Epidemics 43 (2023) 100690

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Optimal capacity sharing for global genomic surveillance

Zsombor Z. Méder [a,*], Robert Somogyi [b,c]

[a] Institute of Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, Netherlands
[b] Faculty of Economic and Social Sciences, Budapest University of Technology and Economic. Muegyetem rkp. 3., H-1111 Budapest, Hungary
[c] Institute of Economics, Centre for Economic and Regional Studies, Toth Kalman utca 4, H-1097 Budapest, Hungary

ARTICLE INFO

Dataset link: https://osf.io/vk4x9/

Keywords:
Capacity constraint
Genomic surveillance
Optimal allocation
International cooperation

ABSTRACT

Recent technological advances and substantial cost reductions have made the genomic surveillance of pathogens during pandemics feasible. Our paper focuses on full genome sequencing as a tool that can serve two goals: the estimation of variant prevalences, and the identification of new variants. Assuming that capacity constraints limit the number of samples that can be sequenced, we solve for the optimal distribution of these capacities among countries. Our results show that if the principal goal of sequencing is prevalence estimation, then the optimal capacity distribution is less than proportional to the weights (e.g., sizes) of countries. If, however, the main aim of sequencing is the detection of new variants, capacities should be allocated to countries or regions that have the most infections. Applying our results to the sequencing of SARS-CoV-2 in 2021, we provide a comparison between the observed and a suggested optimal capacity distribution worldwide and in the EU. We believe that following such quantifiable guidance will increase the efficiency of genomic surveillance for pandemics.

1. Introduction

The SARS-CoV-2 pandemic has increased the scope and magnitude of genomic surveillance. By October 2022, more than 13.2 million genomic sequences of SARS-CoV-2 isolates have been shared through the Global Initiative on Sharing All Influenza Data genomic data repository (GISAID, 2022), originally established to track influenza variants. This international effort engendered by the pandemic allowed researchers to identify and characterize emerging mutations of the virus. However, the share of isolates sequenced presents large inequalities, especially between high/middle-income countries and developing nations (Mestanza et al., 2022; Chen et al., 2022; Crawford and Williams, 2021; Brito et al., 2021; Shey et al., 2020).

In this paper, we build a model that can guide global genomic surveillance strategy. Given the total available sequencing capacity, we derive how many samples should be sequenced in each country. To the best of our knowledge, the optimal distribution of sequencing capacity among countries has not yet been addressed by a formal model. We show how this optimal distribution depends on the prevalent number of infections, and the relative importance of policy goals.

Both health organizations (ECDC, 2021a,b,c,d; WHO, 2021a,b) and experts (Gardy et al., 2015; Gardy and Loman, 2018; Priesemann et al., 2021; Robishaw et al., 2021) have consistently urged countries to strengthen their efforts in genomic surveillance. The European Commission asks EU Member States to sequence at least 5%, and preferably 10% of all SARS-CoV-2 positive test results (EC, 2021). Similarly, in September 2021, the WHO asked African countries to attain a 5% sequencing rate (WHO, 2021c). However, the source of this 5% threshold is typically unspecified. Furthermore, based on simulations relying on Danish data, Vavrek et al. (2021) show that 5% sampling of all positive tests allows the detection of emerging strains when they have a prevalence of 0.1% to 1.0%. However, their model takes into account only variant detection as the goal of sequencing. It also ignores the international aspect of sequencing efforts, which is the main focus of our paper.

Other experts have called for increased international cooperation in the domain of genomic sequencing (Lancet, 2021; Crawford and Williams, 2021; Grubaugh et al., 2021). We aim to reinforce their arguments by building a model that quantifies the advantages gained therefrom, and by providing specific recommendations on how cooperation may maximize these benefits. Our model's contributions are twofold. First, we explicitly identify two goals of sequencing, and show that they lead to different optimal capacity allocations. Second, we give specific guidelines for optimal distribution of sequencing capacity, based on the weights assigned to various goals. We demonstrate that, contrary to existing recommendations, it is generically suboptimal to sequence the same share of isolates in every country. Our model is general in terms of the pathogen concerned, and we believe it can

[*] Corresponding author.
E-mail addresses: z.z.meder@fsw.leidenuniv.nl (Z.Z. Méder), somogyi.robert@gtk.bme.hu (R. Somogyi).

https://doi.org/10.1016/j.epidem.2023.100690
Received 5 December 2022; Received in revised form 10 March 2023; Accepted 21 May 2023
Available online 26 May 2023

provide guidance for genomic surveillance beyond SARS-CoV-2, for future pandemics.

Our paper proceeds as follows. In Section 2, we construct a model of sequencing capacity allocation step by step, by considering two main goals of sequencing first separately, and then jointly. In Section 3, we apply the model to derive optimal sequencing capacity distribution in a global context, and within the European Union. Section 4 concludes and reflects on possible extensions of our framework.

2. A model of sequencing capacity allocation

Variant sequencing is an essential tool for epidemiology for a number of reasons. Based on the literature, we classify these reasons in three classes. First, it provides information on current variant prevalence as a guide for control measures. For example, Brito et al. (2021), Crawford and Williams (2021) and Nadon et al. (2022) emphasize this goal.

Second, it allows the identification and characterization of new mutations as they emerge (see e.g. Burki, 2021; Duarte et al., 2021, 2022; Furuse, 2021; Grubaugh et al., 2021). The WHO uses information gained from sequencing to classify pathogen variants as Emerging Variants or Variants of Concern. Longitudinally, sequencing enables determining the mutation rates of various infectious agents. Sequencing may also be necessary to determine whether mutated pathogens have the ability to escape antibodies or vaccines. Combining sequencing data also enables the construction of phylogenetic trees.

Third, relatively rarely mentioned in the literature, it enables the analysis of transmission networks both between and within species (Quick et al., 2016). In this paper, we focus on the first two objectives, and ignore the third.

The only study we are aware of that takes account of these same two objectives (variant prevalence and variant detection) is Wohl et al. (2022). Their framework aims to calculate appropriate sample sizes for sequencing-based surveillance studies. It ignores, however, the aspect of international cooperation, which is the focus of this study.

We emphasize that our model aims to provide a short-term perspective for the analysis of optimal capacity allocation. This implies that sequencing capacities are fixed at a certain level. Further, our short-term approach allows us to abstract away from the problem of the timeliness of variant identification/detection. It also means that we can ignore the complexities of modeling virus transmission dynamics.

2.1. Goal 1: Estimating variant prevalence

We begin our analysis by focusing on estimating variant prevalence. Assume each country has a capacity constraint of K^j for full genome sequencing that determines the maximum number of sequenced samples, for a total available capacity of $K = \sum_j K^j$.[1] For parsimony, we ignore any costs related to the transportation of the isolates between countries. Suppose the genome of k^j virus-positive samples are sequenced. Within these samples, d_1^j are found to be of variant 1, and $d_2^j = k^j - d_1^j$ of variant 2. The best estimate regarding the prevalence of variant i is given by the sample mean, $\frac{d_i^j}{k^j}$. The actual prevalence for this variant in country j is denoted by p_i^j.

The government of each country is interested in identifying the true variant ratio in the population so that they may adjust public health policy accordingly. For example, consider that one variant poses severe epidemiological risk, while the other does not. If the more risky variant is widespread, optimal response requires strong measures to limit social contacts, strict lock-downs, etc. In contrast, if the less risky variant dominates, the best public health policy may be relatively lenient, forgoing the costs of limiting economic and social activity. Deviation in either direction may be costly for society. This is the reason behind the government's objective of identifying the true variant ratio.

One-country model. For simplicity, we assume that the objective function of the government is to minimize the quadratic difference between the estimated and the actual variant ratios. We call this difference the 'mistake function'. Specifically, country j's decision problem is captured as:[2]

$$\min_{k^j} m(k^j) \equiv E\left[\left(p_1^j - \frac{d_1^j}{k^j}\right)^2\right], \text{ subject to } k^j \leq K^j.$$

The decision variable of the government is the number of samples sequenced k^j. We assume that sequencing within the capacity constraint is free.

Since $E[d_1^j] = k^j p_1^j$, we get that the variance of d_1^j equals $E\left[\left(k^j p_1^j - d_1^j\right)^2\right]$. Therefore, the objective function can be written in terms of $Var(d_1^j)$ and k^j:

$$m(k^j) = E\left[\left(p_1^j - \frac{d_1^j}{k^j}\right)^2\right] = \frac{1}{(k^j)^2} E\left[\left(k^j p_1^j - d_1^j\right)^2\right] = \frac{Var(d_1^j)}{(k^j)^2}$$

As a first approximation of this function, assume that d_1^j is described by a binomial distribution with parameters k^j (number of trials) and p_1^j (probability of finding variant 1 in each trial). This assumes sampling with replacement, the outcome of each trial being independent. In this case, d_1^j has variance $k^j p_1^j(1 - p_1^j)$, and we get:

$$m(k^j) = \frac{p_1^j(1 - p_1^j)}{k^j}.$$

This mistake function is decreasing in k^j, and thus, more sequencing leads to more accurate estimates of the distribution of variants within the infected, and a public policy more adapted to the epidemiological situation. Another key property of the mistake function is that it is convex in the capacity k^j. In other words, the informational benefit of each additional sequenced sample is strictly decreasing. Convexity has two important consequences. First, as Fig. 7 in the Appendix shows, there is an optimal, finite number of samples to be sequenced in case sequencing is costly. Second, in the two-country model, convexity will be key to determining the optimal allocation of sequencing capacity between countries. In Appendix A, we show that these main results hold under the more realistic assumption of sampling without replacement as well.

Finally, as it is apparent from Fig. 1, the closer the distribution of prevalences is to fifty-fifty, the higher the mistake for any given amount sequenced. This result will carry over to the two-country model, and will be discussed in more detail below.

Two-country model. We analyze the optimal allocation of sequencing capacity between two countries (A and B) from the perspective of a social planner. While we show in Appendix B that the results of our model carry over to the case of three or more countries, for expositional simplicity, we here adopt the two-country perspective.

The social planner aims to allocate the total sequencing capacity of K in a way that minimizes the weighted sum of mistakes:

$$\min_{k^A, k^B} E\left[w^A m(k^A) + w^B m(k^B)\right], \text{ subject to } k^A + k^B \leq K.$$

The intuitions behind this formula and the presence of weights are as follows. We assume that during a pandemic, the public health measures taken depend on the relative prevalence of more and less risky variants. The impact of these measures, however, is greater for 'larger' countries, since more people are affected, and the economic

[1] In our short-term perspective, capacity constraints are binding. In Appendix C, we relax this assumption, and instead assume a constant marginal cost for sequencing.

[2] $E[X]$ refers to the expected value of random variable X. It is straightforward to show that it does not matter the prevalence of which of the two variants is estimated.

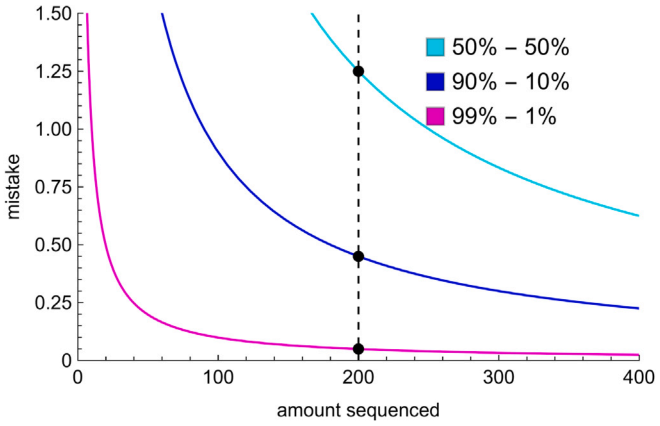

Fig. 1. Mistake (i.e., expected difference between the estimated and actual variant ratios) for three different actual variant ratios at a sequencing capacity of 200 samples. More extreme variant prevalences lead to fewer mistakes.

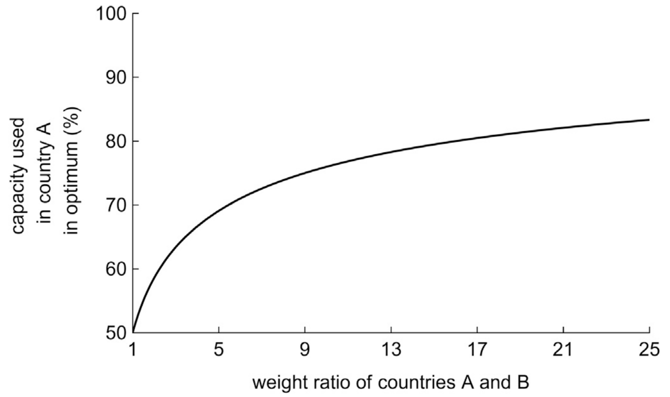

Fig. 2. Share of total available capacity (%) used for sequencing in country A (k_A/K) in optimum as a function of country A's relative weight. Country A is assumed to have a larger weight.

impact is also more significant. Some natural and convenient choices for the weights w^j would be the population size of a country, or the size of its economy. Broadly, how the weights should be chosen is a problem for moral philosophy that also involves intricate empirical considerations,[3] and is therefore beyond the scope of this paper. We believe that the population size of a country provides a good first approximation of appropriate weights to represent the preferences of a 'fair' social planner, and thus, in our empirical analysis in Section 3, we associate the weights w^j with countries' population sizes.

As before, we assume that the sequencing outcomes d_i^j follow a binomial distribution. The objective function of the social planner thus becomes:

$$\min_{k^A,k^B} w^A \frac{p_1^A(1-p_1^A)}{k^A} + w^B \frac{p_1^B(1-p_1^B)}{k^B}, \text{ subject to } k^A + k^B \leq K.$$

Standard optimization leads to:

$$k^A = \frac{K}{1 + \frac{\sqrt{w^B p_1^B(1-p_1^B)}}{\sqrt{w^A p_1^A(1-p_1^A)}}} \quad \text{and} \quad k^B = \frac{K}{1 + \frac{\sqrt{w^A p_1^A(1-p_1^A)}}{\sqrt{w^B p_1^B(1-p_1^B)}}}.$$

We examine the optimal solution by focusing on the ratio of optimal allocations $\frac{k^A}{k^B}$:

$$\frac{k^A}{k^B} = \sqrt{\frac{w^A}{w^B}} \cdot \sqrt{\frac{p_1^A(1-p_1^A)}{p_1^B(1-p_1^B)}}$$

We find that the optimal allocations are determined by two factors, namely, the relative weight of countries and the relative extremeness of prevalences.[4] We disentangle these two effects by assuming, first, equal extremeness and second, equal weights.

For equal prevalence of variants in the two countries ($p_1^A = p_1^B$), which implies equal extremeness, the optimal allocation of capacity simplifies to a square-root rule: $\frac{k^A}{k^B} = \sqrt{\frac{w^A}{w^B}}$, see Fig. 2. This has clear policy-relevant implications. Consider country weights to be chosen according to countries' population size. For example, the population of Spain is approximately *four* times larger than that of Portugal. Our model implies that in the optimal allocation of sequencing capacity, Spain should sequence only *twice* as many samples as Portugal. In general, for determining public policy based on variant prevalence, the optimal allocation of sequencing capacity between countries of uneven weights requires an allocation that is *less than proportional* to country weights.

For equal country weights ($w^A = w^B$), the optimal allocation of sequencing is a function of variant prevalences' extremeness. As $p_i^A(1-p_i^A)$ achieves its maximum at $p_i^A = 0.5$, countries with more evenly shared variants – i.e., where the prevalence of the two variants is closer to fifty-fifty – should receive a higher share of the total sequencing capacity. Fig. 3 illustrates the ratio of optimal allocations as a function of the variant prevalences in the two countries.

To consider the practical implications of our model, consider the following scenario. Initially, a certain variant is fully dominant in countries A and B. A new, more virulent variant appears in country A, starts spreading there first, and appears only later in B. This means that initially, extremeness in A, $p^A(1-p^A)$ will be larger than in B. Thus, initially, more sequencing should be done for isolates from country A, where the variant first appeared. Some time after the new variant takes over in A, reaching near-total prevalence ($p^A \sim 1$), extremeness in A will drop below that in B. From this point on, more sequencing capacity should be allocated for isolates from country B, where there is still epidemiological competition between the new and the old variants.

2.2. Goal 2: Identifying a new variant

So far we have assumed that the governments engage in sequencing in order to estimate the share of different variants among the infected as precisely as possible. In this section, we focus on another objective: detecting emerging variants. Indeed, one of the stated aims of sequencing is to identify emerging variants and potentially label them as "Variants of Interest (VOI)", "Variants of Concern (VOC)" or "Variants of High Consequence (VOHC)".[5] This classification requires full genome sequencing.

We assume that detecting a new variant early on, and classifying it correctly brings a fixed benefit U to society. Moreover, as such a discovery is shared almost instantly all over the world, its benefit is enjoyed by all countries.[6] The benefit incorporates (in monetary terms) all the expected benefits of future research, as well as the ability of governments to adapt to the new epidemiological situation.

[3] For example, it could be argued that a larger weight should be given to countries whose populations are more susceptible or vulnerable to a certain virus due to their demographic characteristics or the poor state of their healthcare system.

[4] Note that we have implicitly assumed the knowledge of the actual prevalences p_i^j to derive the optimal sequencing allocations. This assumption can, however, be relaxed: the p_i^j's in our results may represent expert opinions based on the best available data.

[5] https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.

[6] Such a benefit is a public good, as its consumption is non-excludable and non-rivalrous.

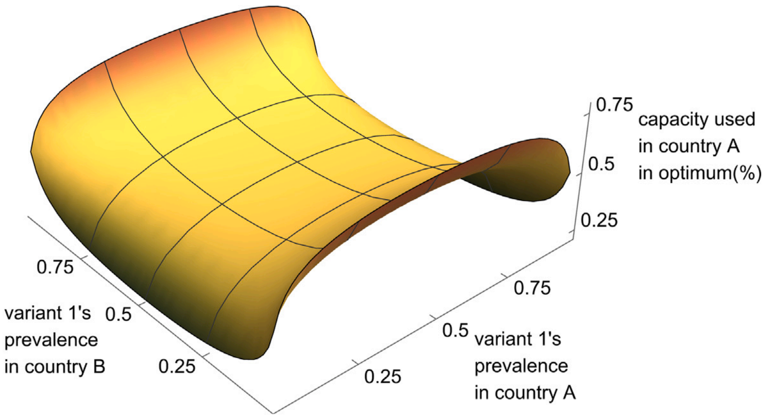

Fig. 3. Share of total available capacity (%) used for sequencing in country A (k_A/K) in optimum as a function of the prevalence of variant 1 in countries A and B. Countries with more extreme variant distributions require a lower share of the capacity.

Assume again that there are two countries A and B. A new variant that can out-compete the existing ones may emerge in either country A, or country B. We regard the possibility of two such variants emerging simultaneously to be vanishingly small. Let s denote the probability that such a new (mutant) variant does not emerge over a unit period of time in either country. If the time period is short, s is very close to one. Assuming that mutations appear randomly, and that the characteristics of the infected population do not differ between the countries, the conditional probability that the variant emerges in country j is proportional to the number of infected in that country, n^j. The probabilities of a new mutant arising in country A and B are thus given by:

$$(1-s)\frac{n^A}{n^A+n^B} \quad \text{and} \quad (1-s)\frac{n^B}{n^A+n^B},$$

respectively.

Suppose that a new mutant indeed emerges in country j. Let q denote the expected share of the new variant among all the infected after one unit of time. In other words, q is proportional to how fast the new variant spreads. Then, if k^j samples are sequenced, the probability that at least one of the mutant-containing samples is sequenced is $1-(1-q)^{k^j}$. Since q is small, this probability can be approximated by qk^j. Substituting in the objective function, we get:

$$\max_{k^A,k^B}\left[(1-s)\frac{n^A}{n^A+n^B}\cdot qk^A\cdot U+(1-s)\frac{n^B}{n^A+n^B}\cdot qk^B\cdot U\right],$$
$$\text{subject to } k^A+k^B\leq K.$$

Using that the total capacity constraint will be binding in optimum, i.e., $k^B=K-k^A$, the maximization problem can be simplified to:

$$\frac{(1-s)q\cdot U}{n^A+n^B}\max_{k^A}\left[(n^A-n^B)k^A+n^BK\right], \text{ subject to } 0\leq k^A\leq K.$$

This leads to a bang–bang solution: if the only objective of sequencing is the detection of new variants, it is optimal to allocate all the sequencing capacity to the country with the larger number of infections. Formally, the optimal number of sequencing in country A satisfies:

$$k^A=\begin{cases}0 & \text{if} \quad n^A<n^B;\\ K & \text{if} \quad n^A>n^B.\end{cases}$$

This result is in stark contrast with the recommendation derived from goal 1, i.e., when the objective is to estimate the prevalence of existing variants. Recall that under that objective, the optimal allocation of sequencing capacity is *less than proportional* to the relative size of countries. When the objective is to detect new variants, the optimal allocation of sequencing is *more than proportional*, in an extreme way: it is all-or-nothing.

2.3. Combined goals

Both choosing a policy that fits the epidemiological situation and detecting new variants are important when determining the allocation of sequencing capacity. In this subsection, we integrate these considerations within a unified, two-country framework.

With the notation of the preceding subsections, the decision problem becomes:

$$\min_{k^A,k^B}\left[w^Am(k^A)+w^Bm(k^B)-\frac{(1-s)q\cdot U}{n^A+n^B}\left(n^Ak^A+n^Bk^B\right)\right],$$
$$\text{subject to } k^A+k^B\leq K.$$

For given parameter values, this problem is solvable with standard numerical methods. To get a qualitative sense of the effects of various principal parameters on the optimal capacity allocation, we adopt two simplifying assumptions. First, we use the share of infected to calculate the weights, in particular, we let $w_A=\frac{n_A}{n_A+n_B}$, $w_B=\frac{n_B}{n_A+n_B}$.[7] Second, we assume that the variant shares are equal across countries, i.e., $p_1^A=p_1^B=p_1$. Let $v=\frac{(1-s)q\cdot U}{p_1(1-p_1)}$, representing the (relative) importance of identifying new variants. Indeed, the more likely mutations are, the faster they spread, and the greater the expected benefit associated with finding them, the larger v becomes; lower extremeness, on the other hand, implies a lower v. Ultimately, the social planner estimates the value of U, and thus, its preferences have a direct influence on the optimal solution by way of v, the importance of identifying new variants.

With these simplifications at hand, using $k^B=K-k^A$, the social planner's problem is equivalent to:

$$\min_{k^A}\left[\frac{n^A}{n^B}\left(\frac{1}{k^A}-vk^A\right)+\left(\frac{1}{K-k^A}-v(K-k^A)\right)\right]$$

Fig. 4A and B contrast the effect of parameter v on the optimal allocation when country A has twice or half as many infected as country B, respectively. In Appendix D, we show that the optimal allocation to country A is increasing in v if and only if country A has more infected than B. This makes intuitive sense, as parameter v captures the relative importance of finding a new variant. Higher values of v lead to a reallocation of the sequencing capacity to the country with more infected. With $v=0$, new mutations are completely irrelevant, and we get back our model from Section 2.1, and the optimal allocation of capacity will follow our square-root rule. Conversely, with very high values of v, we converge to the framework of Section 2.2, and the

[7] In Section 2.1, we argue for determining weights based on population size. However, in the context of this subsection, this would render it impossible to visualize our results.

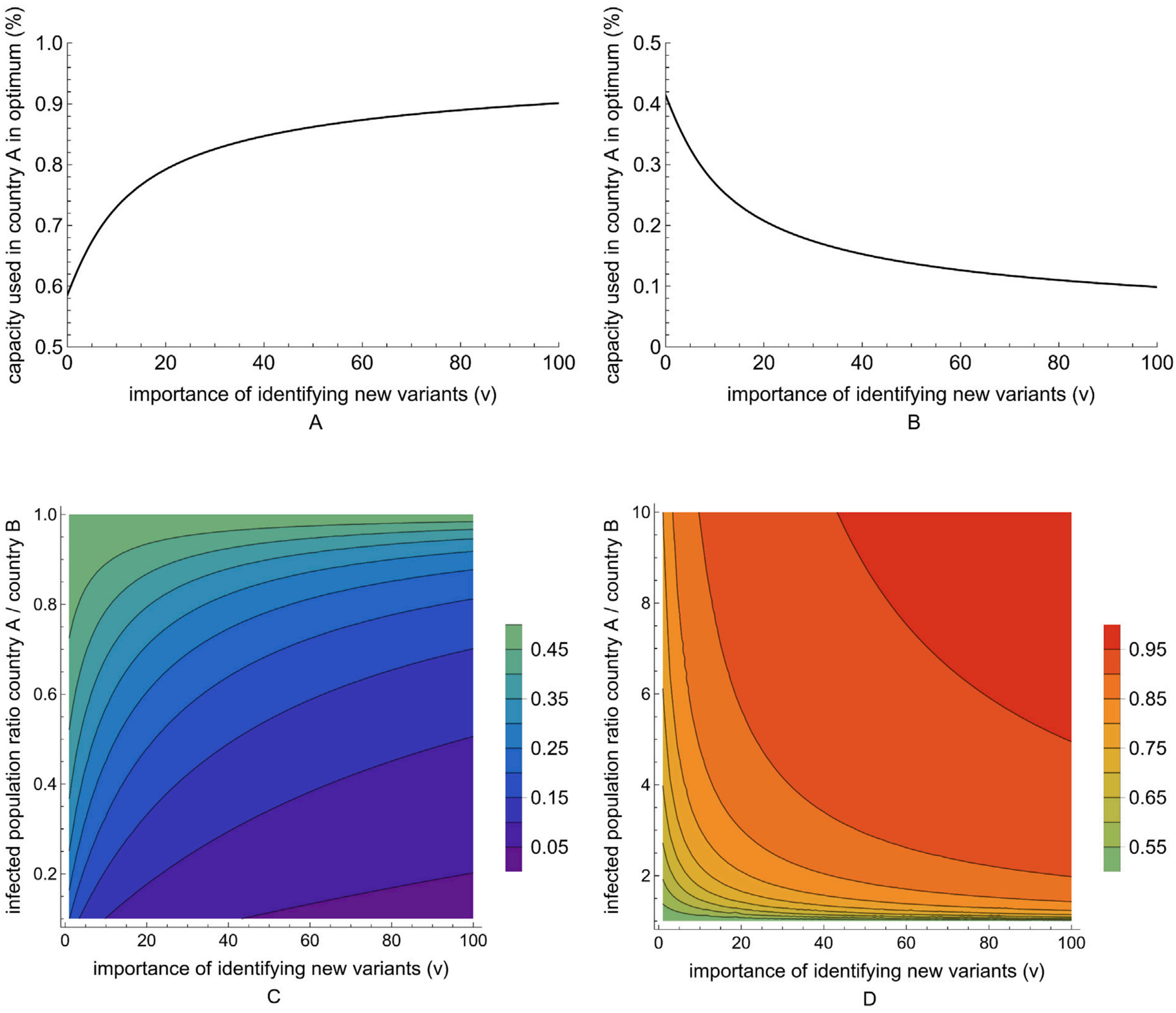

Fig. 4. Share of total available capacity (%) used for sequencing in country A in optimum when both estimating variant prevalence and identifying a new variant are important. Panels **A** and **B**: Share in optimum as a function of the relative importance of identifying a new variant when the number of infected in A are twice (**A**)/half (**B**) as many as in country B. Panels **C** and **D**: Share in optimum as a function of the relative importance of identifying a new variant and relative number of infected. Country A has more (**C**)/less (**D**) infected than B.

entire sequencing capacity will be allocated to the country counting more infected. Fig. 4C and D generalize these relationships to arbitrary n^A/n^B ratios.

3. Genomic surveillance in the case of SARS-CoV-2: Reality and opportunities

While our theoretical model is general, and its insights are applicable to any pathogen, in this section we adapt our results to the SARS-CoV-2 pandemic, based on data from 2021.[8]

According to our datasets, there were approximately 204 million SARS-CoV-2 infections worldwide in 2021. GISAID reports that nearly 6.33 million sequences were submitted to its database, which means that ∼3.1% of all positive samples were sequenced, falling somewhat short of the 5% recommendation of health agencies (ECDC, 2021b,c,d; WHO, 2021a,b). However, there are large inequalities in sequencing efforts, especially between developed countries and the global south, see Fig. 5, Panel A. Out of 183 countries in our dataset, 103 did not sequence even 1% of their sample pool, including, surprisingly, well-off countries such as Saudi Arabia, Taiwan or Cyprus. Only 27 countries managed to reach a sequencing rate of 5%.

In order to derive recommendations based on our model, we focus exclusively on Goal 1, i.e., identifying variant prevalence for public policy. This way, we avoid arbitrarily choosing the relative importance of the two goals. Further, we identify weights with the population size of each country. Fig. 5, Panel B shows the share of global sequencing capacity that should be dedicated to each country based on our model from Section 2.1. The contrast with the actual distribution is apparent.

We acknowledge that transportation costs, legal constraints, as well as other transaction costs may make the global cooperation required to reach the optimum difficult to achieve. Therefore, we next focus on capacity sharing within the European Union, where such hurdles should be easier to overcome. Fig. 6, Panel A shows the actual share

[8] Sequencing data was acquired from GISAID's online repository. Infection data was accessed from the COVID-19 Data Repository of the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (see Dong et al., 2020). We used population and GDP data from Wolfram Mathematica's servers. One limitation of this dataset is that submission to GISAID is voluntary, and it is conceivable that some countries have sequenced more than what appears in GISAID's database. We also assumed that all the genomic sequencing capacity was used for SARS-CoV-2 in 2021, which likely somewhat underestimates the true total capacity. Infection numbers are likely also low estimates, as there may have been significant numbers of unrecorded or unreported SARS-CoV-2 infections, especially in developing and autocratic countries. The datasets and the Wolfram Mathematica code used to derive our results are available at https://osf.io/vk4x9/.

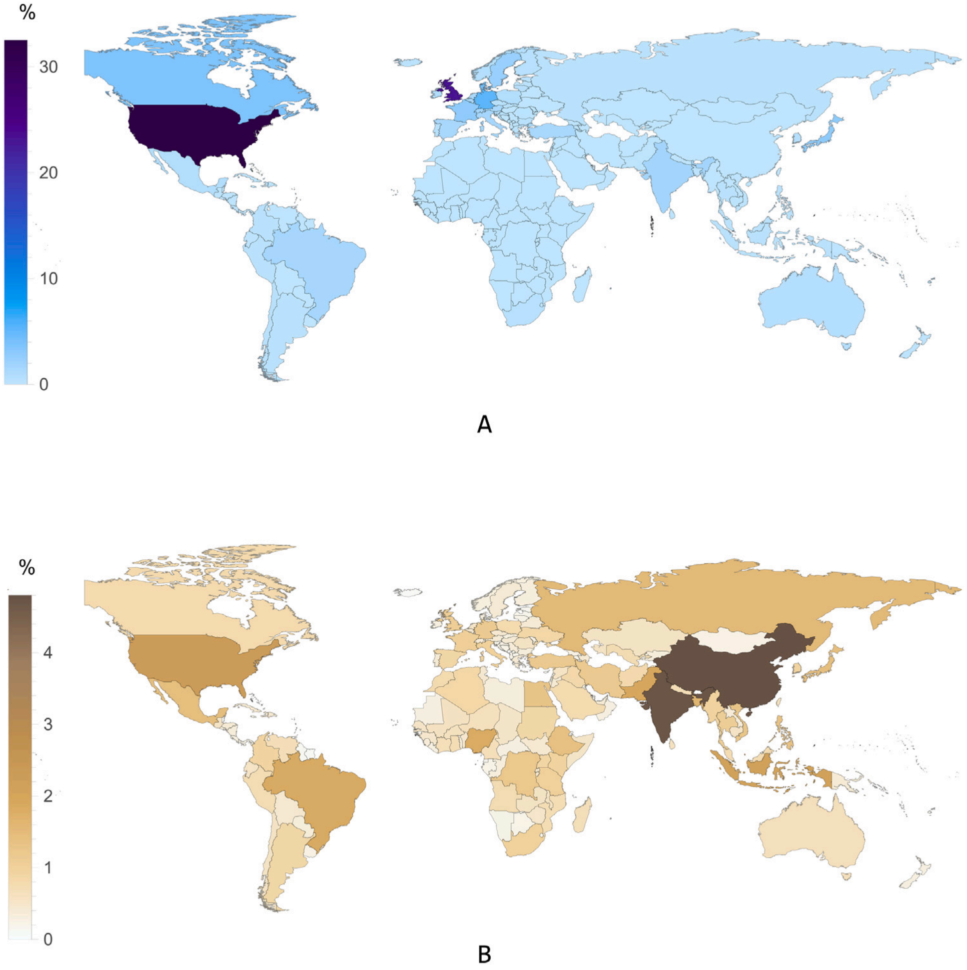

Fig. 5. SARS-CoV-2 sequencing worldwide in 2021. Panel **A**: Actual sequencing as a share of global available capacity. Panel **B**: Optimal sequencing allocation for estimating variant prevalence as a share of global available capacity. Country weights are determined by population size.

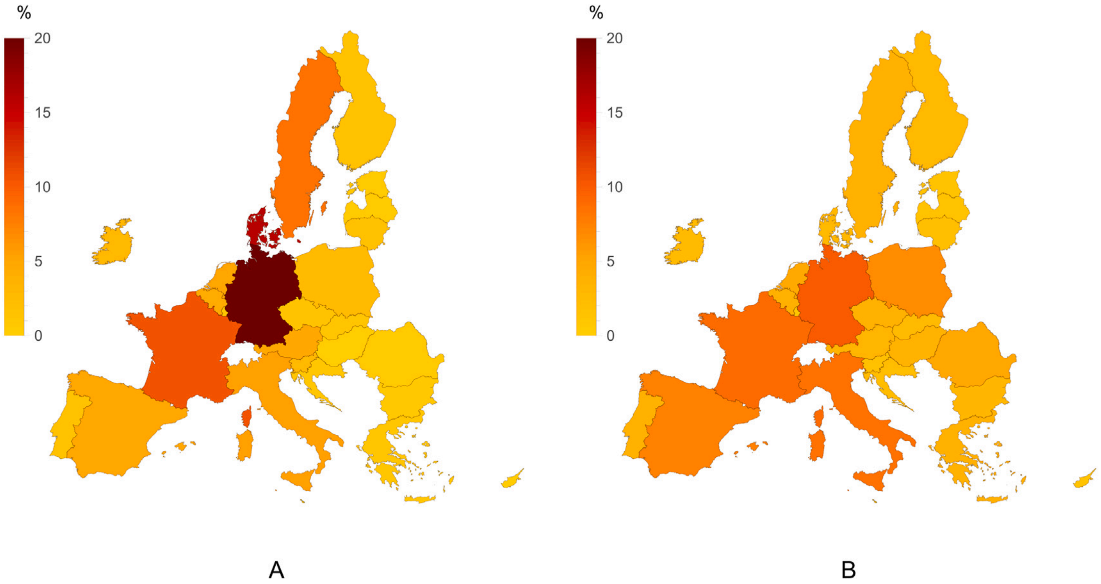

Fig. 6. SARS-CoV-2 sequencing in the European Union in 2021. Panel **A**: Actual sequencing as a share of available EU capacity. Panel **B**: Optimal sequencing allocation for estimating variant prevalence as a share of available EU capacity. Country weights are determined by population size.

Table 1
Genomic sequencing within the European Union in 2021. Second column shows the number of sequences submitted to GISAID. Third column indicates the desired sequencing amounts when the same share of all positive samples are analyzed in each country (3.99%). Fourth column represents the desired sequencing amounts for estimating variant prevalence when country weights are determined by population, and the total capacity is equal to the sequencing capacity of 2021. Last row shows the expected value of the objective function (i.e., the mistake to be minimized) under the different sequencing scenarios.

Country	Actually sequenced	Infection-prop. rule	Square-root rule
Austria	76 506	36 351	53 405
Belgium	73 530	58 073	60 569
Bulgaria	10 241	21 704	46 638
Croatia	14 117	20 057	35 879
Cyprus	754	5 728	19 580
Czechia	20 336	69 497	58 158
Denmark	261 394	25 675	42 819
Estonia	8 042	8 489	20 444
Finland	22 665	9 333	41 831
France	173 774	294 170	143 647
Germany	329 152	215 533	162 668
Greece	12 694	42 702	57 191
Hungary	163	37 112	55 122
Ireland	45 252	27 701	39 643
Italy	88 139	159 278	137 982
Latvia	6 443	9 369	24 265
Lithuania	25 656	15 008	29 126
Luxembourg	17 268	2 286	14 150
Malta	638	1 577	11 817
Netherlands	81 621	94 598	73 594
Poland	39 297	111 695	109 181
Portugal	22 199	38 622	56 629
Romania	8 515	46 739	77 670
Slovakia	17 638	43 284	41 500
Slovenia	44 525	13 555	25 605
Spain	78 705	174 032	121 420
Sweden	137 873	34 970	56 607
EV of obj. function	71 185	6 955	5 128

sequenced by countries in the EU, while Panel *B* represents the optimal distribution, based on the same assumptions as for Fig. 5. Finally, Table 1 compares the actual amounts sequenced with the recommendations of our model, as well as the infection-proportional sharing recommendation of the European Commission. Indeed, the Commission recommends sequencing 5% of all positive samples in each country (ECDC, 2021b,c,d). As the European Union only sequenced 3.99% of positive isolates collectively, indicating a capacity constraint even at the EU level, we use the 3.99% level for the infection-proportional sharing rule.

Three observations can be made based on Table 1. First, countries in the North and West of Europe over-perform, both compared to the 3.99% recommendation, and our proposed distribution; while countries in the South and East of the EU under-perform. There are some positive (Denmark) and negative (Hungary and Cyprus) outliers. Second, unsurprisingly, we find a positive correlation between this sequencing surplus or deficit and the logarithm of per capita GDP, $r = 0.56$. Third, both the infection-proportional and our proposed rule provide an order of magnitude of improvement value of the objective function over the current sequencing allocation. Moreover, our proposed allocation rule entails an improvement of more than 25% over the rule advocated by international health agencies.

4. Conclusion

Our paper provides a model of sequencing capacity sharing by specifying the two main goals of genomic surveillance: variant prevalence estimation and the identification of new pathogen variants. While Section 3 uses SARS-CoV-2 as a case study, our results are general, and do not depend on the type of the pathogen. Due to its novelty and until recently high cost, the principal uses of genomic surveillance were for influenza, Ebola, and SARS-CoV-2. Given the substantially increased probability of extreme epidemics due to environmental change (Marani et al., 2021), the relevance of finding optimal mechanisms for pathogen identification and control will ever increase. Indeed, there is a wide consensus regarding the importance of genomic surveillance for ending the health threat posed by SARS-CoV-2 (Lazarus et al., 2022).

An advantage of our model is that the optimal distribution of sequencing takes into account the relative importance of various public policy goals, which can be parametrized by the policy-maker. For example, in some contexts, only the identification of Variants of Concern (i.e., goal 2) may be policy-relevant. We can get policy recommendations for this scenario as a special case of our model.

One limitation of our model is that it assumes that the transportation of isolates between countries/sequencing centers is costless. When transportation is costly, the optimal distribution of total sequencing capacity will be closer to each country's individual capacity. This point holds not only for financial, but also temporal costs. The more important timeliness of detection is, the less international capacity sharing improves social outcomes. We hope that future work on the problem can address these issues more directly.

Another, related limitation of our framework is that we also ignore other transaction costs, such as legal and political constraints on the international transport of pathogen isolates. Anecdotal evidence suggests that these can create important barriers for international cooperation for genomic surveillance. However, if such barriers are present, the geographic domain of optimal capacity redistribution can be adjusted to the appropriate set within which these barriers are not present, or are manageable (e.g., EU, or NAFTA). Moreover, our model can also be adapted to solve capacity distribution *within* a country, e.g., considering the states of the U.S. or Germany, or the provinces of Canada or China.

We also abstract away from the complexity arising from each country pursuing its self-interest. Instead, our goal is to explore the theoretical maximum of gains on a collective level. A full game-theoretic analysis of these problems is beyond the scope of this paper.

Health experts have already highlighted the necessity of large-scale international cooperation in the efforts to track and control pandemics. Our work quantifies the gains that could be realized from such cooperation. We believe that instead of genomic autarky – i.e., each country focusing its sequencing efforts to infections within its borders –, sequencing capacity sharing can improve outcomes for all parties, especially in the short run. In the long run, countries should aim for building up their sequencing capacities. However, for countries with limited material and human resources, and especially those that do not currently engage in genomic sequencing, this may take a significant amount of time.

In our view, the identification of new pathogen variants, especially variants of concern, should be treated as a global public good. In other words, capacity sharing has positive externalities, and thus, genomic sequencing potentially benefits everyone in the world. Thus, countries that contribute more to global sequencing efforts should not be penalized for identifying new variants, such as has been the case for South Africa for identifying the first Omicron variant. Furthermore, countries with larger capacities should sequence isolates from their neighbors and regional partners.

CRediT authorship contribution statement

Zsombor Z. Méder: Conceived and designed the analysis, Collected the data, Performed the analysis, Wrote the paper. **Robert Somogyi:** Conceived and designed the analysis, Collected the data, Performed the analysis, Wrote the paper.

Declaration of competing interest

The authors declare no conflict of interest

Data availability

Data and code have been made available Open Science Initiative at https://osf.io/vk4x9/.

Acknowledgments

The authors are grateful to Zsuzsa Bakk, Carsten K. W. de Dreu, Guilhem Lecouteux, Ronald Peeters, Olivér Rácz, Luca Sandrini, and Katalin Tamási for comments on early drafts of the manuscript, and to Gábor Kemenesi for his input and advice.

Somogyi thanks the support of the Hungarian National Research Development and Innovation Office (NKFIH) under grant number OTKA FK-142492.

This project has received funding from the Spinoza Award from The Netherlands Science Foundation to Carsten K. W. de Dreu (NWO SPI-57-242).

Appendix A. Sampling without replacement

While the binomial distribution provides a mathematically convenient tool to capture the objective function, considering sampling *without* replacement provides a better approximation of the outcomes of variant sequencing. We can thus assume that d_1^j follows a hypergeometric distribution with parameters N^j (population size), v_1^j (number of infected by variant 1), k^j (number of trials). Then the objective function becomes:

$$m(k^j) = \frac{1}{(k^j)^2} k^j p_1^j (1 - p_1^j) \frac{N^j - k^j}{N^j - 1} = \frac{p_1^j (1 - p_1^j)}{N^j - 1} \left(\frac{N^j}{k^j} - 1 \right),$$

which is also decreasing in k^j. Similar to the binomial distribution, this also yields a convex mistake function. While the hypergeometric distribution would describe the sampling problem more rigorously, in the main text, we decided to deal with the binomial distribution, due to its mathematical convenience.

Appendix B. More than two countries.

Next, we show that the main results in Section 2.1 generalize to the case of more than two countries. The social planner aims to allocate the total sequencing capacity of $K = \sum_{j=1}^{n} K^j$ in a way that minimizes the weighted sum of mistakes of countries $j \in \{1, 2..n\}$:

$$\min_{k^1, k^2, \ldots, k^n} E \left[\sum_{j=1}^{n} w^j m(k^j) \right] = E \left[\sum_{j=1}^{n} w^j \frac{p_1^j (1 - p_1^j)}{k^j} \right],$$

$$\text{subject to } \sum_{i=1}^{n} k^i \leq K.$$

Given that in optimum, all capacity is allocated, we can write the Lagrangian as follows:

$$L(k^1, k^2..k^n) = E \left[\sum_{j=1}^{n} w^j \frac{p_1^j (1 - p_1^j)}{k^j} \right] - \lambda \left(K - \sum_{j=1}^{n} k^j \right).$$

The optimal allocation must, for all countries j, satisfy:

$$0 = \frac{\partial L(k^1, k^2..k^n)}{\partial k^j} = \lambda - \frac{w^j p_1^j (1 - p_1^j)}{(k^j)^2}.$$

Therefore, for any pair of countries A and B, at the optimal allocation:

$$\lambda = \frac{w^A p_1^A (1 - p_1^A)}{(k^A)^2} = \frac{w^B p_1^B (1 - p_1^B)}{(k^B)^2}.$$

Rearranging the equation above, we get that for any A and B:

$$\frac{k^A}{k^B} = \sqrt{\frac{w^A}{w^B}} \cdot \sqrt{\frac{p_1^A (1 - p_1^A)}{p_1^B (1 - p_1^B)}},$$

which generalizes the statement in the main text. □

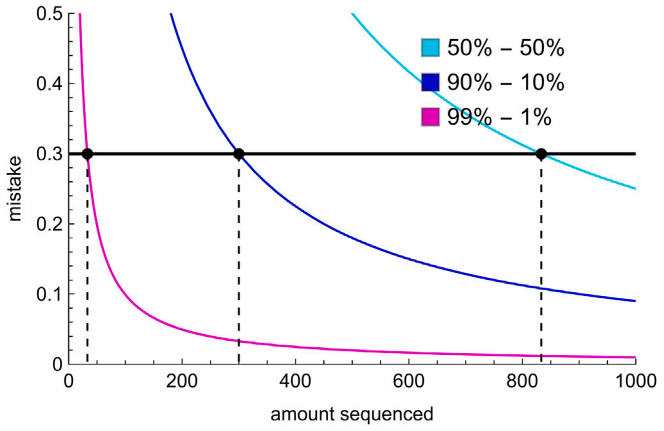

Fig. 7. Optimal amount sequenced for three different actual variant ratios at a constant marginal cost of sequencing of 0.3 relative to a unit cost of mistake.

Appendix C. Costs instead of capacity constraints.

We here show that a model of estimating variant prevalence that includes costs instead of capacity constraints leads to results that are closely analogous to those with our model of capacity constraints.

C.1. One-country case

A single country j aims at estimating variant prevalence. Assume that instead of a capacity constraint, sequencing has a unit cost of c^j, with the cost being expressed relative to the cost of making a unit of mistake in estimating the prevalence. The decision variable is q^j, the number of samples to be sequenced. The decision problem is thus:

$$\min_{q^j} m(q^j) + c^j q^j = E \left[\frac{p_1^j (1 - p_1^j)}{q^j} + c^j q^j \right].$$

Differentiating, we find that the optimal amount is given by:

$$q^j = \frac{\sqrt{p_1^j (1 - p_1^j)}}{\sqrt{c^j}}.$$

Since p_1^j is unknown, its value in the formula for q^j should be substituted by expert estimate. See Fig. 7 for the optimal amount sequenced at various levels of extremeness of prevalences.

C.2. Two, or more countries.

Multiple countries estimate variant prevalence, each with a country weight of w^j, and sequencing unit costs of c^j, deciding on q^j. The social planner aims to minimize the sum of mistakes and costs:

$$\min_{q^1, q^2, \ldots, q^n} E \left[\sum_{j=1}^{n} w^j m(q^j) + c^j q^j \right] = E \left[\sum_{j=1}^{n} w^j \frac{p_1^j (1 - p_1^j)}{q^j} + c^j q^j \right].$$

The objective function is separable in the j's, and for the optimal amounts, we get:

$$q^j = \frac{\sqrt{w^j} \sqrt{p_1^j (1 - p_1^j)}}{\sqrt{c^j}}.$$

For any two countries A and B, we get the following ratio of optimally sequenced quantities:

$$\frac{q^A}{q^B} = \sqrt{\frac{c^B}{c^A}} \cdot \sqrt{\frac{w^A}{w^B}} \cdot \sqrt{\frac{p_1^A (1 - p_1^A)}{p_1^B (1 - p_1^B)}}.$$

If costs are equal, i.e., $c^A = c^B$, we get back the formula derived under capacity constraints in Section 2.1. If sequencing costs differ between countries, their impact is again *less than proportional*, and follows a square-root rule.

Appendix D. Relationship between the relative importance of the two policy goals, the number of infected, and the optimal allocation.

We show that the optimal allocation to country A is increasing in v if and only if country A has more infected than B when considering both goals in Section 2.3. Mathematically, we need to prove that in the optimal allocation denoted k^* given by:

$$\begin{aligned} k^* &= \arg\min_{k^A} f(k^A) \\ &= \arg\min_{k^A} \left[\frac{n^A}{n^B} \left(\frac{1}{k^A} - v k^A \right) + \left(\frac{1}{K - k^A} - v(K - k^A) \right) \right] \end{aligned}$$

satisfies:

$$\frac{\partial k^*}{\partial v} > 0 \text{ if and only if } n_A > n_B.$$

First, we show that the objective function $f(k^A)$ is strictly convex. Indeed, straightforward calculations lead to:

$$f''(k^A) = 2 \left(\frac{n^A/n^B}{(k^A)^3} + \frac{1}{(K - k_A)^3} \right) > 0.$$

Thus the following first-order condition is sufficient for global optimality:

$$f'(k^*) = 0 = \frac{n^A}{n^B} \left(-\frac{1}{(k^*)^2} - v \right) - \frac{1}{(K - k^*)^2} - v.$$

Applying the implicit function theorem to the above equation:

$$\frac{\partial k^*}{\partial v} = -\frac{\frac{\partial f'(k^*)}{\partial v}}{\frac{\partial f'(k^*)}{\partial k^*}} = -\frac{1 - n^A/n^B}{f''(k^*)}.$$

Using that f'' is strictly positive, we conclude that

$$\frac{\partial k^*}{\partial v} > 0 \iff \frac{n^A}{n^B} > 1. \quad \square$$

References

Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U., … Danish, 2021. Covid-19 Genome Consortium. Global disparities in SARS-CoV-2 genomic surveillance., medrxiv.

Burki, T., 2021. Understanding variants of SARS-CoV-2. Lancet 397 (10273), 462.

Chen, Z., Azman, A.S., Chen, X., Zou, J., Tian, Y., Sun, R., . Yu, H., 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nature Genet. 54 (4), 499–507.

Crawford, D.C., Williams, S.M., 2021. Global variation in sequencing impedes SARS-CoV-2 surveillance. PLoS Genet. 17 (7), e1009620.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534.

Duarte, C.M., Jamil, T., Gojobori, T., Alam, I., 2021. Detection of SARS-CoV-2 variants requires urgent global coordination. Int. J. Infect. Dis. 109, 50–53.

Duarte, C.M., Ketcheson, D.I., Eguíluz, V.M., Agustí, S., Fernández-Gracia, J., Jamil, T., . Alam, I., 2022. Rapid evolution of SARS-CoV-2 challenges human defenses. Sci. Rep. 12 (1), 1–8.

European Commission, 2021. Communication from the Commission To the European Parliament, the European Council and the Council. a United Front To Beat COVID-19. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021DC0035. [Available online].

European Centre for Disease Control (ECDC), 2021a. Technical Report: Sequencing of SARS-CoV-2 – First Update. https://www.ecdc.europa.eu/sites/default/files/documents/sequencing-of-SARS-CoV-2.pdf. [Available online].

European Centre for Disease Control (ECDC), 2021b. Technical Report: Detection and Characterisation Capability and Capacity for SARS-CoV-2 Variants Within the EU/EEA. https://www.ecdc.europa.eu/sites/default/files/documents/Detection-and-characterisation-capability-for-SARS-CoV-2-variants-EU%20EEA.pdf. [Available online].

European Centre for Disease Control (ECDC), 2021c. Technical Report: Guidance for Representative and Targeted Genomic SARS-CoV-2 Monitoring. https://www.ecdc.europa.eu/sites/default/files/documents/Guidance-for-representative-and-targeted-genomic-SARS-CoV-2-monitoring-updated-with%20erratum-20-May-2021.pdf. [Available online].

European Centre for Disease Control (ECDC), 2021d. Technical Report: Methods for the Detection and Characterisation Capability and Capacity for SARS-CoV-2 Variants – First Update. https://www.ecdc.europa.eu/sites/default/files/documents/Methods-for-the%20detection-and-characterisation-of-SARS-CoV-2-variants-first-update-WHO-20-Dec-2021.pdf. [Available online].

Furuse, Y., 2021. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. Int. J. Infect. Dis. 103, 305–307.

Gardy, J.L., Loman, N.J., 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. Nature Rev. Genet. 19 (1), 9–20.

Gardy, J., Loman, N.J., Rambaut, A., 2015. Real-time digital pathogen surveillance – the time is now. Genome Biol. 16 (1), 1–3.

GISAID–Initiative, 2022. Global initiative on sharing all influenza data website. https://www.gisaid.org. Accessed September 25, 2022.

Grubaugh, N.D., Hodcroft, E.B., Fauver, J.R., Phelan, A.L., Cevik, M., 2021. Public health actions to control new SARS-CoV-2 variants. Cell 184 (5), 1127–1132.

Lancet, 2021. Genomic sequencing in pandemics. Lancet 397 (445).

Lazarus, J.V., Romero, D., Kopka, C.J., et al., 2022. A multinational delphi consensus to end the COVID-19 public health threat. Nature http://dx.doi.org/10.1038/s41586-022-05398-2, [Available online].

Marani, M., Katul, G.G., Pan, W.K., Parolari, A.J., 2021. Intensity and frequency of extreme novel epidemics. Proc. Natl. Acad. Sci. 118 (35), e2105482118.

Mestanza, O., Lizarraga, W., Padilla-Rojas, C., Jimenez-Vasquez, V., Hurtado, V., Molina, I.S., . Solari, L., 2022. Genomic surveillance of the lambda SARS-CoV-2 variant in a global phylogenetic context. J. Med. Virol. 94 (10), 4689–4695.

Nadon, C., Croxen, N., Knox, J., Tanner, A., Zetner, C., Yoshida, G., Van Domselaar, M., 2022. Public health genomics capacity assessment: readiness for large-scale pathogen genomic surveillance in Canada's public health laboratories. BMC Public Health 22 (1817).

Priesemann, V., Balling, R., Brinkmann, M.M., Ciesek, S., Czypionka, T., Eckerle, I., . Szczurek, E., 2021. An action plan for pan-European defence against new SARS-CoV-2 variants. Lancet 397 (10273), 469–470.

Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., . Carroll, M.W., 2016. Real-time, portable genome sequencing for ebola surveillance. Nature 530 (7589), 228–232.

Robishaw, J.D., Alter, S.M., Solano, J.J., Shih, R.D., DeMets, D.L., Maki, D.G., Hennekens, C.H., 2021. Genomic surveillance to combat COVID-19: challenges and opportunities. Lancet Microb. 2 (9), e481–e484.

Shey, M., Okeibunor, J.C., Yahaya, A.A., Herring, B.L., Tomori, O., Coulibaly, S.O., . Talisuna, A.O., 2020. Genome sequencing and the diagnosis of novel coronavirus (SARS-COV-2) in africa: how far are we? Pan Afr. Med. J. 36 (1).

Vavrek, D., Speroni, L., Curnow, K.J., Oberholzer, M., Moeder, V., Febbo, P.G., 2021. Genomic surveillance at scale is required to detect newly emerging strains at an early timepoint. medRxiv.

World Health Organization, 2021a. Genomic sequencing of SARS-CoV-2 – A guide to implementation for maximum impact on public health. https://apps.who.int/iris/rest/bitstreams/1326052/retrieve. [Available online].

World Health Organization, 2021b. SARS-CoV-2 genomic sequencing for public health goals – interim guidance. https://apps.who.int/iris/rest/bitstreams/1326068/retrieve. [Available online].

World Health Organization, 2021c. Scaling up genomic sequencing in africa. https://www.afro.who.int/news/scaling-genomic-sequencing-africa. [Available online].

Wohl, S., Lee, E.C., DiPrete, B.L., Lessler, J., 2022. Sample size calculations for variant surveillance in the presence of biological and systematic biases. medRxiv, 2021–12.