65
B5	A	POS	28.68	30.06	30.56
B5	B	POS	31.23	33.89	31.51

B5	C	NEG	Undetermined	Undetermined	31.92
B5	D	POS	27.17	28.71	31.3
B5	E	NEG	Undetermined	Undetermined	32.12
B5	F	NEG	Undetermined	Undetermined	30.86
B6	A	POS	32.09	32.96	30.56
B6	B	POS	30.3	32.43	30.22
B6	C	INVALID	Undetermined	Undetermined	Undetermined
B6	D	NEG	Undetermined	Undetermined	31.76
B6	E	POS	32.86	30.42	32.41
B6	F	NEG	Undetermined	Undetermined	32.17
B7	A	POS	22.44	22.61	31.35
B7	B	POS	25.52	27.41	31.98
B7	C	POS	29.26	29.22	31.37
B7	D	NEG	Undetermined	Undetermined	32.12
B7	E	POS	25.09	26.34	31.28
B7	F	INCONCLUSIVE	Undetermined	35.75	31.71
B8	A	NEG	Undetermined	Undetermined	30.74
B8	B	NEG	Undetermined	Undetermined	31.59
B8	C	INCONCLUSIVE	Undetermined	33.89	32.4
B8	D	POS	17.85	18.76	31.24
B8	E	POS	17.71	18.18	30.57
B8	F	POS	24.73	25.65	31.05
B9	A	POS	22.44	22.76	31.27
B9	B	POS	21.17	21.96	32
B9	C	NEG	Undetermined	Undetermined	32.88
B9	D	NEG	Undetermined	Undetermined	32.04
B9	E	NEG	Undetermined	Undetermined	31.93
B9	F	POS	32.67	34.41	32.69
B10	A	POS	23.36	23.73	30.19
B10	B	NEG	Undetermined	Undetermined	31.94
B10	C	NEG	Undetermined	Undetermined	32.56
B10	D	POS	27.69	26.08	32
B10	E	POS	23.34	24.5	32.56

B10	F	POS	27.15	25.63	31.56
Prevalence	P2	POS	21.68	23.21	32.06
Prevalence	P3	POS	33.31	34.66	32.95
Prevalence	P6	POS	22.49	23.25	31.85
Prevalence	P7	POS	20.39	21.43	31.51
Prevalence	P12	POS	23.54	24.93	30.93
Prevalence	P16	POS	21.46	23.02	31.97
Prevalence	P18	POS	19.55	19.63	31.24
Prevalence	P22	POS	22.22	23.61	31.42
Prevalence	P23	POS	34.54	35.93	32.34
Prevalence	P25	INCONCLUSIVE	Undetermined	36.76	32.39
Prevalence	P26	POS	31.53	33.14	32.49
Prevalence	P28	NEG	Undetermined	Undetermined	31.93
Prevalence	P29	POS	21.2	22.61	31.64
Prevalence	P30	NEG	Undetermined	Undetermined	34.92
Prevalence	P35	POS	21.65	22.79	32.19
Prevalence	P37	POS	26.83	28.97	32.36
Prevalence	P38	POS	25.26	26.24	31.01
Prevalence	P40	POS	18.35	19.65	32.31
Prevalence	P48	POS	33.09	34.09	32.79

1033