Inferring the sensitivity of wastewater metagenomic sequencing for pathogen early detection

Simon L. Grimm[1,2,*], Jeff T. Kaufman[1,2,*], Daniel P. Rice[1,2], Charles Whittaker[3], William J. Bradshaw[1,2,✉], & Michael M. McLaren[1,2,✉]

[*] Contributed equally.

[1] Media Laboratory, Massachusetts Institute of Technology, Cambridge, United States

[2] SecureBio, Cambridge, United States

[3] MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, United Kingdom

[4] SecureDNA Foundation, Zug, Switzerland

[✉] Correspondence: wjbrad@mit.edu, mmclaren@mit.edu

Abstract

Detecting novel pathogens at an early stage requires robust early warning that is both sensitive and pathogen-agnostic. Wastewater metagenomic sequencing (W-MGS) could meet these goals, but its sensitivity and financial feasibility depend on the relative abundance of novel pathogen sequences in W-MGS data. Here we collate W-MGS data from a diverse range of studies to characterize the relative abundance of known viruses in wastewater. We then develop a Bayesian statistical model to integrate these data with epidemiological estimates for 13 human-infecting viruses, and use it to estimate the expected relative abundance of different viral pathogens for a given prevalence or incidence in the community. Our results reveal pronounced variation between studies, with estimates differing by one to three orders of magnitude for the same pathogen: for example, the expected relative abundance of SARS-CoV-2 at 1% weekly incidence varied between 10^{-7} and 10^{-10}. Integrating these estimates with a simple cost model highlights similarly wide inter-study and inter-pathogen variation in the cost of W-MGS-based early detection, with a mean yearly cost estimate of roughly $19,000 for a Norovirus-like pathogen and $2.9 million for a SARS-CoV-2-like pathogen at 1% incidence. The model and parameter estimates presented here represent an important resource for future investigation into the performance of wastewater MGS, and can be extended to incorporate new wastewater datasets as they become available.

Introduction

Biosurveillance of viruses present in wastewater has been used as a public health tool for decades[1]. After Poliovirus was first identified in sewage in 1939[2], the virus was frequently tracked in wastewater to aid eradication efforts[1], and polio wastewater surveillance remains an important tool for monitoring and suppressing reintroductions of the disease[3]. Similarly, wastewater surveillance of norovirus has enabled both advance warning[3] and phylogenetic analyses of norovirus evolution[4].

Nevertheless, until very recently, wastewater biosurveillance was largely restricted to a few specific pathogens in particular locales. This changed with the COVID-19 pandemic, when the need to reliably track SARS-CoV-2 led to a surge in wastewater surveillance efforts across the world[5]. Public health agencies and private companies employed qPCR to track the spread of COVID-19[6] and amplicon sequencing to identify and trace SARS-CoV-2 variants[7]. Monitoring efforts are now expanding to track other pathogens, such as Chikungunya[8] , Dengue virus[8], respiratory syncytial virus (RSV)[9] and influenza[9]. However, most such efforts remain targeted, using qPCR[9], amplicon sequencing[10], or other assays[11] to sensitively monitor for a defined list of target pathogens. Targeted methods thus limit the utility of wastewater surveillance for identifying novel pathogens.

In contrast, methods like untargeted metagenomic sequencing are in principle able to detect any nucleic acid present in a sample[12]. A wide diversity of enteric and non-enteric viruses have been found to shed via human feces into wastewater[13], including many pathogens that are not routine targets for wastewater monitoring at present. This broad swath of pathogens therefore presents a promising target for metagenomic sequencing. Additionally, wastewater-based MGS (W-MGS) could detect novel pathogens spreading asymptomatically, thus complementing traditional surveillance approaches like syndromic hospital surveillance.

However, the utility of W-MGS for pathogen surveillance will depend on its sensitivity as a detection tool. This is heavily dependent on the number of pathogen reads present in a given dataset, which in turn is driven by two key parameters: the sequencing effort (number of sequencing reads produced) and the relative abundance of the pathogen (fraction of all reads in the dataset that arise from that pathogen). To a first approximation, the sequencing effort and relative abundance required to achieve a given sensitivity are inversely proportional: the lower the relative abundance, the more sequencing needed to achieve detection. The cost-effectiveness of W-MGS surveillance thus depends critically on pathogen relative abundance in W-MGS data: even given rapidly falling sequencing costs[14], sufficiently low relative abundances would still make W-MGS cost-prohibitive as a surveillance tool.

To better understand the cost of W-MGS for pathogen surveillance, we need to understand the relationship between a disease's epidemiological status (in particular, its incidence or prevalence in the community) and pathogen relative abundance in wastewater. Such relative abundance estimates, however, have not previously been generated in any systematic manner. To address this important

Study	Sequencing Type	Country	Eligible Read Pairs	Included in integrated epidemiological analyses	Reference
Brinch et al. (2020)	DNA	Denmark	4B	Yes	[15]
Crits-Christoph et al. (2021)	RNA	United States	300M	Yes	[10]
Rothman et al. (2021)	RNA	United States	700M	Yes	[16]
Spurbeck et al. (2023)	RNA	United States	2B	Yes	[17]
Bengtsson-Palme et al. (2016)	DNA	Sweden	700M	No	[18]
Brumfield et al. (2022)	DNA, RNA	United States	500M	No	[19]
Maritz et. al. (2019)	DNA	United States	200M	No	[20]
Munk et al. (2022)	DNA	Multiple	40B	No	[21]
Ng et al. (2019)	DNA	Singapore	300M	No	[22]
Yang et al. (2021)	RNA	China	1B	No	[23]

Table 1: Studies included in Figure 1. “Eligible read pairs” refers to the number of Illumina read pairs obtained from raw wastewater using untargeted shotgun metagenomic sequencing, excluding other sample types and libraries that underwent specific target enrichment (e.g. with oligo probe panels).

gap in the biosurveillance literature, we developed a hierarchical Bayesian model that combines publicly available metagenomic datasets with epidemiological data for a range of known viral pathogens, allowing us to estimate the relative abundance of a given virus in wastewater at a given prevalence or incidence. Integrating these estimates with a simple sequencing cost model, we then calculated the cost of pathogen detection with W-MGS at a given target level of sensitivity.

Our results highlight pronounced variability in the relative abundance of different human-infecting viruses across studies, with corresponding variability in the predicted cost of detection with W-MGS. Under more optimistic assumptions, W-MGS appears a viable approach for pathogen detection in the medium-term, while under more pessimistic ones it faces severe obstacles to feasibility.

Results

Relative abundance of human-infecting viruses varies widely among wastewater studies

We performed a literature search for large (>100M read pairs), untargeted W-MGS datasets obtained from raw treatment-plant influent. We identified 10 such datasets (Table 1), obtained the data, and processed it with a Kraken2-based computational pipeline to estimate the relative abundance of human-infecting viruses (Methods). Figure 1a shows the relative abundance of different viral classes in each dataset, calculated as the fraction of all reads that map to the relevant taxa.

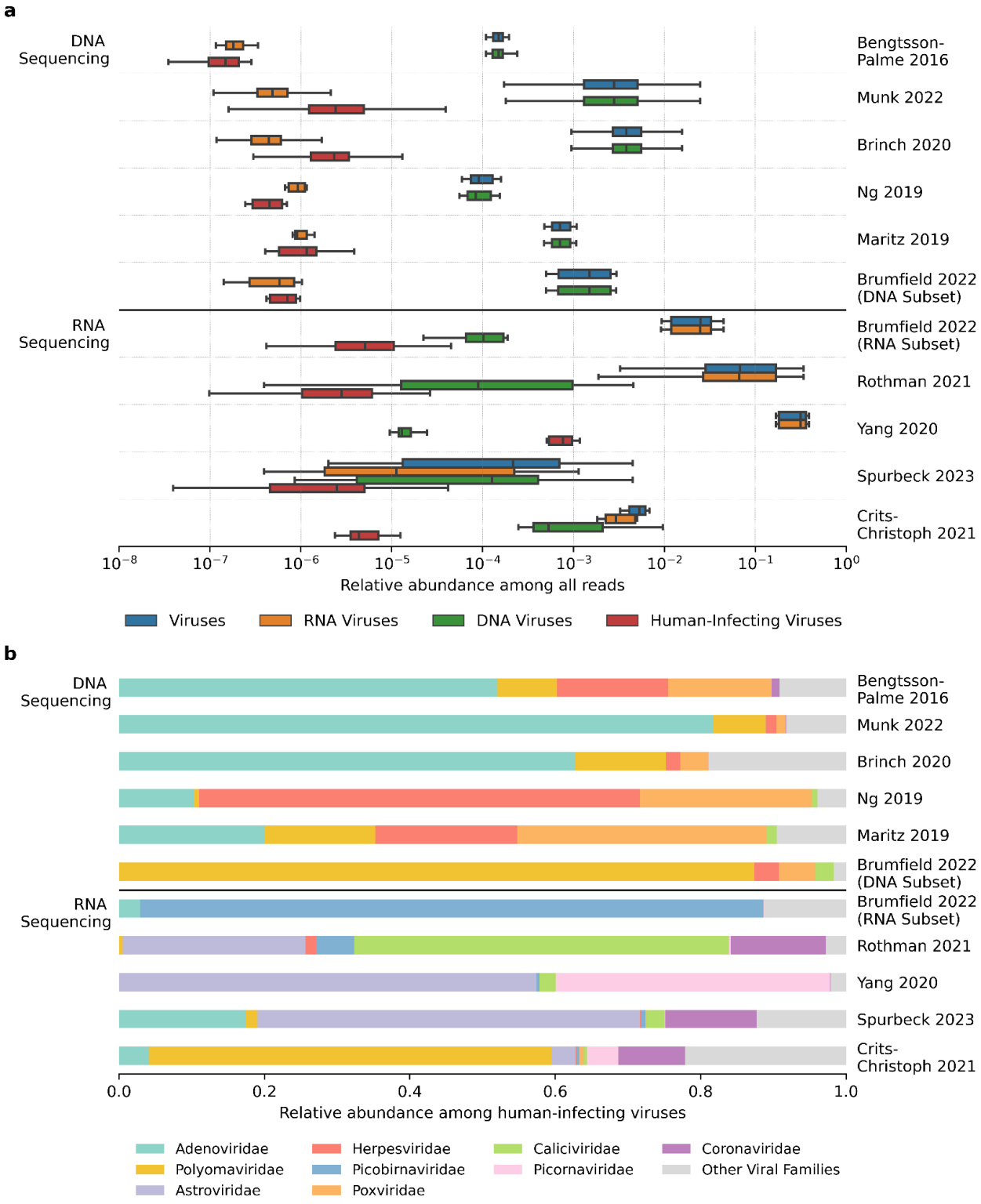

Figure 1: Viral abundance and taxonomic composition vary widely between wastewater metagenomics studies. **(a)** Relative abundance of all viruses (blue), RNA viruses (orange), human-infecting viruses (green), and DNA viruses (red), across samples for each included study. **(b)** Average taxonomic composition of human-infecting viruses in each study, displayed as the arithmetic mean relative abundance of viral families across samples.

The relative abundance of viruses as a whole varied by several orders of magnitude between studies, from roughly 1 in 4 reads (Yang et al. 2021[23]) to 1 in 11,000 (Ng et al. 2019[22]) (Table S1). Unsurprisingly, the relative abundance of RNA viruses was much higher in studies that conducted RNA sequencing, while that of DNA viruses was higher in studies that conducted DNA sequencing; however, the effect was much less dramatic for DNA than RNA viruses (Fig. S1), likely reflecting the presence of DNA virus transcripts in RNA data.

Relative abundance of human-infecting viruses also varied substantially between studies, from roughly 1 in 1,500 reads (Yang et al. 2021[23]) to 1 in 8.3 million (Bengtsson-Palme et al. 2016[18]) (Table S1). On average, human-infecting viruses accounted for roughly 1 in 440,000 reads across all studies (Table S1). The taxonomic composition of human-infecting viruses also varied dramatically (Fig. 1b); for example, the fraction of human-infecting virus reads mapping to Caliciviridae varied from 52% (Rothman et al. 2021[16]) to 3% (Spurbeck et al. 2023[17]), and Polyomaviridae relative abundance varied from 87% (Brumfield et al. 2022[19], DNA Subset) to 7% (Munk et al. 2022[21]).

A multi-level Bayesian model enables flexible combination of epidemiological and metagenomic data

These large inter-study differences in viral abundance and composition could arise from a variety of factors, including differences in catchment demographics, sewershed hydrology, and sample processing methods. One especially critical factor, however, is the number of infected people contributing to the sewershed. We therefore collated public-health data on incidence (number of new infections per unit time) or prevalence (number of infected people as a fraction of the population) for a range of human-infecting viruses, and constructed a hierarchical Bayesian model in Stan[24] to link these metrics to viral relative abundance for a subset of studies (Methods, Table 1).

For this purpose, we selected 16 human-infecting viruses (Table 2), based on availability of relevant public-health estimates as well as either public-health importance (Group 1) or high read counts in our selected datasets (Group 2; see Methods). For viruses that cause chronic infections, we collected estimates of prevalence, while for acutely infecting viruses, we estimated weekly incidence over the studies' coverage period (Fig. 2, Table 2, Methods). We used these estimates to parameterize our model (Fig. 3) to predict expected viral relative abundance in wastewater at a given prevalence or incidence for each virus and study, abbreviated as RA_p and RA_i respectively.

Virus	Abbreviation	Incidence/ Prevalence	Underlying Data for US	Underlying Data for Denmark
Severe Acute Respiratory Syndrome Coronavirus 2	SARS-CoV-2	Incidence	Confirmed and probable COVID-19 cases[25], adjusted by a CDC-provided underreporting factor[26]	No estimate
Influenza A&B Virus	Flu A&B	Incidence	Number of positive tests, adjusted by a yearly underreporting factor[27,28]	No estimate
Norovirus	Norovirus	Incidence	Yearly incidence estimate from 2006[29], adjusted by number of outbreaks per year[30].	No estimate
Herpes Simplex Virus 1	HSV-1	Prevalence	2015 CDC estimates of HSV-1 seroprevalence among 14-49 yos. olds[31].	Extrapolated seroprevalence data based on a 2008-2011 survey of 6627 Germans[32].
Cytomegalo virus	CMV	Prevalence	US-American CMV seroprevalence from NHANES cycle 1999-2004[33].	Extrapolated 2006 seroprevalence measurement of 19,781 Dutch sera[34].
Epstein-Barr Virus	EBV	Prevalence	CDC EBV seroprevalence measurements from NHANES cycles 2003-2010[35], adjusted for current demographics.	1983 EBV seroprevalence measurements, adjusted for current demographics[36].
Human Immunodeficiency Virus (untreated cases)	HIV	Prevalence	2019 CDC estimate of HIV prevalence[37], scaled by share of patients who are not immuno-surpressed[37].	2022 Undiagnosed HIV-positive individuals in Denmark[38], scaled by share of diagnoses made in Copenhagen[38].
Hepatitis C Virus	HCV	Prevalence	2013-2016 Estimate of chronic Hep C prevalence by the CDC[39]	2016 Estimate of total HCV cases in Copenhagen, DK[40]
Human Papillomavirus	HPV	Prevalence	2018 CDC HPV prevalence estimates among 15-59-year-olds[41].	HPV PCR measurements in a 2016 male Danish cohort of 2436 men[42]
Adeno-Associated Virus 2	AAV-2	Prevalence	Extrapolation of global seroprevalence measurements in hemophilia patients[43]	Extrapolation of seroprevalence measurements in Northern-European hemophilia patients[43]
John-Cunningham Virus	JCV	Prevalence	2009 US seroprevalence survey of healthy adult blood donors[44].	2009 Swiss seroprevalence study of 400 healthy adult blood donors. Extrapolated to Denmark[45].
BK Virus	BKV	Prevalence	2009 US seroprevalence survey of healthy adult blood donors[44].	2009 Swiss seroprevalence study of 400 healthy adult blood donors. Extrapolated to Denmark[45].
Merkel Cell Virus	MCV	Prevalence	2009 US seroprevalence study of 451 female control subjects[46].	2018 Dutch seroprevalence study of 1050 serum samples. Extrapolated to Denmark[47].

Table 2: Human-infecting viruses included in this study. See Materials and Methods for additional information on how each estimate was created.

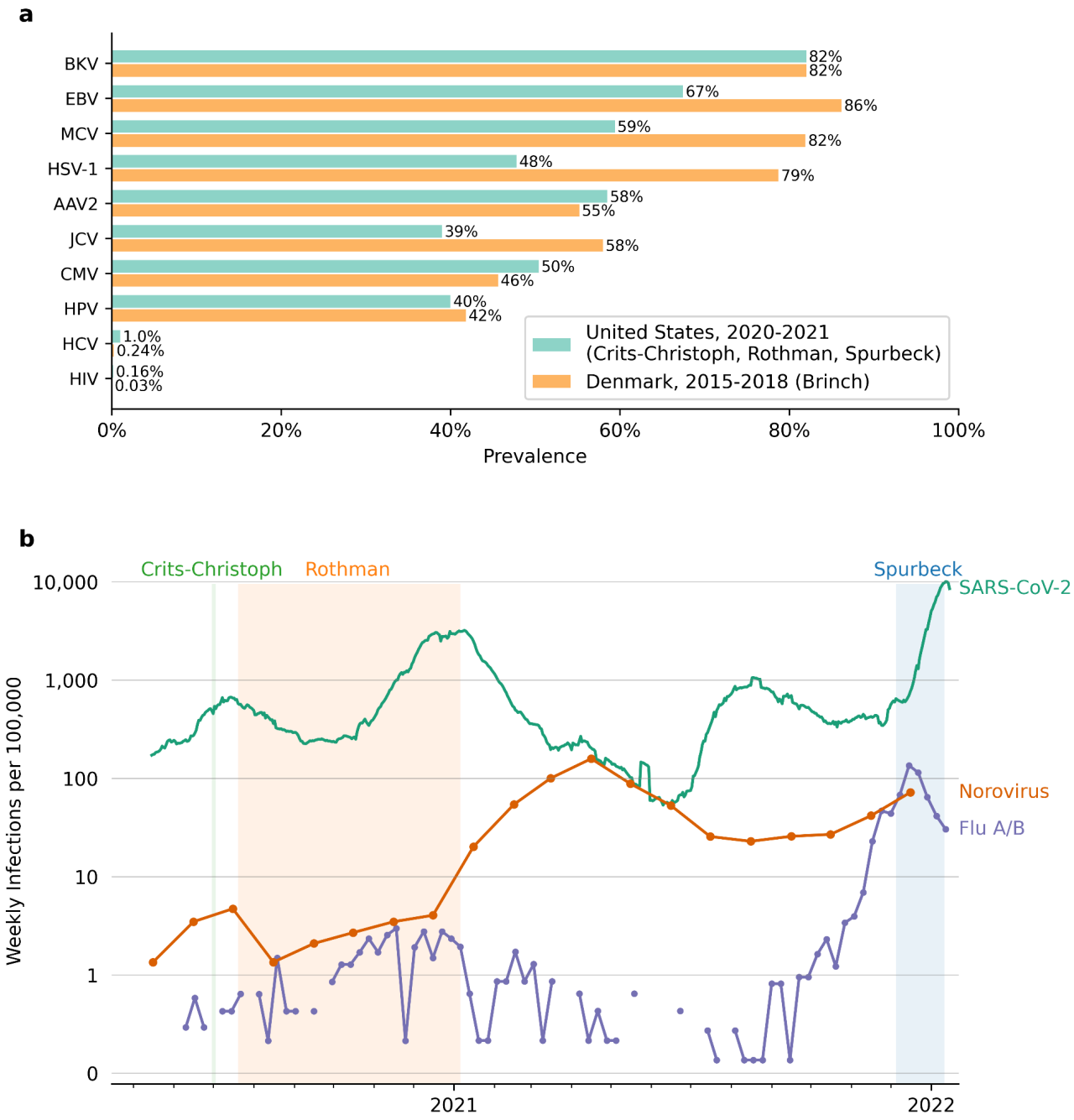

Figure 2: Prevalence and incidence of human-infecting viruses. **(a)** Prevalence of chronic-infecting viruses in the United States 2020-2021 (green), and Denmark 2015-2018 (orange) across samples for each included study. **(b)** Incidence of acute-infecting viruses in the United States, May 2020 to January 2022. SARS-CoV-2 incidence rates (dark red) are a population-weighted average of county-level daily incidences for all counties covered by target metagenomic studies. Influenza incidence rates are population-weighted averages of state-level weekly incidences for the states covered by target metagenomic studies (Ohio and California). Norovirus weekly incidence rates are based on nation-level, monthly data. No estimates for acute-infecting viruses were created for Denmark, as RNA viruses were expected to show zero read counts in DNA sequencing data.

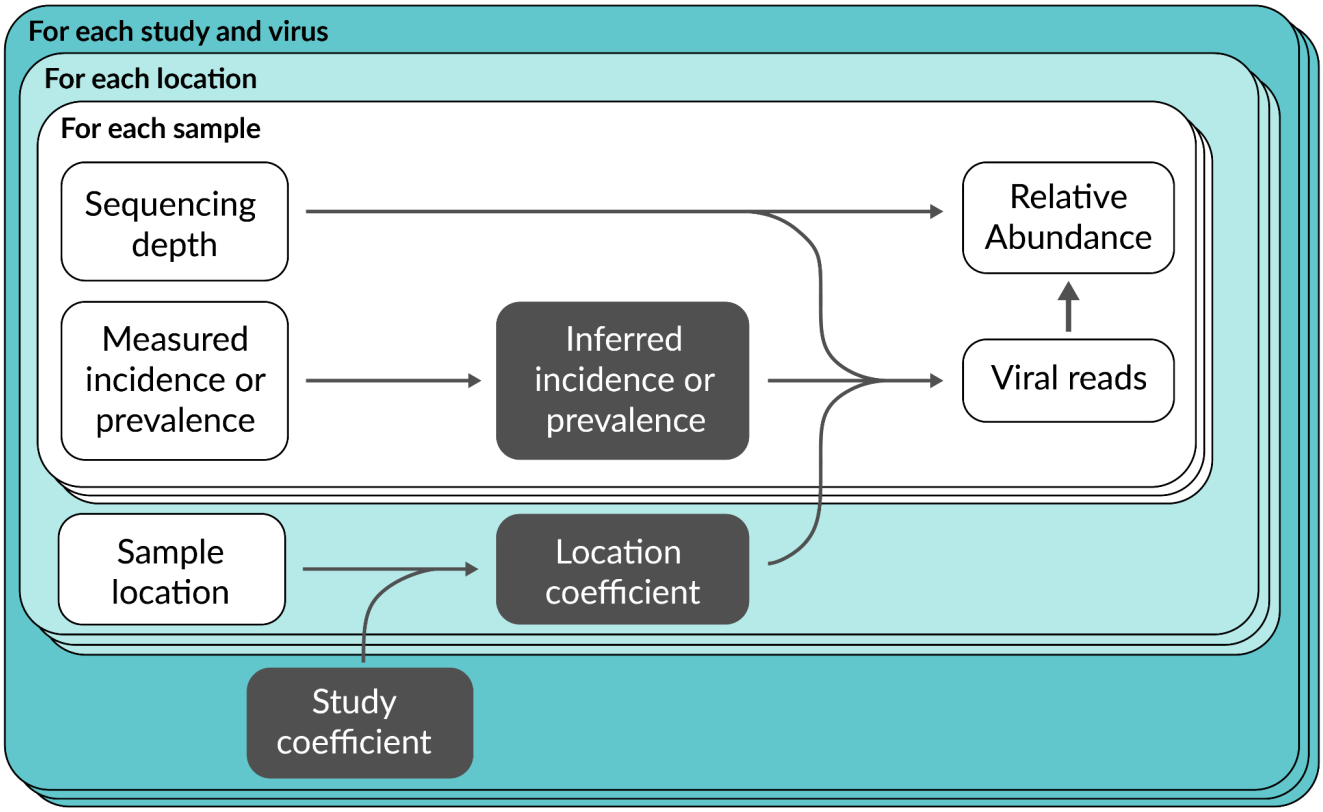

Figure 3: Overview of hierarchical Bayesian model. White boxes are observed variables. Dark boxes are estimated variables. The model uses logistic regression to predict the number of viral reads as a function of the public health predictor (incidence/prevalence) and sample location. Predicted relative abundance is computed as the ratio of the viral read count and overall number of reads (sequencing depth). Coefficients are estimated independently for each study and virus.

Inter-study viral abundance differences persist after incorporating epidemiological data

To allow comparisons between studies and viruses, we focus here on RA_i at 1% incidence (1/100 people infected per week) and RA_p at 1% prevalence (1/100 people currently infected), denoted $RA_i(1\%)$ and $RA_p(1\%)$ respectively. See Figures S2-S4 for results at other incidence/prevalence values.

Even when controlling for viral prevalence and incidence in this way, our results highlight substantial differences in expected relative abundance among viruses and studies (Fig. 4). For example, median estimates of $RA_i(1\%)$ vary by four orders of magnitude between studies for norovirus GI and three orders of magnitude for SARS-CoV-2 (Fig. 4a & S5, Table S5), while median estimates of $RA_p(1\%)$ vary by four orders of magnitude for MCV and two orders of magnitude for BKV (Fig. 4b & S5, Table S5). These differences between viruses across studies were not consistently uniform in direction; for example, Rothman shows higher $RA_i(1\%)$ than Crits-Cristoph for norovirus GI and GII, but lower $RA_i(1\%)$ for SARS-CoV-2 (Fig. 4a, Table S5). Exceptions to the general rule include HIV and CMV, both

of which showed relatively uniform $RA_p(1\%)$ across studies that contained reads for these two viruses (Fig. 4b & S5).

Many chronic viruses (e.g. HCV, CMV, and HPV) showed very low abundance in wastewater data, with zero mapped reads in at least two of the four studies. Influenza showed a similar pattern. In these cases it was only possible to estimate an upper bound on $RA_{p/i}(1\%)$, representing the information we obtain from knowing that a study did *not* detect a virus at some sequencing depth. Such upper bounds differ between studies and viruses based on differences in total read count and epidemiological indicators; for example, HCV and HPV both had zero reads in Crits-Christoph, but their median $RA_p(1\%)$ differed by two orders of magnitude (1.5×10^{-10} vs 3.9×10^{-12}) (Fig. 4, Table S5).

In addition to estimating study-level $RA(1\%)$ values for different viruses, we also estimated these coefficients for different sampling locations within each study (Fig. S6). While more consistent than between studies, intra-study $RA(1\%)$ estimates still showed moderate variation between locations (Fig. S6, Table S6-S7).

Projected costs of metagenomic wastewater surveillance span several orders of magnitude

To learn more about the potential viability of untargeted wastewater-based biosurveillance, we developed a simple model to convert our estimates of $RA_{p/i}(1\%)$ into estimates of the weekly number of sequencing reads required to detect a pathogen by the time it reaches a certain cumulative incidence in the population (Methods). We considered a virus to be detected when the cumulative reads associated with that virus crossed a predefined threshold. Given this detection threshold, a target cumulative incidence, and an $RA_i(1\%)$ value, we can estimate the weekly sequencing depth V required for detection (Methods, Fig. 5):

$$V \approx \frac{detection\ threshold}{100\ RA_i(1\%) \times cumulative\ incidence}$$

Averaging across studies, we find that detection at 1% cumulative incidence, with a read threshold of 100 reads, requires an average of 7×10^7 sequencing reads for norovirus and 1×10^{10} reads for SARS-CoV-2 per week. However, these average values conceal substantial variation in estimates between studies (Fig. 5) which could alter the required read depth by multiple orders of magnitude in either direction. Changing the detection threshold alters the required read depth approximately proportionally.

The results of this model can be used to calculate initial cost estimates for effective pathogen detection via wastewater metagenomics. For example, given a cost of $5,500 per billion reads (Methods) and using the average read depth estimates above, detection at 1% cumulative incidence with a read threshold of 100 would cost roughly $19,000/year for a norovirus-like pathogen and $2,900,000/year for a SARS-CoV-2–like pathogen. Again, however, individual studies different substantially in their projected costs: for the same parameters, Rothman and Crits-Christoph would

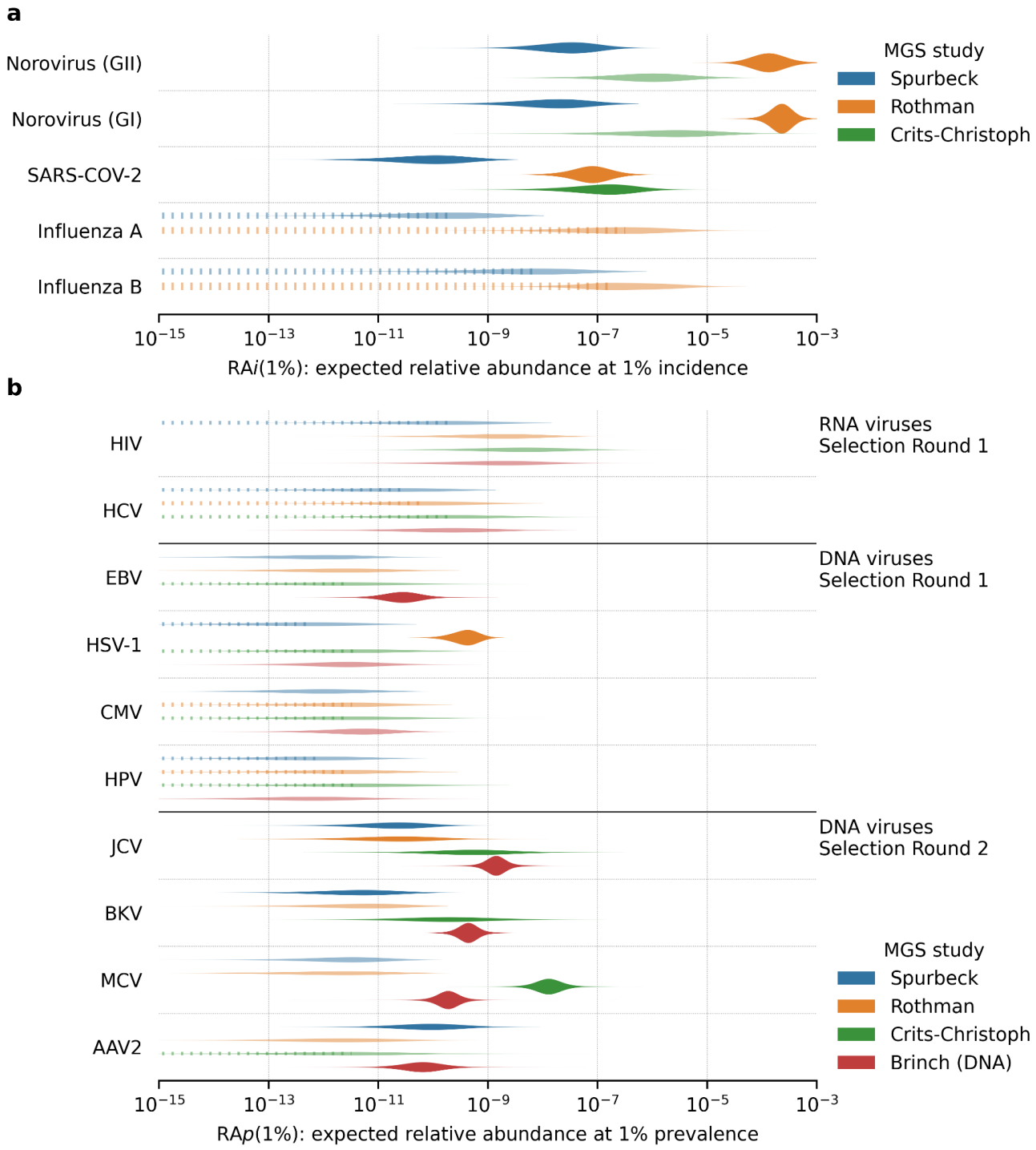

Figure 4. Study-level relative abundance (*RA*) estimates. **(a)** Predicted *RA* of acute viruses at 1% weekly incidence (RA_i(1%)). Influenza RA_i(1%) for Crits-Christoph is not displayed (Methods). **(b)** Predicted *RA* of chronic viruses at 1% prevalence (RA_p(1%)). Each violin represents the posterior predictive distribution of our Bayesian model for a specific study and virus. Transparent violins indicate predictions made with <10 reads mapping to the corresponding virus; dashed lines indicate predictions made with zero mapped reads

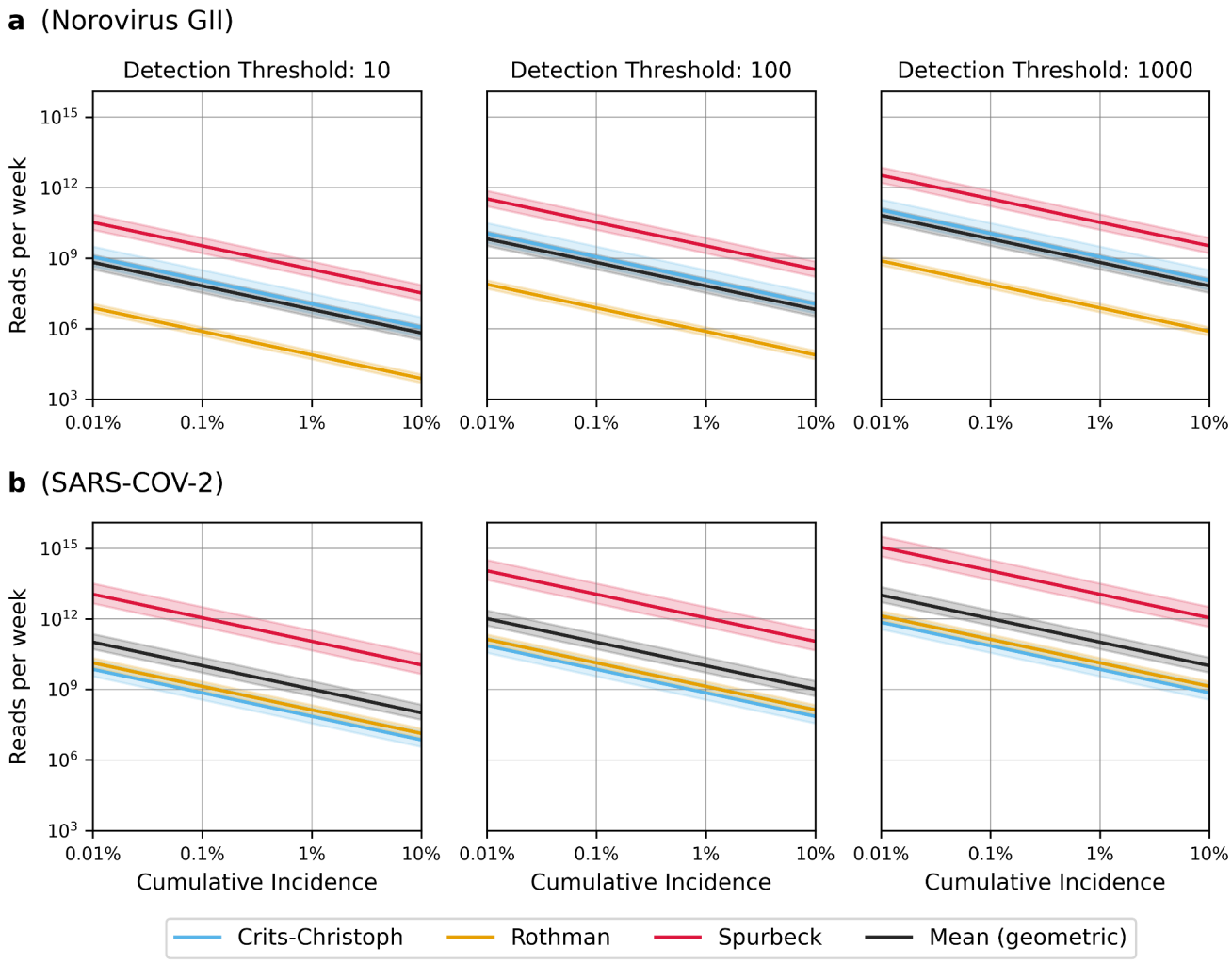

Figure 5: Weekly reads required for detection, given the performance of different study protocols (Crits-Christoph, Rothman, and Spurbeck), and their mean performance. **(a)** Norovirus (GII), **(b)** SARS-CoV-2

require $390,000 and $200,000 in sequencing per year to detect SARS-CoV-2, while Spurbeck's sensitivity would translate into a prohibitively expensive $320,000,000. For norovirus, the cost span increases to four orders of magnitude, ranging from $200 (Rothman) to $960,000 (Spurbeck) per year.

Discussion

Metagenomic surveillance of wastewater (W-MGS) could enable monitoring of a wide range of known and unknown pathogens, strains, and variants without reliance on clinical presentation. However, any untargeted wastewater-based surveillance effort that hopes to detect novel viruses while they are still rare will need to sequence samples to high depth, incurring substantial sequencing costs. Estimating the scale of these costs for different pathogens and sequencing protocols is an important component of any attempt to estimate the cost-effectiveness of metagenomic biosurveillance.

To answer this question, we combined existing W-MGS datasets with epidemiological estimates of different pathogens to estimate the expected relative abundance of different viral pathogens at a given incidence or prevalence for a range of studies and sampling locations. We find that predicted relative abundance varies widely across both pathogens and studies, suggesting that non-epidemiological sources of inter-study variation (e.g. protocol choice, sampling methodology, and sewershed characteristics) will have a dramatic impact on the performance of metagenomic sequencing as a biosurveillance tool.

Using a simple model linking cumulative incidence to sequencing depth, we converted our relative-abundance predictions into estimates of the cost of pathogen detection using W-MGS-based biosurveillance. As with the relative-abundance estimates themselves, these cost estimates varied by orders of magnitude between studies and pathogens. At the optimistic end of this variation, the projected yearly cost to detect 100 cumulative reads of a SARS-CoV-2 type virus at a single location at 1% cumulative incidence was around $200,000 per year, while at the pessimistic end it exceeded $300 million. Further research into the driving factors behind this variability will be important for determining the near- and medium-term feasibility of W-MGS-based biosurveillance; under higher-cost regimes, successful implementation will require improvements in sample preparation methods, highly sensitive threat detection algorithms, and sustained drops in sequencing cost.

While this study represents an important step forward in quantifying the efficacy and cost of metagenomic wastewater surveillance, it nevertheless has important limitations. Most notably, there is the limited amount of available sequencing data from regions with robust public-health estimates. In total, the studies we selected for in-depth analysis comprised roughly 3B RNA-sequencing reads and 4B DNA-sequencing reads: a significant amount compared to most individual studies, but inadequate to quantify viruses at very low relative abundances. Many viruses returned 0 mapped reads in some or all of the included studies, a problem exacerbated by a lack of DNA studies that deliberately enriched for viruses. To address these issues, future research should obtain or generate significantly deeper sequencing data, generated with protocols optimized for viral W-MGS.

The available public health also had important limitations. Many RNA viruses of interest either had little epidemiological information available (metapneumovirus, rhinovirus) or remained close to zero incidence during the periods covered by our data (respiratory syncytial virus), making it impractical to generate estimates for these viruses. Where robust estimates were available, they were often presented as point estimates without adequate or consistent representation of uncertainty. Lastly, many of our estimates incorporated underreporting factors that were applied broadly across time and space (likely underrepresenting real-world variability in reporting) and varied in their methodology between pathogens. Improvements in available public-health estimates would substantially aid future research in this domain.

We see two avenues for future research building on our results. First, our parameter estimates should be integrated into more advanced cost-effectiveness models comparing the feasibility of different approaches to pathogen detection while incorporating other factors influencing W-MGS sensitivity (e.g. detection noise). Second, future work should improve those parameter estimates themselves, by incorporating larger viral MGS datasets, higher-quality public-health data, and additional sequencing platforms. Generating and incorporating large, virally-enriched DNA-sequencing datasets from wastewater would be especially valuable. As sequencing prices continue to decline over the coming years, future research on W-MGS will play a central role in determining whether and when this promising approach to biosurveillance merits widespread implementation.

Materials and Methods

Metagenomic Data

Study selection

We performed a literature search for studies that conducted untargeted shotgun metagenomic sequencing of municipal wastewater influent and generated a large amount of data (>100M read pairs). We identified and obtained data from 10 such studies (Table 1). For selected studies, we discarded data from samples from sources other than municipal influent (e.g. treated sludge, effluent) and from samples sequenced with methods other than untargeted shotgun MGS (e.g. target enrichment).

After investigating viral relative abundance and composition in the 10 selected studies (Fig. 1), we selected a subset of four studies for inclusion in our hierarchical Bayesian model (Fig. 3). Studies were included in this second subset on the basis of (i) taking place in areas with good public health data available for analysis, and (ii) using sample preparation methods well-suited for broad enrichment of viruses, such as size selection with a 0.22 µm filter[16].

Three RNA studies from the US met these criteria: Crits-Cristoph et al. 2021[10], Rothman et al. 2021[16], and Spurbeck et al. 2023[17]. We additionally included Brinch et al. 2020 from Denmark, as this represented the largest eligible DNA study in our set. Inclusion criteria for studies can be found in Table S3. All four of these studies conducted composite sampling of municipal influent (the three RNA studies all used 24-hour composite samples, while Brinch used 12-hour composites) and sequenced processed samples with paired-end Illumina technology.

Data Analysis

FASTQ files for each included study were obtained from the Sequencing Read Archive[48] and analyzed with a custom computational pipeline (see “Data and Code Availability”) as follows:

1. Raw reads were cleaned with AdapterRemoval2[49] v2.3.1 (default settings) to detect and remove adapters, trim low-quality bases, and merge overlapping forward and reverse reads.
2. Cleaned reads were mapped using Kraken2[50] (default settings) to its Standard 16 GB reference database[51](2022-12-09 build).
3. Taxonomic IDs corresponding to human-infecting viruses were identified using the Kyoto University Bioinformatics Center's Virus-Host Database[52] and used to subset the Kraken2 results to human-infecting virus (HV) reads.
4. Kraken2 results were validated by mapping HV reads to their corresponding RefSeq genome[53] using BowTie2[54] (`bowtie2 --local --very-sensitive-local --score-min G,1,0 --mp 2,0`). Alignment scores were normalized by dividing by the natural logarithm of the read length, and reads with normalized alignment scores below a threshold value of 22 (Fig. S7) were discarded.
5. Finally, relative abundances were calculated as the number of validated reads assigned to human-infecting viruses for a given dataset, divided by the total number of reads in that dataset.

Epidemiological Data

Virus Selection

To select viruses for which to perform epidemiological estimates, we performed an exploratory literature review choosing viruses based on public-health importance and availability of applicable incidence or prevalence estimates (Table 2 & S4). We chose a set of 16 "Group 1" viruses for analysis based on these initial criteria. We later identified five further "Group 2" viruses (Table 2) that had both good public health data and significant read counts in our chosen studies. While this allows intra-study, inter-sample site comparison analysis (Fig. S6b), the inclusion of these viruses introduces selection bias[55]: similarly prevalent viruses that shed less into stool or were less amenable to the sequencing approaches used would not have been selected for study. Inter-study comparisons for Group 2 viruses are marked accordingly.

Data Analysis

For viruses that cause chronic infections we collected estimates of prevalence, while for acutely infecting viruses we estimated weekly incidence estimates. SARS-CoV-2 incidence estimates were based on daily confirmed and probable county-level COVID-19 cases (provided by CDC)[25], and adjusted upwards by a uniform underreporting factor[26]. Influenza incidence was estimated using state-level, weekly positive testing data[27], which is adjusted by a custom yearly underreporting factor based on the ratio between reported tests[27] and CDC estimates of total symptomatic flu infections[28].

Norovirus incidence rates were based on the number of monthly, nationwide outbreaks per year[30], which were transformed into case counts using a 2006 US-wide incidence estimate[29]. Both Norovirus[30] outbreak data and Influenza[27] and COVID-19 testing data[25] were provided by CDC. No incidence estimates were generated for the region and period covered by Brinch et al. 2020[15]; this study conducted DNA sequencing, while all incidence viruses had RNA genomes.

For the United States, publicly available prevalence estimates were available for human immunodeficiency virus (HIV)[37], hepatitis C virus (HCV)[39], and herpes simplex virus 1[31]. When prevalence estimates weren't available, seroprevalence data—the share of individuals with antibodies against the virus—was used instead. This was the case for cytomegalovirus (CMV)[33], and epstein-barr-virus (EBV)[35]. PCR testing data was available for human papilloma virus (HPV)[41]. Seroprevalence and PCR data was collected in the National Health and Nutrition Examination Survey (NHANES), a biannual health survey of a US population sample that includes testing for common chronic infections[56].

For viruses of lower public-health concern, estimates were not generally available from the sources above. We thus expanded our search to individual seroprevalence studies for adeno-associated virus 2 (AAV-2)[43], John-Cunningham virus (JCV)[44], BK virus (BKV)[44] and merkel cell virus (MCV)[46].

For Denmark, publicly-available prevalence estimates were available for HCV[39] and HIV[38]. For EBV[36] and HPV[42], we used published seroprevalence and qPCR positivity estimates, respectively, as proxies for overall prevalence. When in-country estimates weren't available, we resorted to the best data source from a country or region with similar demographics, such as Switzerland (BKV)[45], the Netherlands (MCV, CMV)[34,47], Northern Europe (AAV-2)[43], or Germany (HSV-1)[32]. Table 2 describes the method of estimate generation for each virus in detail.

Statistical Model

Derivation

We are interested in modeling the relative abundance of a given virus (henceforth the "focal virus") in W-MGS data as a function of its incidence or prevalence (henceforth "public health predictor") in the population. Relative abundance is the fraction of reads assigned to a focal virus and, as such, is constrained to the interval [0, 1]. We assume that when relative abundance is low, it increases proportionally with the public health predictor. With these properties in mind, a natural model is logistic regression with unit slope:

$$E[relative\ abundance] = logit^{-1}(b + \log public\ health\ predictor)$$

$$\sim e^{b} \times public\ health\ predictor,$$

where the second line holds as the argument of the inverse logit approaches zero. We estimate the parameter b, which governs the relationship between relative abundance and the public health predictor, as described below. We can transform an estimate of b into an estimate of $RA(x)$ by substituting x for the value of the public health predictor into the equation above. For example,

$$RA(1\%) = logit^{-1}(b - \log 100).$$

We estimate b and $RA(1\%)$ for each virus in each metagenomic study separately.

In each study, there are S metagenomic samples taken at various times from L sampling locations. For sample $s \in \{1, ..., S\}$, the data consist of the sampling location $l(s)$, the total number of reads n_s, the number of focal viral reads y_s, and the public health predictor x_s. Because sampling locations vary along several dimensions including sample preparation methods, we use a hierarchical model with a separate term b_l for each sampling location. The model thus produces a joint estimate of location-specific effects and an overall coefficient for each study and virus.

We model the focal viral counts in each sample as a binomial random variable:

$$y_s \sim Binomial\left(n_s,\ logit^{-1}\left(b_{l(s)} + \theta_s\right)\right),$$

$$\theta_s \sim Normal\left(\log x_s, \sigma\right),$$

$$\sigma \sim Gamma(2,\ 1)$$

where θ_s is a latent variable, centered around the estimated log public health predictor, that accounts for three factors:

1. error in the public health predictor
2. differences between the population the public health predictor is derived from (e.g., the entire United States for all of 2021) and the population contributing to the sample (e.g., Orange County on May 21, 2021)
3. unbiased noise in the read counts that is not accounted for by the binomial model.

For each study and virus, we infer the combined magnitude of these effects, σ, from the data.

Finally, we define a location-specific term $b_{l(s)}$, where $l(s) \in \{1, ..., L\}$ is the location of the sample. The hierarchical model of intercept terms has the structure:

$$b_l \sim Normal(b,\ \tau),$$

$$b \sim Normal(\mu,\ 4),$$

$$\tau \sim Gamma(2,\ 1).$$

Here, b is the overall intercept for the study and virus and τ is the standard deviation of the location-specific terms around this overall value. The prior on b is centered on

$$\mu = \log \bar{x} + \log \bar{y} - \log \bar{n},$$

so that $b = \mu$ roughly represents the naive estimate given by dividing the relative abundance in all of the data by the average log public health predictor estimate across samples. The prior on b is weakly informative: it supposes that the naive approximation is not off by more than a few orders of magnitude, but does not contain any other substantive scientific information. We chose the prior standard deviation of b to be as large as possible while still allowing for efficient sampling.

As with σ, τ is learned from the data, allowing for some studies to have a lot of variation between sampling locations and others to have little.

We fit the model using the Stan probabilistic programming language[15]. The code for the model is publicly available online (See "Data and Code Availability").

Pseudocounts

Sometimes the incidence of a virus is zero for a particular sample. This happens when no new cases were reported in the study region during the period overlapping the sample. In order to obtain a finite log-incidence, for the model, we introduced pseudocounts of 0.1 total cases in a given region per week. In order to ensure that our choice of pseudocount did not strongly influence the results, we performed a sensitivity analysis. We found that for most viruses and studies, the effect of pseudocounts is minimal. However, we found that our inference for influenza in Crits-Christoph was dominated by the pseudocounts, so we dropped it from Figure 4.

Pathogens with zero reads

Many viruses of interest returned 0 mapped reads in some or all of the included studies. Based on our model, such an observation still enables us to estimate weak upper-bound estimates of $RA_{p/i}(1\%)$: If two studies sequence at different depths, without detecting a virus with equal prevalence or incidence, the study that sequences more deeply can be expected to have lower $RA_{p/i}(1\%)$.

Model Checking

To assess the posterior sampling, we examined plots of the marginal and joint posterior distributions of the parameters. Posterior distributions that have irregular "spikes" at particular values or are strongly multimodal indicate that the Hamiltonian Monte Carlo sampler may not be exploring the parameter space efficiently, likely due to model specification issues. We also compared the prior and posterior distributions to ensure that the priors did not have undue influence on the posteriors. For example, we found that our initial choice of prior $b \sim Normal(\mu, 2)$ was too narrow, significantly constraining the posterior distribution even for the virus/study combinations with a lot of data.

Conversely, we found that $b \sim Normal(\mu, 10)$ was too broad, leading the sampler to mix poorly. Similarly, we adjusted the priors on σ and τ to avoid the lower bound at zero.

To check the fit of the model, we used the posterior samples of the model to generate simulated datasets. We compared the simulated data to the observed data, looking for features not accurately captured by the model. For example, earlier versions of the model did not include terms for sampling locations within studies. Posterior predictive checks revealed that this model failed to predict the low read counts for SARS-CoV-2 in the Point Loma (PL) location of the Rothman dataset, inspiring us to add location-specific effects.

Cost Model

Derivation

The sequencing effort required to detect a pathogen is a key determinant of the cost of genomic surveillance. To estimate the cost of detecting a virus, we created a simple model predicting the total sequencing reads required for detection as a function of viral incidence and RA(1%). In making this estimate, we did not consider fixed and per-sample costs that do not scale with the number of reads sequenced.

While detection algorithms differ, they all require the presence of reads originating from the virus of interest. Certain methods—such as mapping to a known reference—may detect a pathogen from just a few reads, while less targeted methods may require far more. To represent different computational detection methods, we model a virus as "detected" when the cumulative focal viral reads exceed some chosen threshold.

The expected number of viral reads observed over a given time period can be modeled as a function of viral incidence, which is converted into an estimate of relative abundance using the RA_i values estimated by the model described above.

Let $i(t)$ be the weekly incidence of the focal virus during week t at the location of interest. If n total reads are sequenced at that location each week, then we expect to find $n \times RA_i(i(t))$ reads from the focal virus. The cumulative read count from that virus is calculated as a sum over all previous weeks of monitoring:

$$E[cumulative\ viral\ reads\ at\ week\ t] = \sum_{t' \le t} n \times RA_i(i(t'))$$

When relative abundance is low, the value of RA_i estimated by our statistical model scales roughly proportionally with incidence. As a result, we can compute $RA_i(i(t'))$ as a linear function of $RA_i(1\%)$:

$$E[cumulative\ viral\ reads\ at\ week\ t] \approx \sum_{t' \le t} n \times RA_i(1\%) \frac{i(t')}{1\%}.$$

Pulling out the constant terms gives

$$E[cumulative\ viral\ reads\ at\ week\ t] \approx n \times RA_i(1\%) \times 100 \sum_{t' \le t} i(t')$$

The summation term on the right is just the cumulative incidence at week t. Rearranging the terms, the weekly total reads required for the expected cumulative viral reads to equal the detection threshold is given by:

$$n \approx \frac{detection\ threshold}{100\ RA_i(1\%) \times cumulative\ incidence}.$$

Note that we have not made any assumptions about the incidence time course, $i(t)$. In particular, this last equation is independent of the growth rate. Thus, in our model, the required sequencing depth depends only on the sensitivity of the method (through the detection threshold) and the ability to sample reads from the focal virus (through $RA_i(1\%)$).

Sequencing Cost

Approximate cost of metagenomic sequencing was estimated using pricing information by MIT's BioMicroCenter, Harvard University's Bauer Core Facility, the Dana-Farber Cancer Institute's Molecular Biology Core Facilities (MBCF). The MIT BioMicroCenter charges $23,760 per flow cell for NovaSeq S4 300nt[57], Harvard's Bauer Core facility charges $19,409[58], and Dana Farber's MBCF charges $36,000[59]. Averaging this cost (geometric mean) gives a cost of roughly $25,500. A NovaSeq 6000 with an S4 flow cell is advertised as giving 8B-10B read pairs. Doubling this cost when accounting for library preparation and personnel costs equates to roughly $5,500 per billion read pairs.

Data and Code Availability

Analysis and visualization was performed using Python[60] with the Pystan[61], Numpy[62] and Pandas[63,64] package, and Matplotlib[65] and Seaborn[66] visualization libraries. All sequencing data in this study is available through the European Nucleotide Archive[67], using BioProject accessions listed in Table S2. The raw data used for estimating pathogen prevalence and incidence are listed in the respective Python script for each pathogen. All code used in this study can be found in two Github repositories, https://github.com/naobservatory/mgs-pipeline/tree/p2ra-manuscript (metagenomic data analysis pipeline) and https://github.com/naobservatory/p2ra-manuscript (epidemiological analysis, statistical models, figure generation).

Acknowledgements

We thank Asher Parker-Sartori for help obtaining public-health data and generating prevalence and incidence estimates. We also gratefully acknowledge Lenni Justen for providing helpful feedback on earlier drafts of this preprint.

Author contributions

J.T.K. and M.R.M. conceived the study; S.L.G. and J.T.K. performed literature reviews on pathogen incidence and prevalence and collected and evaluated epidemiological datasets with input from D.P.R.; J.T.K. and W.J.B. performed literature reviews on metagenomics data with feedback from M.R.M.; J.T.K. imported and processed sequencing data; D.P.R. created the study's statistical models with feedback from M.R.M., W.J.B., and J.T.K.; S.L.G., J.T.K., and D.P.R. created figures with feedback from W.J.B.; S.L.G., W.J.B., J.T.K. & D.P.R. wrote the manuscript, with feedback from all authors.

Funding

S.L.G., J.T.K., D.P.R., W.J.B., and M.M.M. were funded for this research project by gifts from Open Philanthropy (to SecureBio) and the Musk Foundation (to MIT). S.L.G. was additionally supported through a grant by the Swiss Scholarship Foundation. C.W. was supported by Sir Henry Wellcome Postdoctoral Fellowship, reference 224190/Z/21/Z.

References

1. Manor, Y. *et al.* Detection of Poliovirus Circulation by Environmental Surveillance in the Absence of Clinical Cases in Israel and the Palestinian Authority. *J. Clin. Microbiol.* **37**, 1670–1675 (1999).
2. Paul, J. R., Trask, J. D. & Culotta, C. S. Poliomyelitic Virus in Sewage. *Science* **90**, 258–259 (1939).
3. Brouwer, A. F. *et al.* Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10625–E10633 (2018).
4. Nenonen, N. P., Hannoun, C., Larsson, C. U. & Bergström, T. Marked Genomic Diversity of Norovirus Genogroup I Strains in a Waterborne Outbreak. *Appl. Environ. Microbiol.* **78**, 1846–1852 (2012).
5. Diamond, M. B. *et al.* Wastewater surveillance of pathogens can inform public health responses.

Nat. Med. **28**, 1992–1995 (2022).

6. Peccia, J. *et al.* Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nat. Biotechnol.* **38**, 1164–1167 (2020).
7. Karthikeyan, S. *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**, 101–108 (2022).
8. Monteiro, S., Pimenta, R., Nunes, F., Cunha, M. V. & Santos, R. Wastewater-based surveillance for tracing the circulation of Dengue and Chikungunya viruses. Preprint at https://doi.org/10.1101/2023.10.30.23297765 (2023).
9. Boehm, A. B. *et al.* Wastewater concentrations of human influenza, metapneumovirus, parainfluenza, respiratory syncytial virus, rhinovirus, and seasonal coronavirus nucleic-acids during the COVID-19 pandemic: a surveillance study. *The Lancet Microbe* **4**, e340–e348 (2023).
10. Crits-Christoph, A. *et al.* Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *MBio* **12**, (2021).
11. Nakamura, T. *et al.* Environmental surveillance of poliovirus in sewage water around the introduction period for inactivated polio vaccine in Japan. *Appl. Environ. Microbiol.* **81**, 1859–1864 (2015).
12. Goldberg, Z., Linder, A. G., Miller, L. N. & Sorrell, E. M. Wastewater Collection and Sequencing as a Proactive Approach to Utilizing Threat Agnostic Biological Defense. *Health security* (2023) doi:10.1089/hs.2023.0075.
13. McCall, C., Wu, H., Miyani, B. & Xagoraraki, I. Identification of multiple potential viral diseases in a large urban center using wastewater surveillance. *Water Res.* **184**, 116160 (2020).
14. The Cost of Sequencing a Human Genome. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

15. Brinch, C. *et al.* Long-Term Temporal Stability of the Resistome in Sewage from Copenhagen. *mSystems* **5**, e00841–20 (2020).
16. Rothman, J. A. *et al.* RNA Viromics of Southern California Wastewater and Detection of SARS-CoV-2 Single-Nucleotide Variants. *Appl. Environ. Microbiol.* **87**, e01448–21 (2021).
17. Spurbeck, R. R., Catlin, L. A., Mukherjee, C., Smith, A. K. & Minard-Smith, A. Analysis of metatranscriptomic methods to enable wastewater-based biosurveillance of all infectious diseases. *Frontiers in Public Health* **11**, (2023).
18. Bengtsson-Palme, J. *et al.* Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci. Total Environ.* **572**, 697–712 (2016).
19. Brumfield, K. D. *et al.* Microbiome Analysis for Wastewater Surveillance during COVID-19. *MBio* **13**, e00591–22 (2022).
20. Maritz, J. M., Ten Eyck, T. A., Elizabeth Alter, S. & Carlton, J. M. Patterns of protist diversity associated with raw sewage in New York City. *ISME J.* **13**, 2750–2763 (2019).
21. Munk, P. *et al.* Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat. Commun.* **13**, 7251 (2022).
22. Ng, C. *et al.* Metagenomic and Resistome Analysis of a Full-Scale Municipal Wastewater Treatment Plant in Singapore Containing Membrane Bioreactors. *Front. Microbiol.* **10**, (2019).
23. Yang, Q. *et al.* Detection of multiple viruses potentially infecting humans in sewage water from Xinjiang Uygur Autonomous Region, China. *Sci. Total Environ.* **754**, 142322 (2021).
24. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **76**, 1–32 (2017).
25. COVID-19/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv at master · CSSEGISandData/COVID-19. *GitHub*

https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv.

26. Estimated COVID-19 Burden | CDC. https://web.archive.org/web/20230722174403/https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.
27. National, Regional, and State Level Outpatient Illness and Viral Surveillance. https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.
28. Estimated Flu-Related Illnesses, Medical visits, Hospitalizations, and Deaths in the United States — 2019–2020 Flu Season | CDC. https://web.archive.org/web/20231004222154/https://www.cdc.gov/flu/about/burden/2019-2020.html (2023).
29. Scallan, E. *et al.* Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg. Infect. Dis.* **17**, 7 (2011).
30. National Outbreak Reporting System (NORS) Dashboard | CDC. https://wwwn.cdc.gov/norsdashboard/.
31. NHANES 2015-2016: Herpes Simplex Virus Type-1 & Type-2 Data Documentation, Codebook, and Frequencies. https://web.archive.org/web/20220707050306/https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/HSV_I.htm.
32. Korr, G. *et al.* Decreasing seroprevalence of herpes simplex virus type 1 and type 2 in Germany leaves many people susceptible to genital infection: time to raise awareness and enhance control. *BMC Infect. Dis.* **17**, (2017).
33. Sl, B., Sc, D. & Mj, C. Cytomegalovirus seroprevalence in the United States: the national health

and nutrition examination surveys, 1988-2004. *Clin. Infect. Dis.* **50**, (2010).

34. Korndewal, M. J. *et al.* Cytomegalovirus infection in the Netherlands: Seroprevalence, risk factors, and implications. *J. Clin. Virol.* **63**, 53–58 (2015).
35. Age-Specific Prevalence of Epstein–Barr Virus Infection Among Individuals Aged 6–19 Years in the United States and Factors Affecting Its Acquisition | The Journal of Infectious Diseases | Oxford Academic. https://academic.oup.com/jid/article/208/8/1286/2192838.
36. Prevalence of Antibodies to Epstein-Barr Virus(EBV) in Childhood and Adolescence in Denmark: Scandinavian Journal of Infectious Diseases: Vol 15, No 4. https://www.tandfonline.com/doi/abs/10.3109/inf.1983.15.issue-4.03.
37. National Profile. https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-26-no-2/content/national-profile.html (2022).
38. HIV 2020. https://en.ssi.dk/surveillance-and-preparedness/surveillance-in-denmark/annual-reports-on-disease-incidence/hiv-2020.
39. Hofmeister, M. G. *et al.* Estimating prevalence of hepatitis C virus infection in the United States, 2013–2016. *Hepatology* **69**, 1020–1031 (2019-3).
40. Hepatitis C prevalence in Denmark in 2016—An updated estimate using multiple national registers | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238203.
41. Lewis, R. M. *et al.* Estimated prevalence and incidence of disease-associated HPV types among 15–59-year-olds in the United States. *Sex. Transm. Dis.* **48**, 273–277 (2021).
42. Hebnes, J. B. *et al.* Human Papillomavirus Infection Among 2460 Men in Denmark: Prevalence in Relation to Age Using 2 Human Papillomavirus DNA Testing Methods. *Sex. Transm. Dis.* **42**,

463–467 (2015).

43. Klamroth, R. *et al.* Global Seroprevalence of Pre-existing Immunity Against AAV5 and Other AAV Serotypes in People with Hemophilia A. *Hum. Gene Ther.* **33**, 432–441 (2022).
44. Kean, J. M., Rao, S., Wang, M. & Garcea, R. L. Seroepidemiology of human polyomaviruses. *PLoS Pathog.* **5**, e1000363 (2009).
45. Egli, A. *et al.* Prevalence of Polyomavirus BK and JC Infection and Replication in 400 Healthy Blood Donors. *J. Infect. Dis.* **199**, 837–846 (2009).
46. Carter, J. J. *et al.* Association of Merkel cell polyomavirus-specific antibodies with Merkel cell carcinoma. *J. Natl. Cancer Inst.* **101**, 1510–1522 (2009).
47. Kamminga, S., van der Meijden, E., Feltkamp, M. C. W. & Zaaijer, H. L. Seroprevalence of fourteen human polyomaviruses determined in blood donors. *PLoS One* **13**, e0206273 (2018).
48. The Sequence Read Archive - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013647/.
49. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
50. Lu, J. *et al.* Metagenome analysis using the Kraken software suite. *Nat. Protoc.* 1–25 (2022) doi:10.1038/s41596-022-00738-y.
51. Index zone by BenLangmead. https://benlangmead.github.io/aws-indexes/k2.
52. Virus-Host Database. https://www.genome.jp/virushostdb/.
53. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/44/D1/D733/2502674.
54. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

55. Tripepi, G., Jager, K. J., Dekker, F. W. & Zoccali, C. Selection Bias and Information Bias in Clinical Research. *Nephron Clin. Pract.* **115**, c94–c99 (2010).

56. National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/index.htm (2023).

57. BioMicroCenter:Pricing - OpenWetWare. https://openwetware.org/wiki/BioMicroCenter:Pricing.

58. NGS Sequencing. https://bauercore.fas.harvard.edu/ngs-sequencing-fees.

59. Illumina NovaSeq 6000. *MBCF at DFCI* http://mbcf.dfci.harvard.edu/genomics/illumina-novaseq-6000.

60. Van Rossum, G., Drake, F. L. & Others. *Python reference manual*. vol. 111 (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

61. Riddell, A., Hartikainen, A. & Carter, M. pystan (3.0.0). Preprint at (2021).

62. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

63. The pandas development team. *pandas-dev/pandas: Pandas*. (Zenodo, 2023). doi:10.5281/ZENODO.3509134.

64. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). doi:10.25080/majora-92bf1922-00a.

65. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (May-June 2007).

66. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

67. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39**, D28–31 (2011).

Supplementary Information

Supplementary Figures

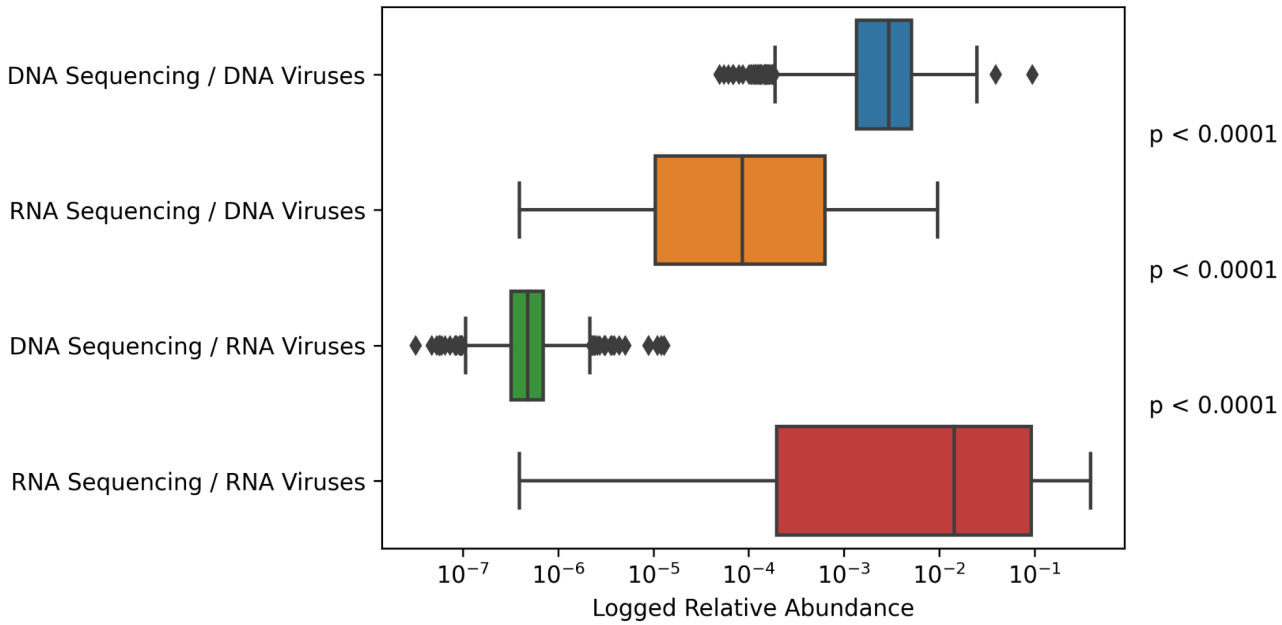

Figure S1: Relative abundance of RNA viruses and DNA viruses in samples processed with DNA vs RNA sequencing, across all studies included in Figure 1. P-values indicate significant differences between the designated abundance distributions, as determined by a Mann-Whitney U Test.

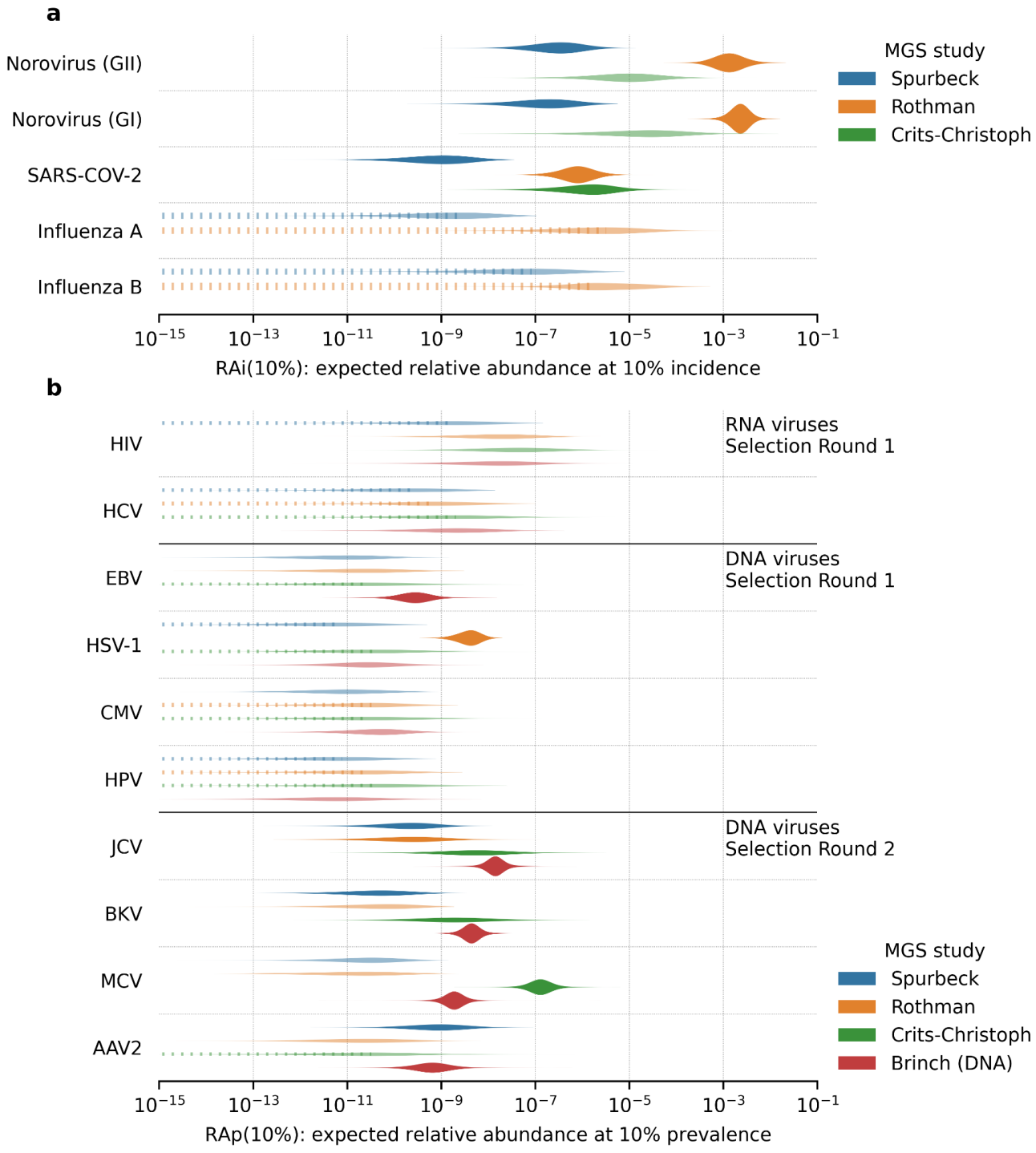

Figure S2: Study-level relative abundance (RA) as a function of epidemiological indicators. **(a)** Predicted relative abundance of acute viruses at 10% weekly incidence. **(b)** Predicted relative abundance of chronic viruses at 10% prevalence.

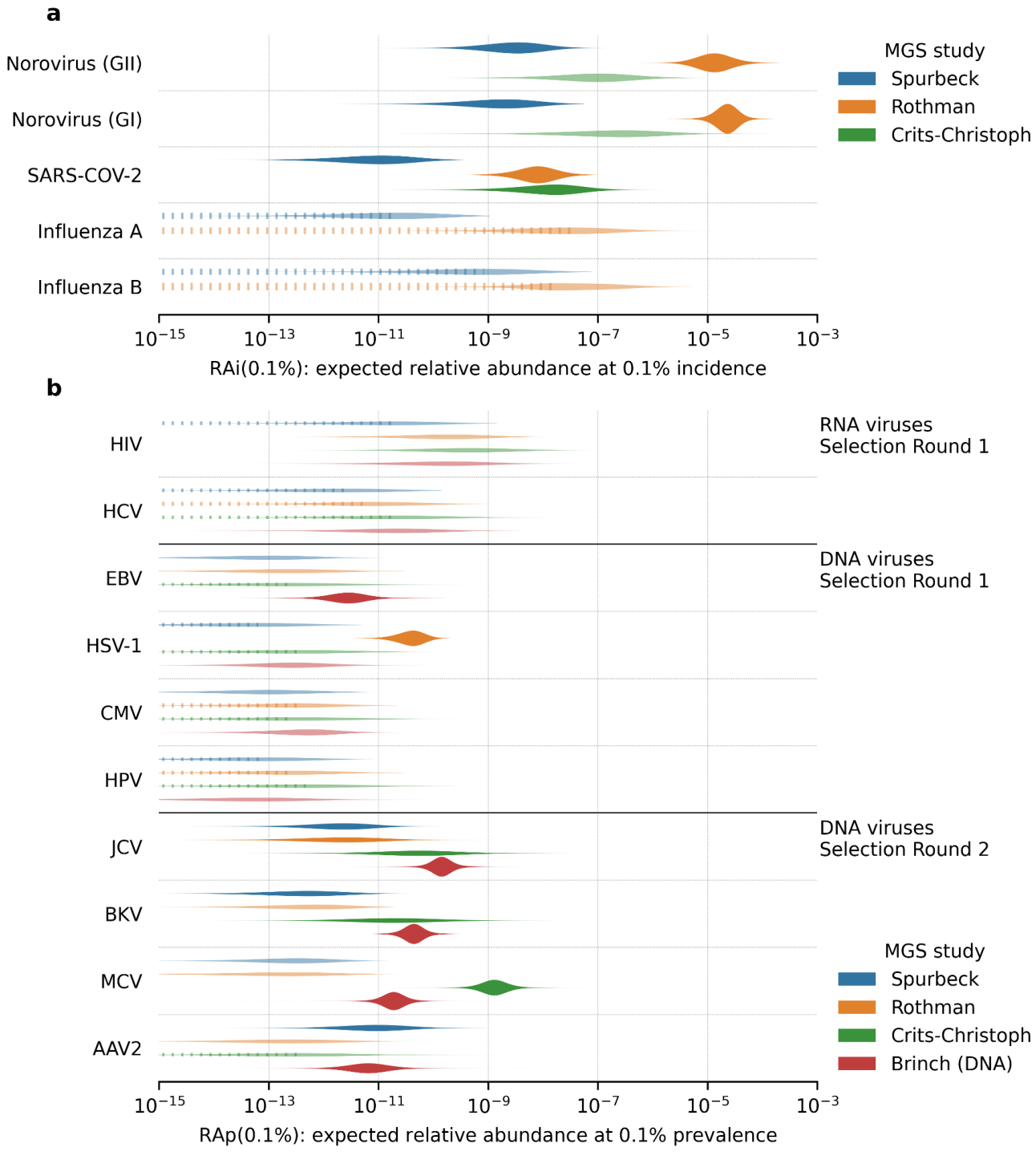

Figure S3: Study-level relative abundance (RA) as a function of epidemiological indicators. **(a)** Predicted relative abundance of acute viruses at 0.1% weekly incidence. **(b)** Predicted relative abundance of chronic viruses at 0.1% prevalence.

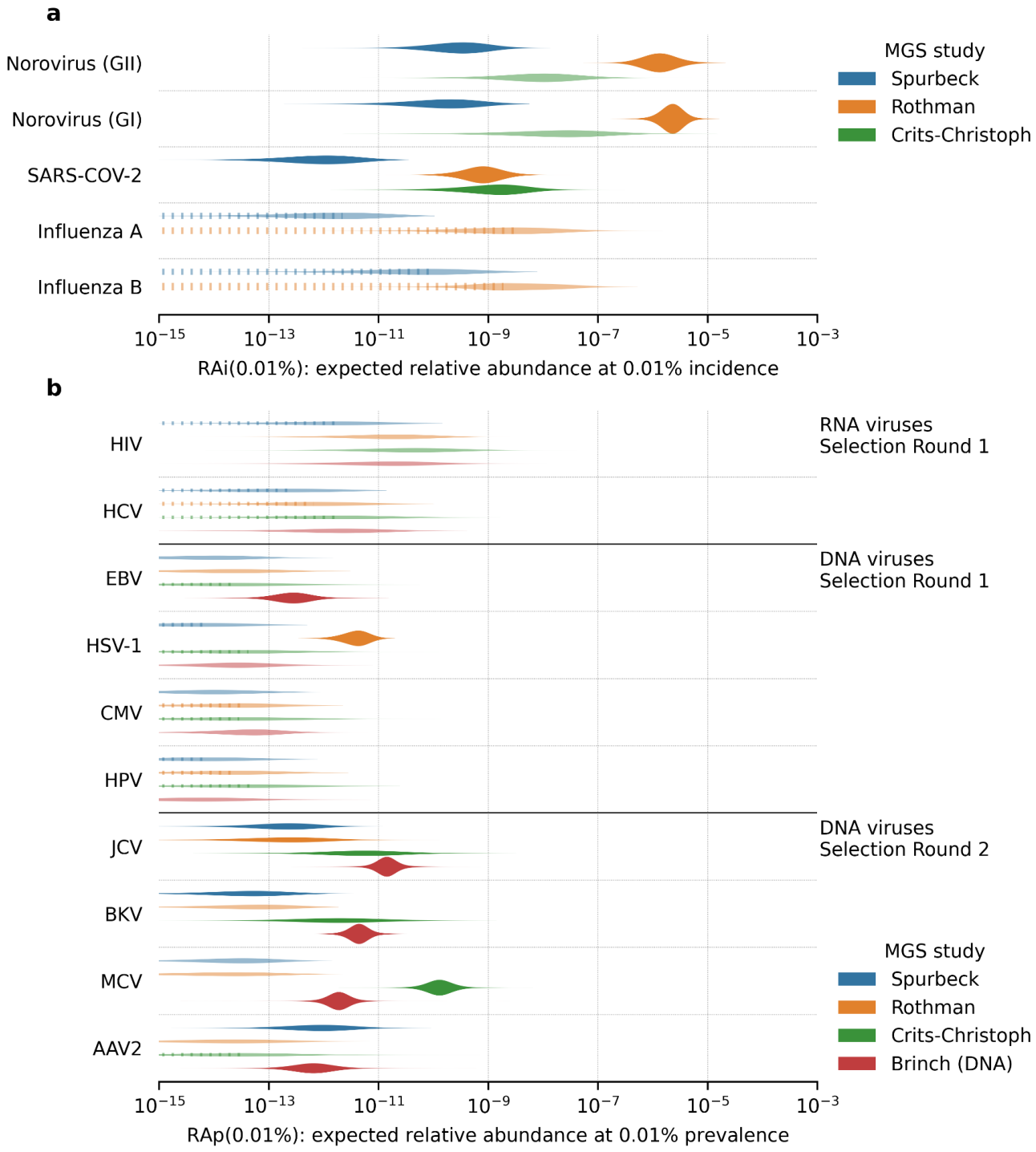

Figure S4: Study-level relative abundance (RA) as a function of epidemiological indicators. (a) Predicted relative abundance of acute viruses at 0.01% weekly incidence. (b) Predicted relative abundance of chronic viruses at 0.01% prevalence.

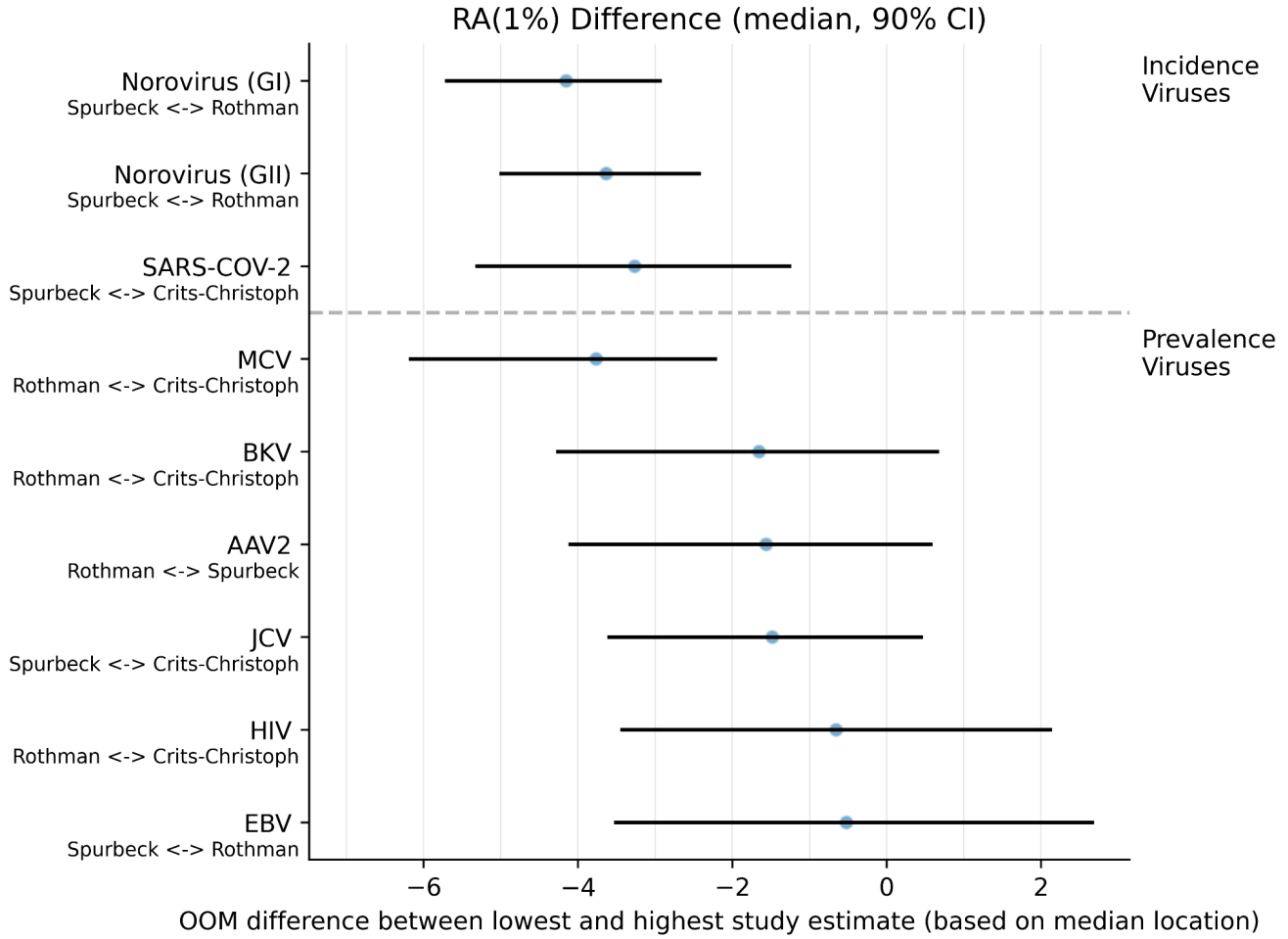

Figure S5: Inter-study variability (RA), as determined by the difference between the study with the lowest median *RA*(1%) and the study with the highest median *RA*(1%). The bars represent the difference between the two studies' median; the range spans i) the low-end of the lower estimate divided by the high-end of the higher estimate and ii) the high-end of the lower estimate divided by the low-end of the high estimate. Differences are displayed in order-of-magnitudes. Selected studies are displayed on the y-axis, studies to the left have lower RA(1%) median.

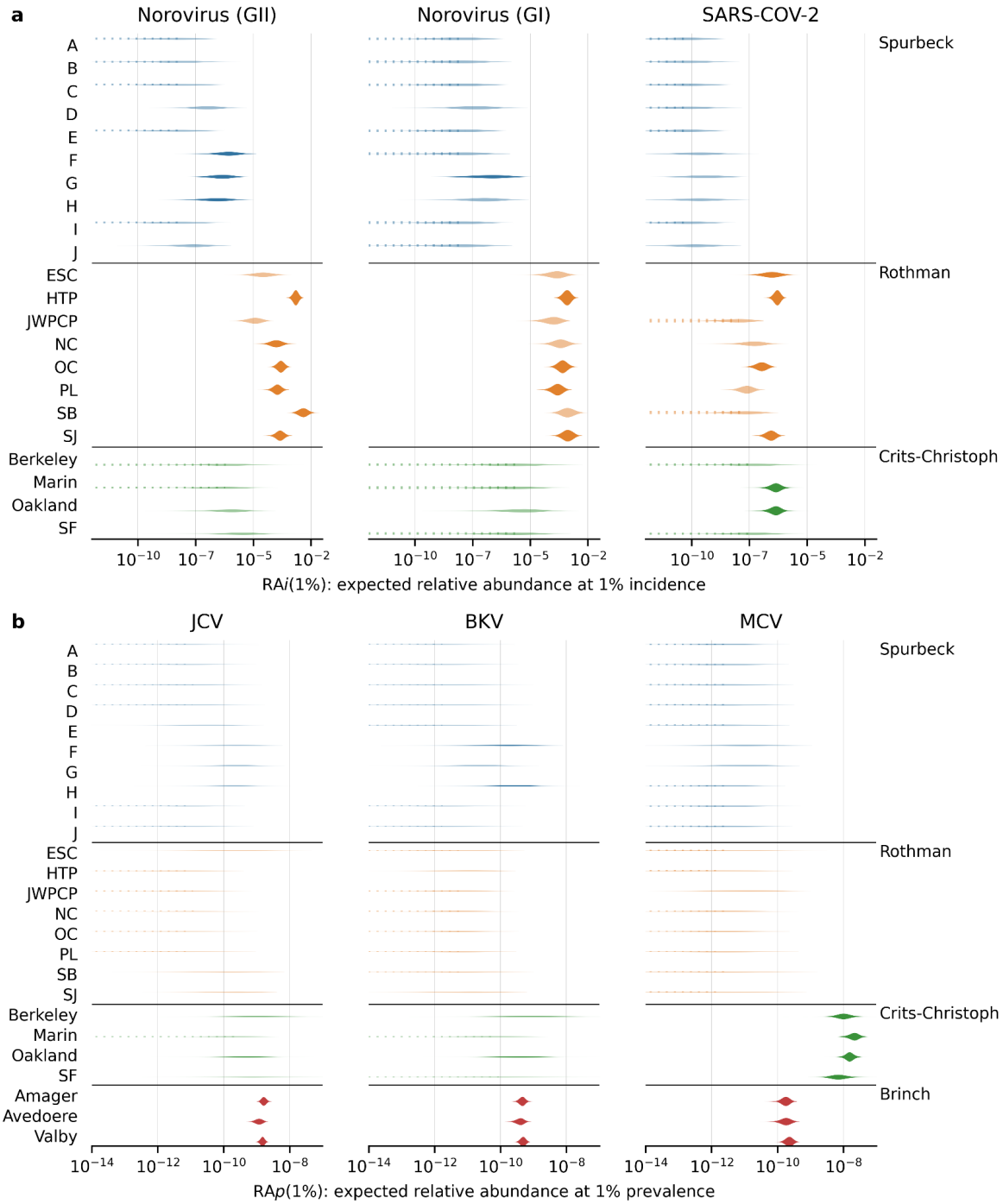

Figure S6. Location-level relative abundance (RA) estimates for selected viruses. **(a)** Predicted RA of selected acute viruses at 1% weekly incidence. **(b)** Predicted RA of selected chronic viruses at 1% prevalence. Each violin represents the posterior predictive distribution of our Bayesian model for a specific study, location, and virus. Transparent violins indicate predictions made with <10 reads mapping to the corresponding virus; dashed lines indicate predictions made with zero mapped reads.

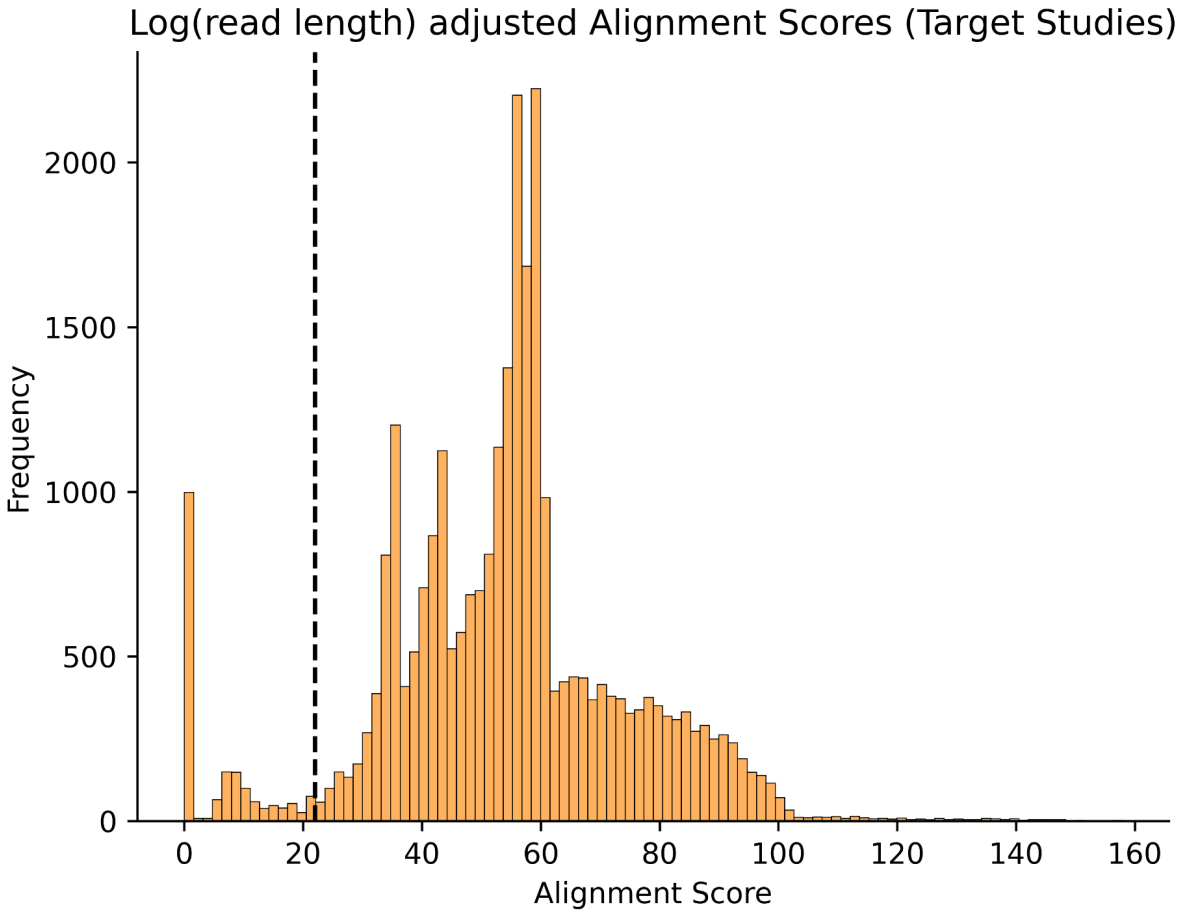

Figure S7: Normalized alignment scores of all reads of target studies (Brinch, Rothman, Spurbeck, and Crits-Christoph). Alignment scores are normalized by dividing the BowTie2 alignment score by the natural logarithm of the read length. The black, dashed line shows the cut-off alignment score under which Kraken read assignments were discarded.

Supplementary Tables

Study	Relative abundance (fraction of all reads)	
	All viruses	Human-infecting viruses
Bengtsson-Palme 2016	1.53×10^{-4} (1 in 6553)	1.2×10^{-7} (1 in 8324122)
Brinch 2020	3.59×10^{-3} (1 in 278)	2.24×10^{-6} (1 in 446026)
Brumfield 2022	5.27×10^{-3} (1 in 189)	1.52×10^{-6} (1 in 658766)
Crits-Christoph 2021	5.37×10^{-3} (1 in 186)	5.01×10^{-6} (1 in 199529)
Maritz 2019	8.49×10^{-4} (1 in 1177)	9.23×10^{-7} (1 in 1083027)
Munk 2022	2.42×10^{-3} (1 in 413)	2.73×10^{-6} (1 in 366316)
Ng 2019	9.04×10^{-5} (1 in 11059)	4.02×10^{-7} (1 in 2487173)
Rothman 2021	4.2×10^{-2} (1 in 23)	2.43×10^{-6} (1 in 410935)
Spurbeck 2023	8.38×10^{-5} (1 in 11927)	1.36×10^{-6} (1 in 732757)
Yang 2020	2.08×10^{-1} (1 in 4)	6.38×10^{-4} (1 in 1567)
All studies	1.93×10^{-3} (1 in 516)	2.26×10^{-6} (1 in 441830)

Table S1: Average relative abundance of all viruses (left) and human-infecting viruses (right) in all studies included in Figure 1. Relative abundance values for each study are given as the geometric mean relative abundance across all samples in that study. The “all studies” row gives the geometric mean across the average values for each study,

Study	Bioprojects
Bengtsson-Palme 2016	PRJEB14051
Brinch 2020	PRJEB13832, PRJEB34633
Brumfield 2022	PRJNA812772
Crits-Christoph 2021	PRJNA661613
Maritz 2019	PRJEB28033
Munk 2022	PRJEB13831, PRJEB27054, PRJEB27621, PRJEB40798, PRJEB40815, PRJEB40816, PRJEB51229
Ng 2019	PRJNA438174
Rothman 2021	PRJNA729801
Spurbeck 2023	PRJNA924011
Yang 2020	PRJNA645711

Table S2: Bioproject IDs for all included studies:

Study	Metagenomic protocol	Sample type	Sampling location	Included
Rothman 2021	✓	✓	✓	Yes
Crits-Christoph 2021	✓	✓	✓	Yes
Spurbeck 2023	✓	✓	✓	Yes
Brinch 2020	×	✓	✓	Yes
Bengtsson-Palme 2016	×	✓	✓	No, no viral enrichment
Brumfield 2022	✓	×	✓	No, sampled from manhole
Maritz 2019	×	✓	✓	No, no viral enrichment
Munk 2022	×	✓	✓	No, no viral enrichment
Ng 2019	×	×	✓	No, protocol selects against viruses
Yang 2020	✓	✓	×	No, Samples originate from Xinjiang China, with little available public health data

Table S3: Inclusion criteria for study subselection. Studies were included if they sampled untreated wastewater at a wastewater treatment plant, and performed metagenomic sequencing that is biased toward viruses, while not biased toward a viral subset (e.g., no amplification of a distinct set of human viruses). The 2020 study by Brinch et al. was included because it is the largest DNA study that sourced its samples from a single location. This contrasts with other studies like the one by Munk, which collected samples from many locations around the globe, which complicates the creation of public health estimates.

Virus	Selection Group	Incidence/Prevalence	Exclusion Criterion
SARS-CoV-2	Group 1	Incidence	Not excluded
Influenza A&B Virus	Group 1	Incidence	Not excluded
Norovirus	Group 1	Incidence	Not excluded
Respiratory Syncytial Virus	Group 1	Incidence	Heavily suppressed during the coverage period of RNA studies
Rhinovirus	Group 1	Incidence	No precise public health data available
Enteric Adenovirus	Group 1	Incidence	No precise public health data available
Metapneumovirus	Group 1	Incidence	No precise public health data available
Herpes Simplex Virus 1 (HSV-1)	Group 1	Prevalence	Not excluded
Herpes Simplex Virus 2 (HSV-2)	Group 1	Prevalence	Not excluded
Cytomegalovirus (CMV)	Group 1	Prevalence	Not excluded
Epstein-Barr-Virus (EBV)	Group 1	Prevalence	Not excluded
Human Immunodeficiency Virus (HIV)	Group 1	Prevalence	Not excluded
Hepatitis B Virus (HBV)	Group 1	Prevalence	Not excluded
Hepatitis C Virus (HCV)	Group 1	Prevalence	Not excluded
Human Papilloma Virus (HPV)	Group 1	Prevalence	Not excluded
Adeno-Associated Virus 2	Group 2	Prevalence	Not excluded
John-Cunningham Virus (JCV)	Group 2	Prevalence	Not excluded
BK Virus (BKV)	Group 2	Prevalence	Not excluded
Merkel Cell Virus (MCV)	Group 2	Prevalence	Not excluded
Hepatitis A Virus (HAV)	Group 1	Prevalence	No precise public health data available

Table S4: Exclusion criteria for human-infecting viruses

Virus	Study	Median	5th Percentile	95th Percentile
SARS-COV-2	Rothman	7.40×10^{-8}	1.91×10^{-8}	2.32×10^{-7}
SARS-COV-2	Crits-Christoph	1.39×10^{-7}	1.15×10^{-8}	7.66×10^{-7}
SARS-COV-2	Spurbeck	8.97×10^{-11}	5.15×10^{-12}	6.18×10^{-10}
Norovirus (GI)	Rothman	2.27×10^{-4}	1.03×10^{-4}	4.37×10^{-4}
Norovirus (GI)	Crits-Christoph	2.12×10^{-6}	5.14×10^{-8}	2.54×10^{-5}
Norovirus (GI)	Spurbeck	1.63×10^{-8}	1.10×10^{-9}	1.06×10^{-7}
Norovirus (GII)	Rothman	1.29×10^{-4}	3.88×10^{-5}	4.18×10^{-4}
Norovirus (GII)	Crits-Christoph	8.88×10^{-7}	4.15×10^{-8}	7.67×10^{-6}
Norovirus (GII)	Spurbeck	2.99×10^{-8}	3.82×10^{-9}	1.51×10^{-7}
Influenza A	Rothman	2.32×10^{-7}	6.14×10^{-9}	3.02×10^{-6}
Influenza A	Crits-Christoph	3.33×10^{-4}	4.46×10^{-6}	9.83×10^{-3}
Influenza A	Spurbeck	1.55×10^{-10}	3.22×10^{-12}	1.99×10^{-9}
Influenza B	Rothman	2.86×10^{-7}	1.86×10^{-8}	3.91×10^{-6}
Influenza B	Crits-Christoph	4.04×10^{-4}	3.89×10^{-6}	9.40×10^{-3}
Influenza B	Spurbeck	5.78×10^{-9}	8.60×10^{-11}	9.60×10^{-8}
AAV2	Rothman	2.06×10^{-12}	4.49×10^{-14}	3.20×10^{-11}
AAV2	Crits-Christoph	2.38×10^{-12}	2.65×10^{-14}	6.76×10^{-11}
AAV2	Spurbeck	8.35×10^{-11}	8.40×10^{-12}	5.45×10^{-10}
AAV2	Brinch	5.92×10^{-11}	1.48×10^{-11}	2.57×10^{-10}
BKV	Rothman	5.34×10^{-12}	1.58×10^{-13}	4.14×10^{-11}
BKV	Crits-Christoph	1.86×10^{-10}	8.72×10^{-12}	2.93×10^{-9}
BKV	Spurbeck	4.34×10^{-12}	2.24×10^{-13}	3.53×10^{-11}
BKV	Brinch	4.44×10^{-10}	2.09×10^{-10}	8.58×10^{-10}
CMV	Rothman	2.67×10^{-12}	5.55×10^{-14}	2.84×10^{-11}
CMV	Crits-Christoph	2.58×10^{-12}	2.90×10^{-14}	8.30×10^{-11}
CMV	Spurbeck	8.95×10^{-13}	2.46×10^{-14}	1.24×10^{-11}
CMV	Brinch	5.53×10^{-13}	1.30×10^{-14}	8.99×10^{-12}
EBV	Rothman	1.84×10^{-12}	3.69×10^{-14}	2.62×10^{-11}
EBV	Crits-Christoph	1.77×10^{-12}	1.50×10^{-14}	5.51×10^{-11}
EBV	Spurbeck	7.75×10^{-13}	1.68×10^{-14}	7.95×10^{-12}
EBV	Brinch	2.71×10^{-11}	7.14×10^{-12}	9.35×10^{-11}

HCV	Rothman	6.37×10^{-11}	1.33×10^{-12}	1.03×10^{-9}
HCV	Crits-Christoph	1.59×10^{-10}	2.09×10^{-12}	3.73×10^{-9}
HCV	Spurbeck	1.62×10^{-11}	3.22×10^{-13}	2.84×10^{-10}
HCV	Brinch	1.58×10^{-10}	5.70×10^{-12}	1.91×10^{-9}
HIV	Rothman	1.49×10^{-9}	3.45×10^{-11}	1.42×10^{-8}
HIV	Crits-Christoph	3.82×10^{-9}	1.17×10^{-10}	5.09×10^{-8}
HIV	Spurbeck	1.33×10^{-10}	2.18×10^{-12}	2.21×10^{-9}
HIV	Brinch	9.71×10^{-10}	2.40×10^{-11}	1.38×10^{-8}
HPV	Rothman	1.92×10^{-12}	3.50×10^{-14}	3.03×10^{-11}
HPV	Crits-Christoph	4.08×10^{-12}	3.39×10^{-14}	1.00×10^{-10}
HPV	Spurbeck	5.81×10^{-13}	1.17×10^{-14}	9.85×10^{-12}
HPV	Brinch	4.35×10^{-13}	8.79×10^{-15}	8.73×10^{-12}
HSV-1	Rothman	4.01×10^{-10}	1.41×10^{-10}	8.23×10^{-10}
HSV-1	Crits-Christoph	3.00×10^{-12}	4.21×10^{-14}	7.60×10^{-11}
HSV-1	Spurbeck	4.47×10^{-13}	8.07×10^{-15}	7.02×10^{-12}
HSV-1	Brinch	1.02×10^{-12}	5.20×10^{-14}	1.16×10^{-11}
JCV	Rothman	2.21×10^{-11}	1.08×10^{-12}	1.96×10^{-10}
JCV	Crits-Christoph	5.24×10^{-10}	3.30×10^{-11}	5.57×10^{-9}
JCV	Spurbeck	1.94×10^{-11}	1.64×10^{-12}	1.07×10^{-10}
JCV	Brinch	1.45×10^{-9}	6.22×10^{-10}	3.22×10^{-9}
MCV	Rothman	1.61×10^{-12}	2.31×10^{-14}	2.38×10^{-11}
MCV	Crits-Christoph	1.24×10^{-8}	5.15×10^{-9}	2.97×10^{-8}
MCV	Spurbeck	2.56×10^{-12}	8.84×10^{-14}	2.05×10^{-11}
MCV	Brinch	1.96×10^{-10}	8.48×10^{-11}	4.70×10^{-10}

Table S5: $RA_{i/p}$(1%) for selected viruses. Viruses covered: Norovirus GI, GII, SARS-CoV-2, HIV, MCV, BKV, JCV)

Virus	Difference at 5%	Difference at 25%	Difference at 50%	Difference at 75%	Difference at 95%
BKV	-2.56	-1.78	-1.46	-1.23	-1
JCV	-2.29	-1.56	-1.28	-1.1	-0.93
MCV	-0.6	-0.38	-0.29	-0.28	-0.27
Norovirus (GI)	-0.76	-0.56	-0.51	-0.48	-0.42
Norovirus (GII)	-1.75	-1.36	-1.17	-1.01	-0.88
SARS-COV-2	-0.76	-0.49	-0.42	-0.41	-0.44

Table S6: Log-fold RA(1%) difference between un-enriched (A, B, C, D, I, J) and enriched (E, F, G, H) Spurbeck 2023 samples, shown between 5th, 25th, 50th, 75th, and 95th percentiles respectively. Viruses covered: Norovirus (GI), Norovirus (GII), SARS-COV-2, MCV, JCV, BKV

Virus	Difference at 5%	Difference at 25%	Difference at 50%	Difference at 75%	Difference at 95%
BKV	-0.68	-0.38	-0.22	-0.17	-0.09
JCV	1.2	0.86	0.78	0.75	0.7
MCV	0.45	0.24	0.22	0.19	0.22
Norovirus (GI)	-0.46	-0.33	-0.27	-0.2	-0.13
Norovirus (GII)	-0.97	-0.86	-0.78	-0.72	-0.63
SARS-COV-2	-1.48	-1.12	-0.94	-0.8	-0.62

Table S7: Log-fold RA(1%) difference between the HTP site and the geometric mean of all other Rothman 2022 sampling sites, shown between 5th, 25th, 50th, 75th, and 95th percentiles respectively. Viruses covered: Norovirus (GI), Norovirus (GII), SARS-COV-2, MCV, JCV, BKV

Virus	Study	Median	25th Percentile	75th Percentile	Detection Threshold (reads)
Norovirus (GII)	Crits-Christoph	1.19×10^{7}	3.32×10^{7}	4.89×10^{6}	10
Norovirus (GII)	Rothman	7.57×10^{4}	1.17×10^{5}	4.92×10^{4}	10
Norovirus (GII)	Spurbeck	3.26×10^{8}	7.15×10^{8}	1.62×10^{8}	10
Norovirus (GII)	Mean (geometric)	6.65×10^{6}	1.41×10^{7}	3.39×10^{6}	10
SARS-COV-2	Crits-Christoph	6.07×10^{7}	1.49×10^{8}	2.54×10^{7}	10
SARS-COV-2	Rothman	1.28×10^{8}	2.09×10^{8}	8.21×10^{7}	10
SARS-COV-2	Spurbeck	1.12×10^{11}	3.07×10^{11}	4.81×10^{10}	10
SARS-COV-2	Mean (geometric)	9.54×10^{8}	2.12×10^{9}	4.64×10^{8}	10
Norovirus (GII)	Crits-Christoph	1.19×10^{8}	3.32×10^{8}	4.89×10^{7}	100
Norovirus (GII)	Rothman	7.57×10^{5}	1.17×10^{6}	4.92×10^{5}	100
Norovirus (GII)	Spurbeck	3.26×10^{9}	7.15×10^{9}	1.62×10^{9}	100
Norovirus (GII)	Mean (geometric)	6.65×10^{7}	1.41×10^{8}	3.39×10^{7}	100
SARS-COV-2	Crits-Christoph	6.07×10^{8}	1.49×10^{9}	2.54×10^{8}	100
SARS-COV-2	Rothman	1.28×10^{9}	2.09×10^{9}	8.21×10^{8}	100
SARS-COV-2	Spurbeck	1.12×10^{12}	3.07×10^{12}	4.81×10^{11}	100
SARS-COV-2	Mean (geometric)	9.54×10^{9}	2.12×10^{10}	4.64×10^{9}	100
Norovirus (GII)	Crits-Christoph	1.19×10^{9}	3.32×10^{9}	4.89×10^{8}	1000
Norovirus (GII)	Rothman	7.57×10^{6}	1.17×10^{7}	4.92×10^{6}	1000
Norovirus (GII)	Spurbeck	3.26×10^{10}	7.15×10^{10}	1.62×10^{10}	1000
Norovirus (GII)	Mean (geometric)	6.65×10^{8}	1.41×10^{9}	3.39×10^{8}	1000
SARS-COV-2	Crits-Christoph	6.07×10^{9}	1.49×10^{10}	2.54×10^{9}	1000
SARS-COV-2	Rothman	1.28×10^{10}	2.09×10^{10}	8.21×10^{9}	1000
SARS-COV-2	Spurbeck	1.12×10^{13}	3.07×10^{13}	4.81×10^{12}	1000
SARS-COV-2	Mean (geometric)	9.54×10^{10}	2.12×10^{11}	4.64×10^{10}	1000

Table S8: Weekly sequencing required for detection at 1% cumulative incidence. Viruses covered: Norovirus GII, SARS-CoV-2 Thresholds covered: 10 reads, 100 reads, 1000 reads

Virus	Study	Median	25th Percentile	75th Percentile	Detection Threshold (reads)
Norovirus (GII)	Crits-Christoph	$3,221	$8,955	$1,306	10
Norovirus (GII)	Rothman	$22	$33	$15	10
Norovirus (GII)	Spurbeck	$95,722	$209,318	$46,321	10
Norovirus (GII)	Mean (geometric)	$1,896	$3,962	$958	10
SARS-COV-2	Crits-Christoph	$20,543	$48,448	$10,390	10
SARS-COV-2	Rothman	$38,667	$62,425	$24,295	10
SARS-COV-2	Spurbeck	$31,877,527	$92,092,175	$13,087,461	10
SARS-COV-2	Mean (geometric)	$293,650	$653,058	$148,936	10
Norovirus (GII)	Crits-Christoph	$32,207	$89,550	$13,063	100
Norovirus (GII)	Rothman	$221	$332	$145	100
Norovirus (GII)	Spurbeck	$957,216	$2,093,181	$463,207	100
Norovirus (GII)	Mean (geometric)	$18,957	$39,623	$9,580	100
SARS-COV-2	Crits-Christoph	$205,432	$484,481	$103,903	100
SARS-COV-2	Rothman	$386,666	$624,247	$242,950	100
SARS-COV-2	Spurbeck	$318,775,271	$920,921,750	$130,874,613	100
SARS-COV-2	Mean (geometric)	$2,936,498	$6,530,585	$1,489,363	100
Norovirus (GII)	Crits-Christoph	$322,067	$895,496	$130,630	1000
Norovirus (GII)	Rothman	$2,210	$3,319	$1,453	1000
Norovirus (GII)	Spurbeck	$9,572,164	$20,931,813	$4,632,075	1000
Norovirus (GII)	Mean (geometric)	$189,567	$396,230	$95,798	1000
SARS-COV-2	Crits-Christoph	$2,054,321	$4,844,811	$1,039,032	1000
SARS-COV-2	Rothman	$3,866,659	$6,242,473	$2,429,503	1000
SARS-COV-2	Spurbeck	$3,187,752,706	$9,209,217,497	$1,308,746,129	1000
SARS-COV-2	Mean (geometric)	$29,364,976	$65,305,846	$14,893,629	1000

Table S9: Yearly sequencing cost for detection at 1% cumulative incidence. Cost is set at $5,500 per billion read pairs. Viruses covered: Norovirus GII, SARS-CoV-2 Thresholds covered: 10, 100, 1000.