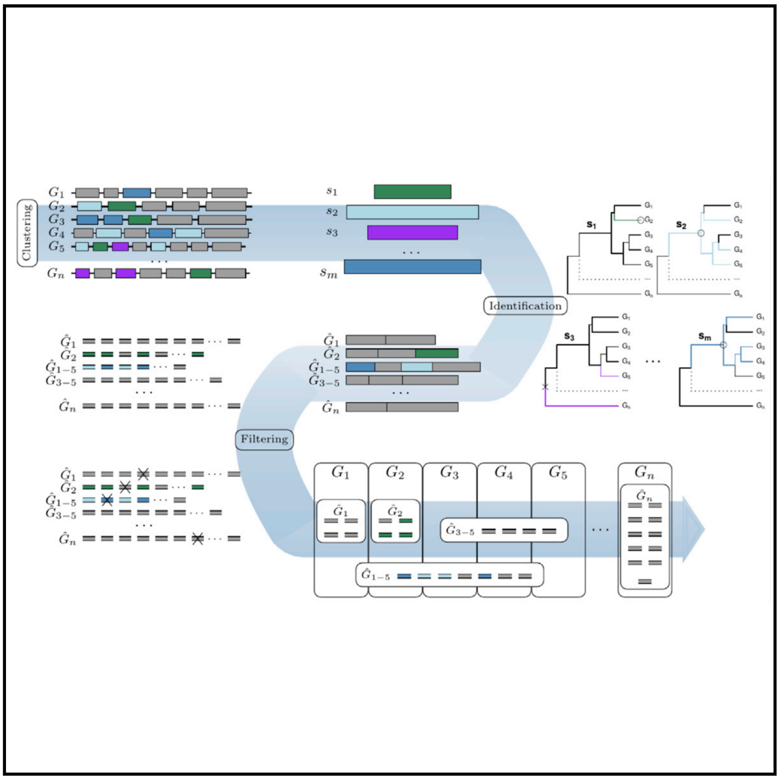
Cell Reports
Methods

Probe design for simultaneous, targeted capture of diverse metagenomic targets

Graphical abstract



Authors

Zachery W. Dickson, Dirk Hackenberger, Melanie Kuch, ..., Michael G. Surette, Geoffrey Brian Golding, Hendrik Poinar

Correspondence

dicksoz@mcmaster.ca (Z.W.D.),
poinarh@mcmaster.ca (H.P.)

In brief

Dickson et al. present HUBDesign: a pipeline that can be used to design probes for targeted DNA capture in contexts where high levels of background sequences are expected or it is unknown which of a broad set of target organisms of interest will be present.

Highlights

- HUBDesign identifies probes for specifically capturing a wide range of targets
- Probe sets were designed and validated targeting coronaviruses and sepsis pathogens
- Significant, specific, and simultaneous enrichment was observed for all targets

Dickson et al., 2021, Cell Reports Methods *1*, 100069
October 25, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.crmeth.2021.100069

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Probe design for simultaneous, targeted capture of diverse metagenomic targets

Zachery W. Dickson,[1,10,*] Dirk Hackenberger,[2,3] Melanie Kuch,[4] Art Marzok,[2,3,5] Arinjay Banerjee,[3,5,6,7] Laura Rossi,[2,3] Jennifer Ann Klowak,[8] Alison Fox-Robichaud,[9] Karen Mossmann,[3,5,9] Matthew S. Miller,[2,3,5] Michael G. Surette,[2,3,9] Geoffrey Brian Golding,[1] and Hendrik Poinar[2,3,4,*]

[1]Department of Biology, McMaster University, Hamilton, ON L8S 4K1, Canada
[2]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON L8S 4K1, Canada
[3]Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON L8S 4K1, Canada
[4]McMaster aDNA Center, Department of Anthropology, McMaster University, Hamilton, ON L8S 4L9, Canada
[5]McMaster Immunology Research Center, McMaster University, Hamilton, ON L8S 4K1, Canada
[6]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON L8S 4K1, Canada
[7]Vaccine and Infectious Disease Organization, Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, SK S7N 5E3, Canada
[8]Department of Pediatrics, McMaster University, Hamilton, ON L8S 4K1, Canada
[9]Department of Medicine, McMaster University, Hamilton, ON L8S 4K1, Canada
[10]Lead contact
*Correspondence: dicksoz@mcmaster.ca (Z.W.D.), poinarh@mcmaster.ca (H.P.)

https://doi.org/10.1016/j.crmeth.2021.100069

MOTIVATION A wide array of metagenomic research efforts are hampered by the same challenge: low concentrations of targets of interest combined with overwhelming amounts of background signal. Although PCR or naive DNA capture can be used when there are a small number of organisms of interest, design challenges become untenable for large numbers of targets. We present HUBDesign, a bioinformatic pipeline that designs probes for targeted DNA capture, which leverages sequence homology to identify probe sets that maximize the breadth of coverage for targets while maintaining specificity.

SUMMARY

The compounding challenges of low signal, high background, and uncertain targets plague many metagenomic sequencing efforts. One solution has been DNA capture, wherein probes are designed to hybridize with target sequences, enriching them in relation to their background. However, balancing probe depth with breadth of capture is challenging for diverse targets. To find this balance, we have developed the HUBDesign pipeline, which makes use of sequence homology to design probes at multiple taxonomic levels. This creates an efficient probe set capable of simultaneously and specifically capturing known and related sequences. We validated HUBDesign by generating probe sets targeting the breadth of coronavirus diversity, as well as a suite of bacterial pathogens often underlying sepsis. In separate experiments demonstrating significant, simultaneous enrichment, we captured SARS-CoV-2 and HCoV-NL63 in a human RNA background and seven bacterial strains in human blood. HUBDesign (https://github.com/zacherydickson/HUBDesign) has broad applicability wherever there are multiple organisms of interest.

INTRODUCTION

Several critical monitoring, clinical, and research efforts are hampered by the same challenge: low concentrations of targets of interest combined with overwhelming amounts of background signal. Whether it be monitoring the reservoirs, disease ecology, and transmission of zoonotic infections, such as COVID-19 (Rodriguez-Morales et al., 2020; Boni et al., 2020), or attempting to determine which of a huge array of potential pathogens is present in a patient displaying sepsis, the combination of low signal in a high background presents a significant challenge. Attempts to overcome this have been stymied by the difficulty associated with detecting or culturing these microbes (Wade, 2002; Papafragkou et al., 2014).

The advent of next-generation sequencing and the ever-declining cost of sequencing have made feasible a wide variety of research, including transcriptomics, ancient genomics, and microbial metagenomics. It is now viable to use RNA or DNA sequencing to rapidly identify organisms and characterize the diversity of nucleic acids in heterogeneous samples (Wang et al., 2019). However, there remain limits to sequencing depth and cost. In some cases, rare and interesting microbes might remain undetected.

Rare taxa can be clouded by high backgrounds from host or environmental sources, which often make up 99% of sequencing depth. In addition to the obscuring effects from sample backgrounds, differentiating true signals from contaminants becomes increasingly difficult, as the organisms of interest often make up a small fraction of the sample.

A naive approach of simply sequencing deeper is an unbiased, yet costly, way to overcome these issues. Pathogens in clinical or wildlife settings can easily make up less than 1 millionth of a sample, especially in the early stages of infection, where detection would be most useful for patients (Opota et al., 2015). Even with inexpensive sequencing costs, it becomes extremely wasteful and inefficient to spend sometimes critical time and resources to acquire and analyze these data when the majority is ultimately uninformative.

One way to alleviate the issue of cost is to bias detection toward targets of interest. Polymerase chain reaction (PCR) is one such technique used in many rapid detection systems (Tatti et al., 2011; Benirschke et al., 2019), including those used to detect individual sepsis pathogens and SARS-CoV-2, the causative agent of COVID-19 (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020). The technique relies on primers that bind to nucleic acid sequences specific to an organism or group of organisms. Although capable of sensitive, rapid detection and quantification of a particular target, PCR is limited when multiple loci are targeted by primers. Identifying "barcoding" regions has been used to amplify related organisms (Stahlberg et al., 2017; Adamowicz, 2015), and multiplexed PCR can allow for the amplification of multiple disparate targets (Hayden et al., 2008). The former is possible only for closely related groups, and the latter can be prone to bias and interference between the various primers in use (Elnifro et al., 2000). In addition, PCR is susceptible to failure in rapidly evolving organisms like viruses, where mutations occurring at priming sites can prevent amplification, as seen in SARS-CoV-2 (Rahman et al., 2020).

Another important technique in this area is microarrays. Oligonucleotide probes are designed to specifically hybridize to sequences of interest. These probes are then immobilized such that each probe sequence is in a known position, and the entire array is exposed to a DNA sample. Target sequences will be retained, whereas the remainder are washed away. Fluorescently labeling the libraries allows captured targets to be visualized to determine the presence of key taxa within a sample (Brown and Botstein, 1999). Such microbial detection arrays have demonstrated effectiveness (Gardner et al., 2010). However, they are limited to detection and identification of only known sequences, and there are challenges in efficiently designing probes to capture the targets of interest.

A complementary solution is targeted enrichment. Oligonucleotide probes are designed to hybridize to target nucleic acids; however, "capture" is performed in solution and preferentially retains them over non-target sequences (Mertes et al., 2011). This leads to an enrichment of the target in relation to the background and less effort and fewer resources expended on sequencing and identifying uninformative molecules. A major advantage over PCR is the ability to design probes capturing multiple loci simultaneously, like those designed to capture ~2,000 antimicrobial resistance genes (Guitor et al., 2019). Where identification and detection are important, capturing multiple independent loci in a genome provides more confidence of an organism's presence. Having multiple loci also assists in tracking variation between strains as they emerge and evolve.

The simplest probe design for a single organism is to select probes with a window that slides along the genome. Typically, each subsequent window overlaps the previous one. The resulting overlapping probes tile the target and are likely to be more effective than non-overlapping probes (Bertone et al., 2006). This approach can be extended to multiple organisms; however, the number of probes increases rapidly as more genomes are targeted, and this method makes no effort to ensure the probes are specific to the organisms of interest. Each additional genome adds its length in probes, increasing the chances for probe sequences to match multiple genomes. These matches are most often due to sequence homology between related organisms. As hybridization between probe and target is not perfectly specific (Mason et al., 2011), imperfect matches increase the chance of cross-reactivity and makes most of the probes generated in this manner redundant.

Fortunately, sequence homology and variable hybridization are beneficial for the design of more efficient probes that are capable of specifically and simultaneously capturing targets from known and novel members of a group of organisms. Although a probe will preferentially hybridize to its exact complement in a competitive environment, hybridization to sequences up to 20% divergent is possible (Mason et al., 2011; Delsuc et al., 2016). This has been used to design probes on the basis of sequences from extant organisms, facilitating the capture and enrichment of DNA from distantly related extinct taxa (Wagner et al., 2014; Enk et al., 2016; Delsuc et al., 2016). Increased success was obtained by designing probes on the basis of ancestral reconstructions (Delsuc et al., 2016). Ancestral reconstruction infers past character states from the diversity of modern states (Joy et al., 2016) and, in the context of probe design, might be seen as constructing a sequence representing the diversity of a set of input sequences. The representation might also capture diversity that is not represented by existing nodes on a tree. More generally, the concept of representative sequences will be used to design probes capable of capturing a broad set of sequences.

Genomes, especially those of bacteria and viruses, are mosaic in nature, and different genes and genomic regions offer unique evolutionary histories (Pedulla et al., 2003; Martin, 1999). As a result, hierarchical trees constructed on the basis of sequence similarity for entire organisms might differ from those constructed for individual genes (Goodman et al., 1979). Given a gene tree, a representative sequence can be constructed for each node in the tree by collapsing the sequences of all tips of the tree descended from that node. We have developed a pipeline that designs probes on the basis of representative sequences at multiple hierarchical levels (e.g., family, genus, species). The resulting probes specifically target and enrich

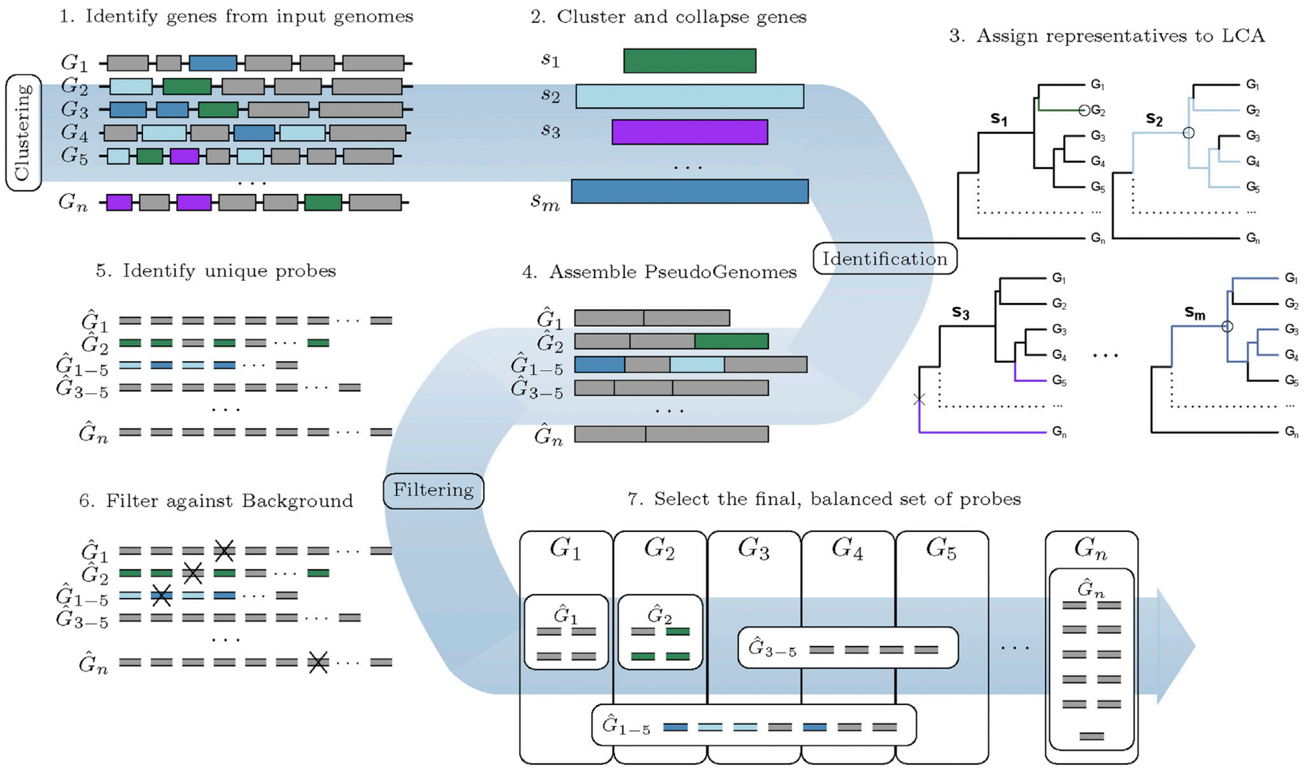


Figure 1. The workflow of the HUBDesign pipeline

HUBDesign takes annotated genomes as input. In step 2, genes with sequences that are at most 15% divergent (noted by color) are collapsed into representative sequences. In step 3, representative sequences are assigned to the lowest common ancestor (LCA) of organisms that possess the sequences represented. s_1 represents only one organism and is assigned to the leaf node. The LCA of organisms represented by s_1 and s_m is the same node, and both represent the majority of the five descendants of this node; they are considered valid representatives. The LCA for s_3 is the root of the tree; however, it is only representing a small fraction of the organisms in the tree. It is excluded in subsequent steps. In step 4, all representatives assigned to a node are concatenated into a pseudo-genome for that organism; note that s_2 and s_m were both assigned to the same node, and therefore are in the same pseudo-genome. In step 5, SA_BOND is run to identify all probes unique to each pseudo-genome. In step 6, probes that would capture off-target sequences are removed. In the final step, a set of probes that balances the number of probes per input genome is selected from the candidate probes.

nested clades, allowing for enrichment and identification of sequences from known and novel organisms.

Here we present and describe HUBDesign, a bioinformatic pipeline that leverages sequence homology and flexible DNA hybridization to design probes that can efficiently target sequences from a broad selection of organisms while maintaining specificity (Figure 1). To demonstrate the capabilities and effectiveness of HUBDesign, we have designed and tested two probe sets: a coronavirus probe set capable of simultaneously detecting all sequenced coronaviruses, and a set of probes targeting bacterial pathogens associated with sepsis.

RESULTS

Probe design

Multiple methods of designing probes were performed, and information on each is detailed in Table 1. Given the differences in breadth and depth of coverage, a comparable metric of efficiency was calculated as the average number of distinct genomes any given probe maps to. For the relatively small coronavirus dataset, the runtime (52 min) and effectiveness (1.87) of HUBDesign fall within the performance range of CATCH, given reasonable hybridization parameters (44–102 min and 1.13–4.23, respectively). However, HUBDesign is more memory efficient, which allows it to scale to the much larger sepsis dataset, for which CATCH failed with all tested parameter sets. Both methods produce more compact and efficient probe sets than a naive strategy.

The HUBDesign probe set for coronaviruses was tiled such that each taxon was targeted by approximately 400 probes. As seen in Figure 2, all genomes, where possible, are targeted by a minimum of 200 probes. The four lowest probe counts are for two gammacoronaviruses (turkey coronavirus, txid11152, 23 probes, and infectious bronchitis virus, txid11120, 8 probes) and two alphacoronaviruses (BtRf-AlphaCoV/YN2012, txid1503293, and *Rhinolophus* bat coronavirus HKU2, txid693998) with zero probes.

The majority (62.5%) of the probes have targets that are specific to one virus. Of the probes targeting multiple viruses, most (78.1%) target two or three. The remaining three sets of probes target loci specific to merbecoviruses and embecoviruses (both are *Betacoronavirus* subgenera) and loci common to the *Deltacoronavirus* genus. Both SARS-CoV-2 and HCoV-NL63 have probes at two levels in the hierarchy. For SARS-CoV-2 there are nearly 400 probes that target sequences common to

Table 1. Statistics for various probe sets produced

Dataset	Method	Number of probes	Nucleotide coverage (%)	Depth of coverage	Efficiency	Runtime (h)	Peak memory (GB)
Coronavirus	HUBDesign	13,500	25.0%	4.72*x*	1.87	0.87	0.5
	CATCH strict	3,846	20.0%	1.01*x*	1.13	1.7	6
	CATCH permissive	1,474	22.6%	1.29*x*	4.23	0.73	4
	naive	21,267	20%	5*x*	1	0.02	0.02
Sepsis	HUBDesign	26,870	2.09%	3.64*x*	29.3	6.1	7
	naive	2 million	2%	5*x*	1	3.5	7

SARS-CoV-2 and SARS-CoV-1 and an additional 400 probes that target *Sarbecovirus* sequences in general. Although there are no probes that target SARS-CoV-2-specific loci, the virus is easily differentiated by its sequence at those bait positions. HCoV-NL63 has 400 probes targeting *Setracovirus* sequences and an additional 4 probes that specifically target HCoV-NL63 loci.

The HUBDesign probe set for sepsis pathogens contained 26,870 probes targeting bacterial pathogens, covering 2.09% of all nucleotides in the input dataset at an average depth of coverage of 3.64*x*. A naive tiling achieving 2% coverage at 5*x* would require over 2 million probes. All 1,926 bacterial strains are targeted by probes that are at least at the genus level, and 53.3% of strains are targeted at the species level. The only genus that did not have any probes was *Clostridium*, but all strains in the genus were targeted at the species level. These species, *C. botulinum*, *C. perfringens*, and *C. tetani*, also had the lowest probe counts at 12, 44, and 71, respectively. The next lowest were *Rickettsia prowazekii* and *Borrellia burgdorferi* at 53 and 90 probes, respectively. All other species had at least 100 probes and an overall median of 478 probes per species. The seven spiked strains were targeted by at least 110 probes (*S. sanguinis*) and up to 564 probes (*B. multivorans*). *S. sanguinis* was the only spiked strain targeted at only the genus level, as it was not included in the dataset used to design the probes. Details on the numbers of probes per genus and species are in Table S6.

Coronavirus probe validation

Figure S1 shows how the amplicon levels compare with the rest of the coronavirus genomes. Although there is a peak in the coverage in this region, it is within the variability of nearby genomic regions. The estimated copy masses of HCoV-NL63 and SARS-CoV-2 are 29.5 and 10.8 ag/copy, respectively. Note that the copy mass of HCoV-NL63 was nearly 3× higher and therefore the nominal ratios based on copy number will not be represented in the sequencing results. For example, the "equal high" (EH) sample was prepared with an amount of viral extract expected to result in 20,000 copies of each virus. However, based on the shotgun baseline, the actual amounts of viral RNA were 215 and 589 fg of SARS-CoV-2 and HCoV-NL63, respectively. This gives a ratio closer to 1:3 rather than the original PCR estimated ratio of 1:1.

The proportion of reads assigned to SARS-CoV-2, HCoV-NL63, human, or otherwise can be seen for each sample in Figure 3. The proportion of viral reads is significantly and highly enriched in relation to a shotgun sample with the same amount of viral RNA. The combined number of reads assigned to either of the two spiked viruses was considered to be the number of on-target reads. This and the number of off-target reads were used to perform logistic regression. We observed fold enrichment on target reads of 97.6*x* (95% confidence interval [CI] 96.4*x*–98.9*x*). The summary of the logistic regression can be found in Table S8.

To examine the performance of individual probes, the two genomes were divided into alternating regions with and without probes. For example, the SARS-CoV-2 genome was broken into 20 regions, the first of which covers the first 4,950 bp in the genome and was targeted by 166 probes. It is followed by a 956 bp region not targeted by any probes, which is in turn followed by a 773 bp region with one probe at its center. Region boundaries were defined as 350 bp upstream and downstream of overlapping probes. This allows the analysis to account for the extended field of influence of probes resulting from capturing fragments with significant overhang. The following values were calculated for each region: the GC content, the number of probes, the average level of divergence from the genome sequence across the probes, and the fold enrichment. Differences in library size were accounted for by adjusting the enriched library's read counts by the relative size of the paired shotgun library. The fold enrichment for each region was calculated by dividing the observed read count in the enriched sample by the adjusted read count in the paired shotgun sample.

Because of the low sample concentration of viral RNA, there were several regions with no read coverage in the shotgun samples. The shotgun baseline was used to adjust shotgun read counts to reduce zeros. For each region, the proportion of reads from the shotgun baseline in that region was calculated. The adjusted read count for each region was the proportion in the shotgun baseline multiplied by an estimate of total reads across the genome taken from a weighted average across all regions.

Based on linear regression, increases in GC content from the genomic mean are negatively associated with fold enrichment. An increase in GC content of 11.3% is associated with a halving of fold enrichment; however, this effect is not significant in genomic regions targeted by probes. This amelioration of negative relationships when probes are present holds across all predictors. Another linear regression was also done, which used only regions that had probes. Surprisingly, GC content, probe divergence, and shotgun baseline levels had little if any significant effect. However, probe density was significantly and positively associated with fold enrichment. Every additional 77 probes/kb resulted in a doubling of fold enrichment. Both

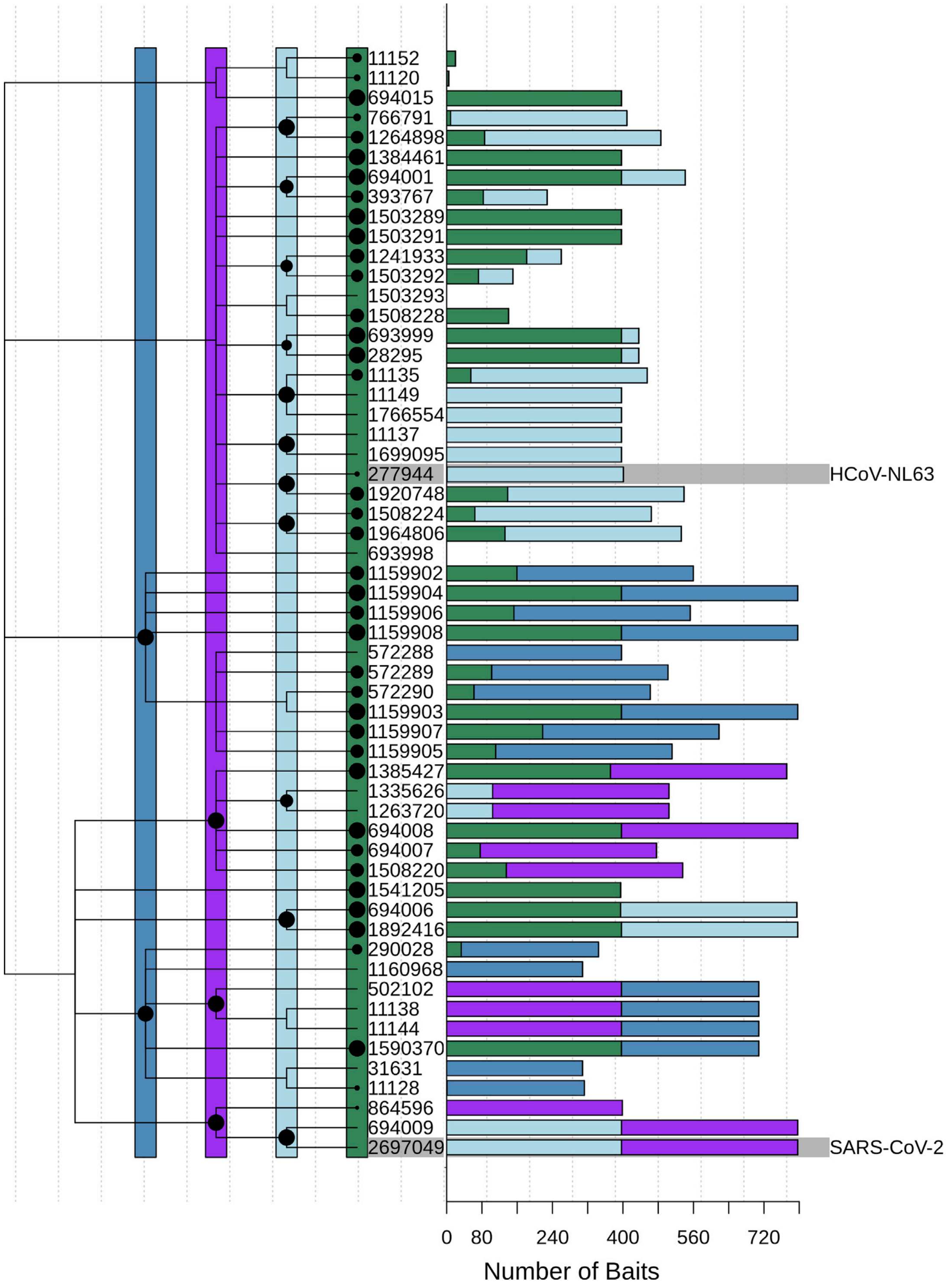


Figure 2. Probe density of each hierarchical level for CoV genomes

Tip labels indicate taxon IDs for the viruses, and the size of each node and the width of the matching bar give the number of probes targeting that genome. The hierarchical level of the probes is color coded according to the node height in the dendrogram. The two viruses used are highlighted with gray bars.

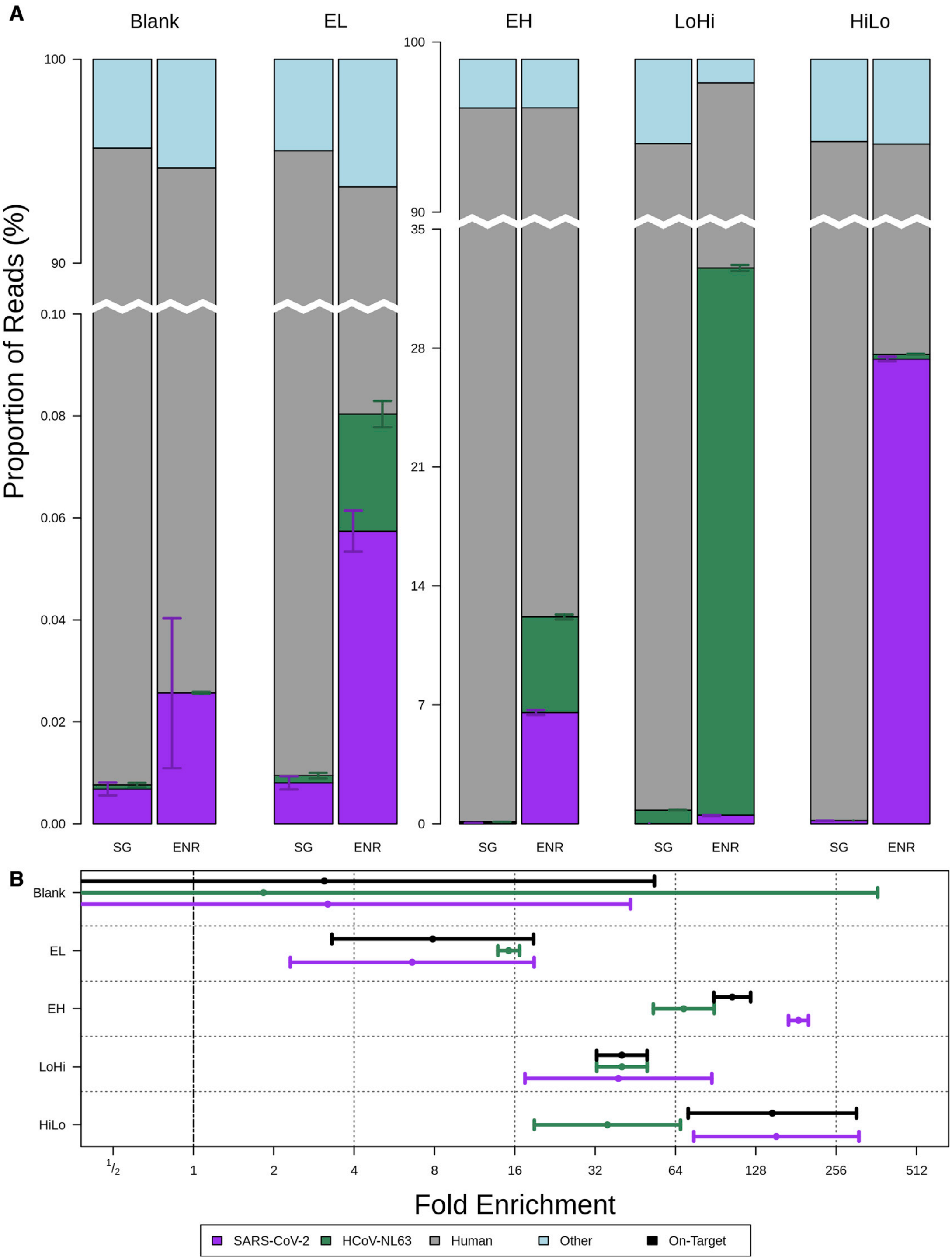


Figure 3. Fold enrichment of on-target CoV reads

(A) The proportion of reads assigned to the two spiked viruses and the human background. Reads assigned to any other taxa are grouped together. The right column in each pair represents the sample enriched with the probes. Error bars are the 95% confidence interval on the proportion. The blank and equal low (EL) samples are on a different scale to show that enrichment is observable even at the lowest tested viral concentrations.

(B) The fold enrichment of on-target viral RNA in each sample. Whereas the proportion of SARS-CoV-2 in the blank is significantly greater in the enriched sample, the calculated fold enrichment is insignificant. See also Table S8.

regressions showed a positive relationship with viral load over the range of viral loads tested. The relationship was weaker for HCoV-NL63, which had more RNA per nominal copy, consistent with expected diminishing returns in fold enrichment at high viral loads. A complete summary of these regressions can be found in Table S9. GC content and baseline shotgun levels have a significant effect only in probe-free regions. It is likely that these factors increase the number of sequenced reads overall, rather than affecting the enrichment specifically.

Figure 4 shows the fold enrichment, probe coverage, and GC content at each position in both genomes. There is no discernible relationship between GC content and fold enrichment, especially given the correlation between probe coverage and GC content. HUBDesign's selection of probes is based on finding unique sequences, and biased nucleotide content makes unique sequences less likely. Coronaviruses have an average of 30% GC content overall, but there are genomic regions with GC content at parity with AT content. As unique sequences are more likely to be found in these regions, this explains the correlation between increasing GC content and higher numbers of probes.

Sepsis probe assessment

The proportion of reads assigned to each of the spiked strains, with all *Streptococcus* spp. grouped together, as well as the proportion of human or other organisms, can be seen in Figure 5. Enrichment of the spiked taxa, but not the human background, is observed in the blood blanks. We detected genomic sequences from every spiked strain in the blank shotgun samples, and these contaminant sequences were captured by the probes intended to do so. Enrichment of sequences targeted by probes, but unintentionally present in the samples is also apparent when examining the ''Other'' category. The majority of these reads are assigned to probes targeting *Shigella* (69%) and *Escherichia* (30%) sequences. Adjusting for library size, there were 466*x* more reads for these two genera in the water blank samples than the blood blank samples. Fold enrichments estimated with logistic regression were 11.8*x* (95% CI 8.87*x*–15.7*x*) in the Low sample, 64.3*x* (95% CI 40.1*x*–103*x*) in the Medium sample, and 18.6*x* (95% CI 12.4*x*–27.9*x*) in the High sample.

To assess the difference in performance between probes targeting at the genus and species levels, all reads were remapped competitively to the genomes of the spiked bacterial strains, and BLASTn was used to disambiguate reads that mapped to multiple positions within a genome or reads that mapped to multiple genomes. The genomes were broken up into regions targeted by probes at each taxonomic level and one large region composed of all untargeted genomic regions. Within each region the log ratio of enriched reads to shotgun reads was calculated. The difference in library depth was accounted for by adjusting read counts in the larger library down by the ratio in size between the two libraries. Linear regression was used to account for properties of the baits and assess the difference between species level and genus level probes. This difference was significant only for *S. aureus*, which also had the greatest disparity between the number of regions targeted at the genus and species level (Table S10). In all cases, the variation due to properties of the probes, especially probe density and probe divergence, was larger than the variation due to taxonomic level. The performance in the probe regions across the spiked strains can be seen in Figure 6.

DISCUSSION

The HUBDesign pipeline was able to rapidly design a compact and efficient probe set covering almost every one of the targeted coronaviruses. Overall design time was less than a day, the majority of which was spent exhaustively filtering candidate probes against the human genome. The collapsing of the genomes into representative sequences required less than an hour, and the identification of candidates was completed in under a minute and required less than 1 GB of memory. However, processing 56 viral genomes is a minor task compared with the capabilities of the pipeline. Memory requirements for candidate identification scale linearly with genome size and number of organisms, but time requirements grow much more quickly. The pipeline has also been tested on three other input sets. Table S5 details the performance of SA_BOND on each set of organisms. These datasets come from different stages in the development of the HUBDesign pipeline, but the SA_BOND step has remained constant, and it is also the most memory-intensive step. It should be noted that the amount of diversity within a set of organisms is very important. For example, a dataset composed of 1,473 common gut bacteria was tested. Although there were fewer genomes than in the sepsis dataset, those genomes were spread over 161 genera compared with 35 for sepsis. As a result, there was less overall sharing of sequences, allowing for less information to be collapsed into representative sequences, and the corresponding runtime of SA_BOND was longer.

There were two viruses for which no candidate probes were found at any taxonomic level: txid1503293 and txid693998. In the tree used during assignment of gene clusters, these two viruses are labeled as alphacoronaviruses, without any indication of being more closely related (Figure 2). However, all but one of their genes consistently formed clusters that were distinct from all other viruses. Given the tree, the apparent lowest common ancestor node for these two is the *Alphacoronavirus* node. As the two viruses represent only 8.7% of the strains in the genus, their clusters did not meet the penetrance threshold; thus, their sequences were not included in the pseudo-genome. Without any sequence to select from, no probes were found for these viruses. We have recognized this and the current version of HUBDesign constructs a new tree to be used during cluster assignment, which is based on the observed clusters and optionally guided by a user-provided tree to resolve ambiguities.

Eleven mutations across SARS-CoV-2 isolates have been identified, which can be used to classify the virus into five clades (Guan et al., 2020). Our probes, which target SARS-CoV-2, cover these mutations well. Four of the loci are in positions directly covered by the probes, five more have a probe within 100 bp, and the remaining two are 267 and 546 bp from the nearest probe. The validation results demonstrate that enrichment of genomic regions adjacent to the probes occurs at least as far as 350 bp, if not farther (Figure 4). For these loci, all positions demonstrated an average of at least 3-fold enrichment across all samples, with the two loci farthest from a probe averaging 17- and 19-fold enrichment, respectively.

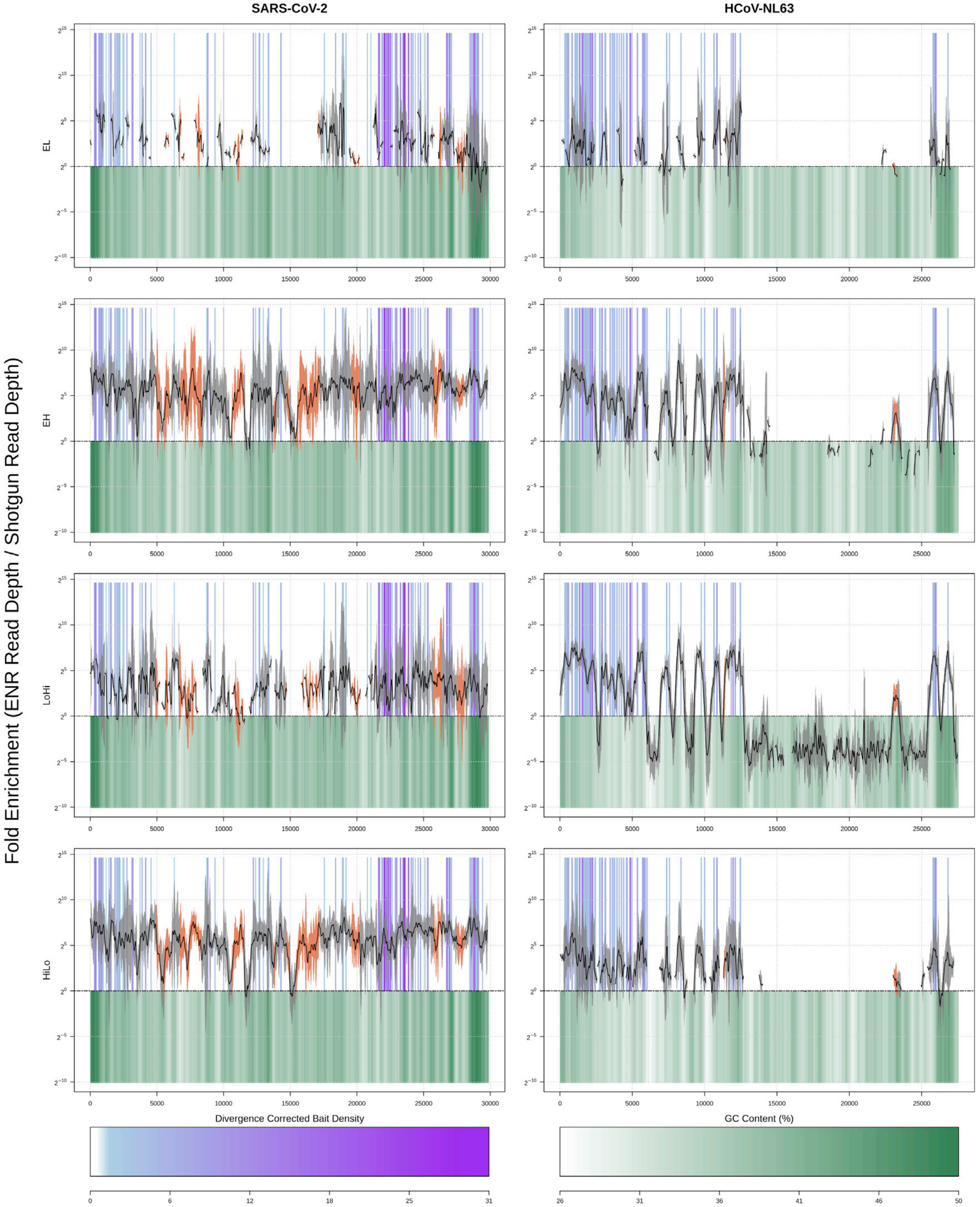
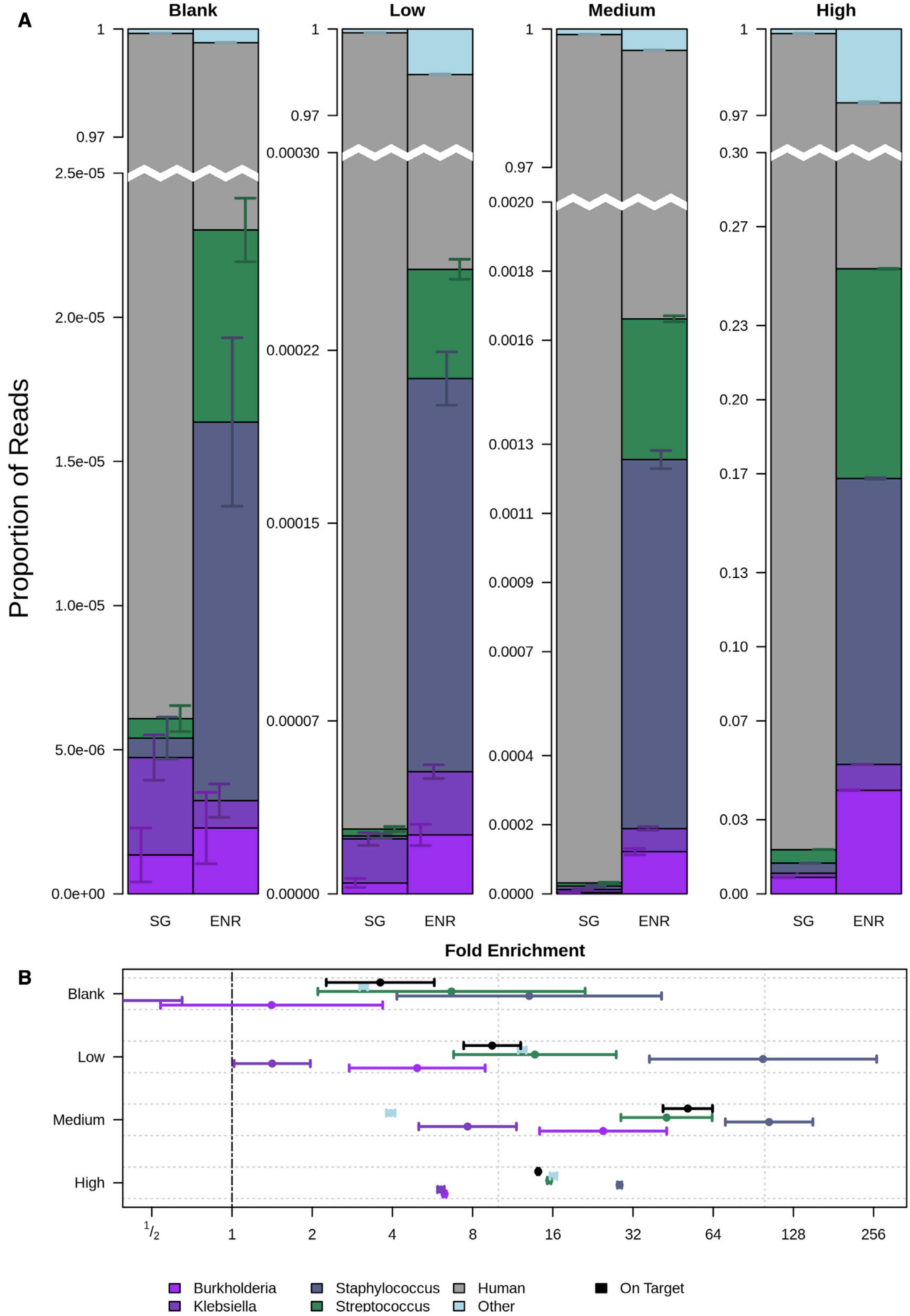


Figure 4. Fold enrichment of CoV genomic positions

Each row is a single set of samples in the order EL, EH, LoHi, and HiLo. The gray area indicates 3 standard deviations around the mean. The number of probes covering a particular position is indicated by the blue-purple intensity. This number of probes is adjusted by the divergence, such that more divergent probes contribute to the density less. GC content is indicated by the green intensity. Breaks in the line indicate that there were no reads covering that position in enrichment samples. Regions with apparent enrichment without being targeted by a probe are highlighted with orange confidence intervals; see also Table S7. Large gaps, or low fold enrichment regions, for HCoV-NL63 are strongly correlated with regions with no probe coverage. Probes cover a larger proportion of the SARS-CoV-2 genome, and there are fewer gaps or low fold enrichment regions of the genome.

See also Table S9 and Figure S2.



(legend on next page)

During analysis, we removed reads that mapped to amplicons generated during qPCR of the viruses. This was performed to reduce potential contamination from any amplicons that had escaped into the environment, which can occur relatively easily throughout the procedures (Rys and Persing, 1993). Analysis of the reads filtered out suggested that our approach was quite conservative, as of the blank samples (a total of 6 million reads), only one read mapped to an amplicon. The number of reads mapping to the amplicons correlates well (cor = 0.997) with the nominal copy number of SARS-CoV-2 in the sample, with one read mapping to an amplicon for every 10.45 nominal copies of the virus present ($p < 0.0001$). Despite conservative removal of these reads, which were mostly true viral reads, high levels of enrichment were observed.

In preparing each sample's read data, only one read was carried forward for each unique sequence observed. All other copies of that sequence were string duplicates, which could be true biological reads. When multiple copies of the genome are randomly fragmented, identical sequences can be produced. However, given the experimental setup, it is much more likely that these reads are the result of PCR duplication. To evaluate the effect of duplication, all analyses were performed again without the deduplication step. The mean ± SEM duplication rates in human and viral reads were observed at 8.2% ± 0.7% and 15.8% ± 5.5%, respectively, in shotgun samples. In enriched samples duplication rates were 20.3% ± 3.5% for human reads and 54.2% ± 5.6% for viral reads. Fold enrichment of viral reads was approximately doubled when reanalyzed without duplication, consistent with the approximately doubled rates of duplication in viral reads. The duplication rates for both human and viral reads were elevated in the enriched samples, consistent with going through additional rounds of PCR. As the viral genomes are much shorter than the human genome, it is much more likely for identical fragments to arise by chance. This might explain the elevated duplication rates in relation to human. Deduplication would then be reducing the true read count on viral reads, but failure to deduplicate artificially inflates read counts in the enrichment in relation to the shotgun. Despite our conservative approach by removing duplicates, we still observed significant enrichment of both SARS-CoV-2 and HCoV-NL63.

Although the specifics of individual probe performance vary depending on the resolution and method of analysis, the clear enrichment of sequences from both viruses is robust and apparent with every analysis we performed. The target sequences of probes at all relevant hierarchical levels were significantly enriched. Fold enrichment levels were highest in the EH pool and lowest, but still significant, in the equal low (EL) pool, the sample with the lowest viral input. The relationship does not appear to be linear. Fold enrichment is about double for SARS-CoV-2 compared with HCoV-NL63; however, there were also nearly twice as many probes targeting the former (796 probes) as the latter (402 probes). Although the number of probes targeting a single locus was not observed to have a significant effect, the global number of probes targeting an organism does have an effect. This emphasizes the importance of balancing probe numbers across organisms. The fact that there were nearly twice as many probes targeting loci in SARS-CoV-2, and that these loci were more evenly spread across the genome, contributes to the apparent difference in enrichment profile observed in Figure 4. Enrichment across the SARS-CoV-2 genome was relatively even, potentially because of the closely spaced probes.

There is evidence of enrichment in regions that are not targeted by probes for SARS-CoV-2 or HCoV-NL63. A notable example is the peak visible near position 23k of HCoV-NL63 in Figure 4. Three potential explanations for this are an extended field of influence for nearby probes, off-target capture between viruses, and within-genome off-target capture. We identified 14 regions that were significantly enriched and which were also at least 350 bp away from the nearest probe. Two were in HCoV-NL63, and the remainder in SARS-CoV-2. These regions are highlighted in Figure 4 and are detailed in Table S7.

We identified 45,791 unique molecules in these regions across all enriched samples, making up 7.1% of all viral molecules. The large majority (92.4%) of these molecules best match SARS-CoV-2, with nearly half of those mapping in the vicinity of on-target probes, consistent with the range of enrichment around a probe being larger than 350 bp. This can be explained by overhanging fragments. For example, if a 75 bp probe perfectly matches the end of a 300 bp fragment, this leaves 225 bp of the fragment to potentially capture overlapping fragments from the opposite strand. This allows the probe to enrich beyond its immediate target. If this were occurring, we would expect to see bias in the strandedness of our captured viral sequence data. The probes are designed to capture negative-strand cDNA synthesized by reverse transcribing the positive-strand RNA of the virus. Most of the captured fragments should then be from the negative strand, but fragments pulled down indirectly should be positive stranded. As can been seen in Figure S2, we do indeed see that most reads are negative stranded, but in regions flanking enriched areas the strand bias flips. These indirectly enriched areas in some cases explain regions where enrichment is occurring without targeted probes.

All off-target-enriched reads were remapped to probe regions excluding probes that capture HCoV-NL63 and SARS-CoV-2. Only 763 (1.7% of all reads in noted regions) of these molecules successfully mapped, but they always did so to a probe in the correct genera (HCoV-NL63 reads mapping to other alphacoronavirus probes, and SARS-CoV-2 to other betacoronavirus

Figure 5. Fold enrichment of on-target sepsis reads

(A) The proportion of reads assigned to four different spiked genera and the human background. Reads assigned to other taxa are grouped together. The right column in each pair represents the sample enriched with the probes. Error bars are the 95% confidence interval of the proportion. Each sample is on a different scale to show the enrichment of the targeted genera.

(B) The fold enrichment of non-human sequences in each sample. The error bars indicate 95% confidence intervals of the difference between log read counts in the enriched and matching shotgun libraries. The only insignificant enrichment among the spiked genera is for *Burkholderia* and *Klebsiella* in the blanks. The blanks in both (A) and (B) refer to libraries prepared from blood only without any spiked bacteria.

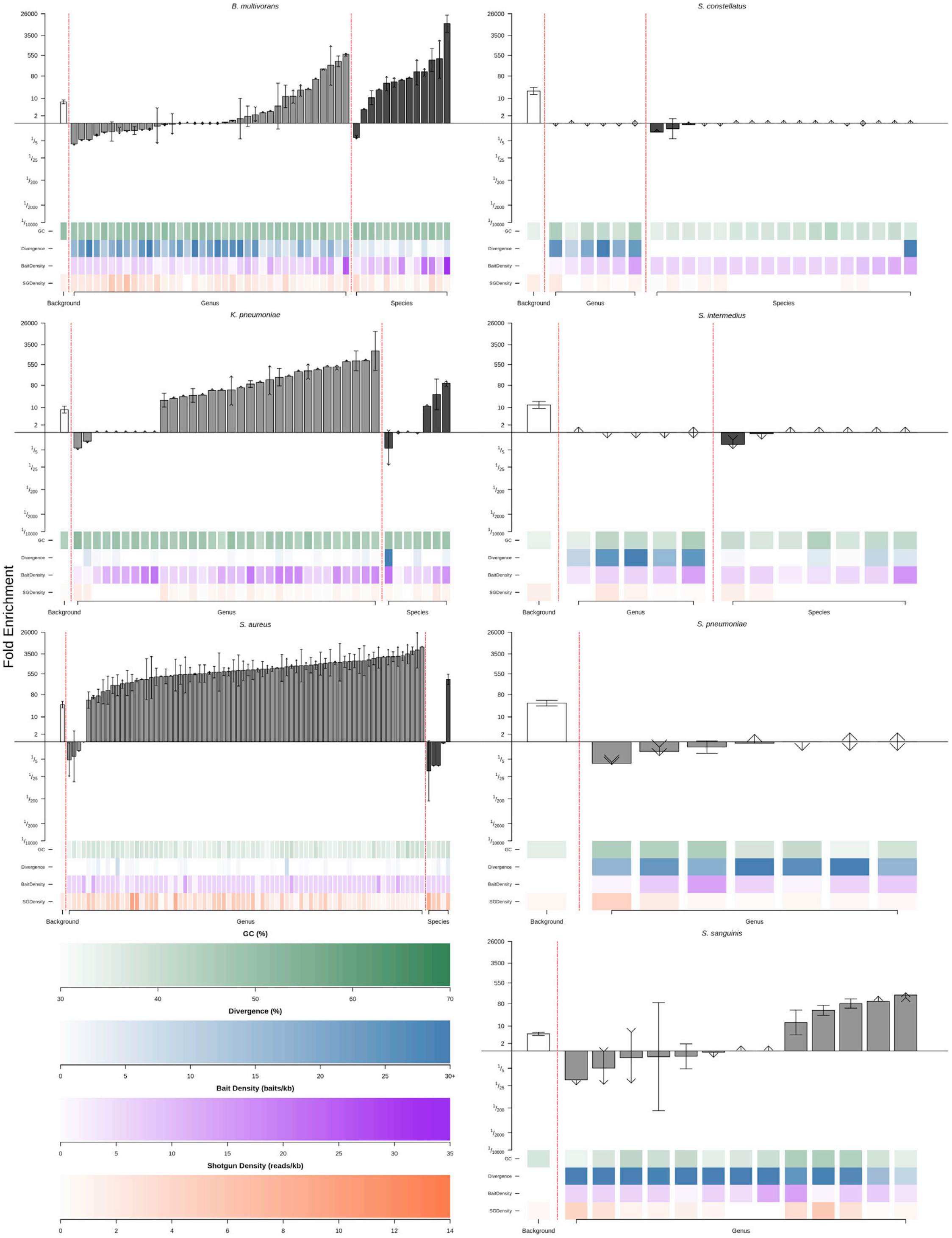


Figure 6. The fold enrichment within regions targeted by the sepsis probe set for each spiked strain

Bar heights indicate the fold enrichment observed for each genomic region. Bars are in ascending order of fold enrichment. Error bars represent the 95% confidence level of the difference between the log read counts of enriched and shotgun samples. The arrows on the error bars are an attempt to incorporate incomplete information. Upward-pointing arrows indicate that in some replicates reads were present in the enriched sample and not the shotgun sample, and therefore the finite values observed in the other replicates indicate the minimum fold enrichment. Downward-facing arrows conversely indicate that the finite observed values are the maximum fold enrichment. Arrows in both directions indicated both scenarios were observed,

(legend continued on next page)

probes). When the remappings are broken down by their region in the genome, only 3 of the 14 regions appear to be enriched by off-target probes. The region from positions 23k to 23.3k in the HCoV-NL63 genome appears to have been captured by a group of 27 probes targeting the same region of the camel alphacoronavirus (txid1699095), which is about 25% divergent from HCoV-NL63. Two regions of the SARS-CoV-2 genome covering positions 15.4k–17.2k appear to have been captured by three sets of probes all targeting similar genomic regions. The first set of 118 probes targets the betacoronavirus HKU24 (txid1590370), the second set of 36 probes targets rat coronavirus Parker (txid502102), and the third set is a single probe targeting *Rousettus* bat coronavirus HKU9 (txid694006). Due to the overlapping nature of probes, it is difficult to say which specific probe is responsible, except in the case of the txid694006 probe, which had 100 reads attributable to it. In general, the fold enrichment in these off-target regions is lower than in on-target regions, and the off-target effects would be expected to diminish if the true target of the probes were present. When not present, these probes can still be an advantage when the goal is hunting for novel organisms or identifying and monitoring members of a community.

The third possibility for apparent enrichment at a distance from a probe is off-target capture by probes within the same genome. The reads that map to the genome in probe-free enrichment regions were mapped directly to probes, and the positions targeted by those probes were compared with the original genomic position of the read. Of the 15,669 (34.2%) reads that mapped to a probe, almost all (97.9%) mapped to a probe directly adjacent to a probe-free enrichment region. This again indicates an extended field of influence of the probes. However, there were 330 reads that mapped to a probe targeting a position at least 1,000 bp away from the read's genomic position, with the farthest being nearly 27,000 bp away. These 330 reads fall exclusively into 13 of the 14 probe-free enrichment regions identified. The only apparent probe-free enrichment not at least partly explained by within-genome off-target effects is the region from 23k to 23.3k in HCoV-NL63.

A final concern for off-target enrichment is inadvertent capture of the background. Of the nearly 29 million reads that mapped to the human genome, only 57 (0.0002%) also mapped to a probe or the 350 bp region immediately up- or downstream of the probe. Broken down by sample, these reads are found more in the HiLo and LoHi samples (46 reads) than in the lower viral load samples (11 reads), and none were found in the negative samples. This is the opposite of what would be expected if the probes were enriching the human background. The amount of human input is at least 10,000 times higher than viral input in the samples with the highest concentration (and over 10 million times higher in the lowest). Therefore, the number of off-target reads would be expected to either be constant or decrease with higher viral load, as competition between the probe's true target and a partial human match would favor capture of the virus. As there is an increase with viral load, it is more likely that our human filtration step overzealously filtered out reads from the spiked viruses. Human background enrichment does not appear to be occurring for the sepsis probes either, as we observed 0.00055% of human reads also mapping to a probe. Most (67%) are in the blood blank samples, followed by the blood-free positive controls (22%). Whereas the same counterpattern as seen for the coronavirus probes was not observed, given the overwhelming amount of human DNA present, the low number of reads at worst indicates extremely inefficient off-target enrichment.

We also observed enrichment in untargeted genomic regions of the spiked bacterial species. These samples were prepared with a double-stranded library preparation protocol, and thus, the same strand bias as was seen for the coronaviruses would not be expected. Instead, we calculated the minimum distance to the nearest on-target probe for each read, and calculated the fold enrichment of reads at each distance. Although there was clear enrichment far from probe regions, there was almost none observed near to, but outside of, targeted regions (Figure S3). This indicates that daisy-chain enrichment was not a significant factor for these enrichments. This might be because of differences in library prep and the strandedness of the input nucleic acids. Another important factor is that the coronavirus genomes are orders of magnitude smaller, and the pool of fragments from which to sample hybridizations during enrichment is also less diverse. This makes it far more likely for a complementary fragment to be pulled down during enrichment.

The majority of observed off-target enrichment is the result of probes meant for other taxa targeting the spiked strains. There were 215 probes (0.8% of probes) that mapped to the spiked genomes but had different nominal targets. Of these, 178 nominally targeted another species in the correct genus. There were 31 probes that targeted at various levels within the non-*Klebsiella* members of the Enterobacteriaceae family. Most notable were probes targeting *Enterobacter aerogenes*; however, since these probes were designed, this bacterium has since been classified as a member of the *Klebsiella* genus. The remaining six probes are five *Streptococcus intermedius* probes mistargeting *K. pneumoniae*, and one *Klebsiella* probe mistargeting *S. sanguinis*. The Enterobacteriaceae probes highlight a flaw in the design for the sepsis probe set: the partitioning of clusters primarily on the basis of the nominal taxonomy, rather than observed sequence similarity. The sequences targeted by these probes are shared at the family level, but the design considered the genera independently and selected candidate probes that were not truly specific to their nominal targets. These flaws present in the older version of HUBDesign have already been corrected in newer versions of HUBDesign, as we continue to develop and improve it. The same cluster partitioning issue is likely the reason for some highly divergent probes being included in the probe set, especially for the *Streptococcus* spp. Despite these points to improve, the probes were able to enrich *S. sanguinis*, a strain

and there is very likely no enrichment or depletion. Color tracks below each bar plot indicate the bait or genomic properties likely to affect fold enrichment; note that bait divergence ranges from 0% to 30% or more. Although there are differences between the performances of genus-and-species-level baits, the direction of this is inconsistent, as the properties of the baits, especially bait divergence and bait density, have larger effects than bait level. See also Table S10 and Figure S3.

"unknown" to HUBDesign, and some genomic regions were enriched over 100× (Figure 6).

Although not among the bacteria intentionally spiked into the samples, we also observed significant enrichment (2×−16×) of reads mapping to *Shigella* and *Escherichia* probes. The latter is a common reagent contaminant (Salter et al., 2014), and these two bacteria are closely related. As a result of the cluster partitioning described above, the more numerous *Shigella* probes (Table S6) are also likely capturing contaminating *Escherichia* sequences. Both are also common human pathogens included in the design dataset for the sepsis probes. To avoid the capture of reagent contaminants, backlist databases of common contaminant sequences could be provided to HUBDesign during the filtration phase.

Overall, the fold enrichment observed in total reads for target organisms ranged between 10*x* and 100*x* for both probe sets. With individual regions of the viral genomes having mean fold enrichment up to 1,000*x* for the coronavirus probes and up to 20,000*x* for some *S. aureus* probe regions. This is comparable to the reported performance of CATCH (Metsky et al., 2019). When using a set of ~350,000 probes targeting 356 viral species on 30 patient samples with known viral infections, the median fold enrichment at genomic positions ranged from 1*x* to 53*x*. They observed fold changes for the number of reads for a virus within a sample as high as 1,000*x*. The enrichment achieved by HUBDesign's probes can be translated in savings to sequencing costs. To attain the same depth of coverage that we observed in our enrichments with a shotgun library, one would need to sequence 10*x* to 100*x* more deeply, with a commensurate increase in sequencing costs!

Limitations of the study

The validation experiments used artificial samples generated by pooling the desired background with genomic extracts from the targets of interest. Nucleic acids in patient samples and environmental extracts might have damage or modifications that reduce their availability for capture, reducing efficacy without altering specificity. As the performance of probes is dependent on the sequence properties of the targets, the level of enrichment will vary for each probe set produced.

HUBDesign relies on annotated genomes to efficiently cluster sequences. Probes cannot be designed for targets with unknown, or unannotatable genomes. HUBDesign might be able to capture these sequences by targeting closely related taxa. Development on the pipeline is ongoing to improve efficiency and reduce barriers to use.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Cell lines
  - ○ Viruses
  - ○ Bacteria
- METHOD DETAILS
  - ○ HUBDesign pipeline
  - ○ Coronavirus probe design
  - ○ Sepsis probe design
  - ○ Comparative probe design
  - ○ Viral RNA extraction
  - ○ Coronavirus sample preparation
  - ○ Bacterial sample preparation
  - ○ Sample analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100069.

ACKNOWLEDGMENTS

This work was partially supported by CIHR COVID-19 Rapid Response funding to M.S.M. M.S.M. was supported in part by a CIHR New Investigator Award and an Early Career Researcher Award from the Government of Ontario. A.B., H.P., G.B.G., and Z.W.D. were funded through the Natural Sciences and Engineering Research Council of Canada. This work was partially funded by a CIHR COVID grant to the principal applicant K.M. and co-applicant A.B. H.P. was generously funded by the Boris Family Fund, the Michael G. DeGroote Institute for Infectious Disease Research, and the CIFAR Humans and the Microbiome Program.

AUTHOR CONTRIBUTIONS

Conceptualization, A.F.-R., G.B.G., H.P., M.G.S., and Z.D.; methodology, H.P., D.H., M.K., H.P., and Z.W.D.; software, Z.W.D.; formal analysis, Z.W.D.; investigation, D.H. and M.K.; resources A.B., A.M., L.R., K.M., M.S.M., and M.G.S.; data curation, J.A.K. and A.F.-R.; writing – original draft, Z.W.D., D.H., A.B., and A.M.; writing – review & editing, all authors; visualization, Z.W.D.; supervision, A.F.-R., G.B.G., H.P., K.M., M.S.M., and M.G.S.; project administration, H.P.; funding acquisition, G.B.G., H.P., M.S.M., and K.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 22, 2020
Revised: June 10, 2021
Accepted: August 5, 2021
Published: September 15, 2021

REFERENCES

Adamowicz, S. (2015). International Barcode of Life: Evolution of a global research community. Genome *58*, 151–162.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Arbor Biosciences (2018). myBaits - Hybridization Capture for Targeted NGS, v4.01 (Daicel Arbor Biosciences).

Banerjee, A., Nasir, J., Budylowski, P., Yip, L., Aftanas, P., Christie, N., Ghalami, A., Baid, K., Raphenya, A., Hirota, J., et al. (2020). Isolation, sequence, infectivity, and replication kinetics of severe acute respiratory syndrome coronavirus 2. Emerg. Infect. Dis. *26*, 2054–2063.

Benirschke, R., McElvania, E., ThomsonRB, J., Kaul, K., and Das, S. (2019). Clinical impact of rapid point-of-care PCR influenza testing in an urgent care setting: a single-center study. J. Clin. Microbiol. *57*, e01281-18.

Bertone, P., Trifonov, V., Rozowsky, J., Schubert, F., Emanuelsson, O., Karro, J., Kao, M., Snyder, M., and Gerstein, M. (2006). Design optimization methods for genomic DNA tiling arrays. Genome Res. *16*, 271–281.

Boni, M., Lemey, P., Jiang, X., Lam, T., Perry, B., Castoe, T., Rambaut, A., and Robertson, D. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. *5*, 1408–1417.

Brown, P., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nat. Genet. *21*, 33–37.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.

Davis, J., Wattam, A., Aziz, R., Brettin, T., Butler, R., Butler, R., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E., et al. (2020). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res. *48*, D606–D612.

Delsuc, F., Gibb, G., Kuch, M., Billet, G., Hautier, L., Southon, J., Rouillard, J., Fernicola, J., Vizcaino, S., MacPhee, R., and Poinar, H. (2016). The phylogenetic affinities of the extinct glyptodonts. Curr. Biol. *26*, R155–R156.

Elnifro, E., Ashshi, A., Cooper, R., and Klapper, P. (2000). Multiplex PCR: optimization and application in diagnostic virology. Clin. Microbiol. Rev. *13*, 559–570.

Enk, J., Devault, A., Widga, C., Saunders, J., Szpak, P., Southon, J., Rouillard, J.M., Shapiro, B., Golding, G.B., Zazula, G., et al. (2016). Mammuthus population dynamics in late pleistocene North America: divergence, phylogeography, and introgression. Front. Ecol. Evol. *4*, 42.

Fehr, A., and Perlman, S. (2015). Coronaviruses: an overview of their replication and pathogenesis. Methods Mol. Biol. *1282*, 1–23.

Felsenstein, J. (1989). Phylip - phylogeny inference package (version 3.2). Cladistics *5*, 164–166.

Gardner, S., Jaing, C., McLoughlin, K., and Slezak, T. (2010). A microbial detection array (MDA) for viral and bacterial detection. BMC Genomics *11*, 668.

Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from Globin sequences. Syst. Zoolog. *28*, 132–163.

Guan, Q., Sadykov, M., Mfarrej, S., Hala, S., Naeem, R., Nugmanova, R., Al-Omari, A., Salih, S., Mutair, A., Carr, M., et al. (2020). A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int. J. Infect. Dis. *100*, 216–223.

Guitor, A., Raphenya, A., Klunk, J., Kuch, M., Alcock, B., Surette, M., McArthur, A., Poinar, H., and Wright, G. (2019). Capturing the resistome: a targeted capture method to reveal antibiotic resistance determinants in metagenomes. Antimicrob. Agents Chemother. *64*, e01324-19.

Hayden, M., Nguyen, T., Waterman, A., and Chalmers, K. (2008). Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. BMC Genomics *9*, 80.

Ilie, L., Mohamadi, H., Golding, G., and Smyth, W. (2013). BOND: basic Oligo-Nucleotide design. BMC Bioinformatics *14*, 69.

Joy, J., Liang, R., McCloskey, R., Nguyen, T., and Poon, A. (2016). Ancestral reconstruction. Plos Comput. Biol. *12*, e1004763.

Katoh, K., and Standley, D. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kim, D., Lee, J., Yang, J., Kim, J., Kim, V., and Chang, H. (2020). The Architecture of SARS-CoV-2 transcriptome. Cell *181*, 914–921.e10.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Martin, W. (1999). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays *21*, 99–104.

Mason, V., Li, G., Helgen, K., and Murphy, W. (2011). Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res. *21*, 1695–1704.

Mertes, F., Elsharawy, A., Sauer, S., vanHelvoort, J., vanderZaag, P., Franke, A., Nilsson, M., Lehrach, H., and Brookes, A. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief. Funct. Genomics *10*, 374–386.

Metsky, H., Siddle, K., Gladden-Young, A., Qu, J., Yang, D., Brehio, P., Goldfarb, A., Piantadosi, A., Wohl, S., Carter, A., et al. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nat. Biotechnol. *37*, 160–168.

Morgulis, A., Gertz, E., Schaffer, A., and Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comput. Biol. *13*, 1028–1040.

O'Leary, N., Wright, M., Brister, J., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.

Opota, O., Jaton, K., and Greub, G. (2015). Microbial diagnosis of bloodstream infection: towards molecular diagnosis directly from blood. Clin. Microbiol. Infect. *21*, 323–331.

Papafragkou, E., Hewitt, J., Park, G., Greening, G., and Vinje, J. (2014). Challenges of culturing human norovirus in three-dimensional organoid intestinal cell culture models. PLoS One *8*, e63485.

Pedulla, M., Ford, M., Houtz, J., Karthikeyan, T., Wadsworth, C., Lewis, J., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N., et al. (2003). Origins of highly mosaic mycobacteriophage genomes. Cell *113*, 171–182.

Rahman, M.S., Islam, M.R., Ul Alam, A.S.M.R., Islam, I., Hoque, M.N., Akter, S., Rahaman, M.M., Sultana, M., and Hossain, M.A. (2020). Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein (N protein) and its consequences. bioRxiv. https://doi.org/10.1101/2020.08.05.237339.

Reed, L., and Muench, H. (1938). A simple method of estimating fifty per cent Endpoints. Am. J. Epidemiol. *27*, 493–497.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. Trends Genet. *16*, 276–277.

Rodriguez-Morales, A., Bonilla-Aldana, D., Balbin-Ramon, G., Rabaan, A., Sah, R., Paniz-Mondolfi, A., Pagliano, P., and Esposito, S. (2020). History is repeating itself: probable zoonotic spillover as the cause of the 2019 novel Coronavirus Epidemic. Infez Med. *28*, 3–5.

Rys, P., and Persing, D. (1993). Preventing false positives: quantitative evaluation of three protocols for inactivation of polymerase chain reaction amplification products. J. Clin. Microbiol. *31*, 2356–2360.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Salter, S., Cox, M., Turek, E., Calus, S., Cookson, W., Moffatt, M., Turner, P., Parkhill, J., Loman, N., and Walker, A. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864.

Schoch, C., Ciufo, S., Domrachev, M., Hotton, C., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) *2020*, baaa062.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Stahlberg, A., Krzyzanowski, P., Egyud, M., Filges, S., Stein, L., and Godfrey, T. (2017). Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat. Protoc. *12*, 664–682.

Tatti, K., Sparks, K., Boney, K., and Tondella, M. (2011). Novel multitarget real-time PCR assay for rapid detection of Bordetella species in clinical specimens. J. Clin. Microbiol. *49*, 4059–4066.

Wade, W. (2002). Unculturable bacteria–the uncharacterized organisms that cause oral infections. J. R. Soc. Med. *95*, 81–83.

Wagner, D., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J., Enk, J., Birdsell, D., Kuch, M., Lumibao, C., et al. (2014). Yersinia pestis and the plague of Justinian 541-543 AD: a genomic analysis. Lancet Infect. Dis. *14*, 319–326.

Wang, P., Yan, Z., Yang, S., Wang, S., Zheng, X., Fan, J., and Zhang, T. (2019). Environmental DNA: an emerging tool in ecological assessment. Bull Environ. Contam. Toxicol. *103*, 651–656.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
SARS-CoV-2	Arinjay Banerjee	SARS-CoV-2/SB3-TYAGNC
HCoV-NL63	BEI Resources, NIAID, NIH	NR-470
Burkholderia Multivorans	ATCC	Cat#17616
Klebsiella pneumoniae	Micheal G Surette	N25C9
Sthaphylococcus aureus	Micheal G Surette	IIDRC0017
Streptococcus constellattus	Micheal G Surette	C1050
Streptococcus intermedius	Micheal G Surette	B196
Streptococcus pneumoniae	Micheal G Surette	R6
Streptococcus sanguinis	Micheal G Surette	GC83
Chemicals, peptides, and recombinant proteins		
FBS	Sigma Aldritch	Cat#F1051
DNase I	NEB	Cat#M0303S
Superscript III Reverse Transcriptase	Thermo Fisher Scientific	Cat#18080093
AMPure XP beads	Beckman Coulter	Cat#A63880
Critical commercial assays		
Luna Universal Probe One-Step RT-qPCR Kit	NEB	Cat#E3006S
Qubit RNA HS Assay Kit	Thermo Fisher Scientific	Cat#Q32852
NEBNext rRNA Depletion Kit	NEB	Cat#E6310S
SRSLY Nanoplus kit	Claret Biosciences	Cat#CBS-K150B-24
High Pure Viral Nucleic Acid Extraction large volume kit	Rocje Life Science	Cat#05114403001
NEBNext Ultra II DNA Library Prep Kit	NEB	Cat#E7645S
Deposited data		
Repository of Raw Sequencing data generated	SRA	BioProject: PRJNA674643
Experimental models: Cell lines		
Vero E6 cells	ATCC	Cat#CRL-1586
Oligonucleotides		
Primers for quantification of Viral RNA - See Table S3	IDT	RxnReady® Oligos
CoV Probes	Ann Arbor Biosciences myBaits custom DNA-Seq	https://github.com/zacherydickson/HUBDesign/tree/main/probes/Coronavirus
Sepsis Probes	Ann Arbor Biosciences myBaits custom DNA-Seq	https://github.com/zacherydickson/HUBDesign/tree/main/probes/Sepsis
Software and algorithms		
HUBDesign	this work	Zenodo: https://doi.org/10.5281/zenodo.5156877
distmat	Rice et al., 2000	http://emboss.sourceforge.net/download/
neighbor	Felsenstein,1989	https://evolution.genetics.washington.edu/phylip/getme-new1.html
BOND	Ilie et al., 2013	https://www.csd.uwo.ca/~ilie/BOND/
BLAST	Altschul et al., 1990	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
sdust	Morgulis et al., 2006	https://github.com/lh3/sdust
MAFFT	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/software/
Prokka	Seeman 2014	https://github.com/tseemann/prokka
CATCH	Metsky et al., 2019	https://github.com/broadinstitute/catch
Fastp	Chen et al., 2018	https://github.com/OpenGene/fastp
Prinseq	Schmieder and Edwards, 2011	http://prinseq.sourceforge.net/
Bwa	Li and Durbin, 2009	https://github.com/lh3/bwa
SAMtools	Li et al., 2009	http://www.htslib.org/download/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Zachery W Dickson (dicksoz@mcmaster.ca).

Materials availability

- This study did not generate new unique reagents.
- Generated Probe Sequences and associated metadata can be found at https://github.com/zacherydickson/HUBDesign/probes.

Data and code availability

- All sequencing data generated in the course of this work is available on the Sequence Read Archive under the BioProject accession PRJNA674643.
- The source code for HUBDesign is available under the terms of the GPL-3.0 license at https://github.com/zacherydickson/HUBDesign. https://doi.org/10.5281/zenodo.5156877.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

VeroE6 cells were cultured in Dulbec's modified Eagle medium containing 10% fetal bovine serum, 0.02M L-glutamine, 1,000 Units/mL Penicillin, and 1,000 μg/mL Streptomycin, as previosuly described (Banerjee et al., 2020)

Viruses

A clinical isolate of SARS-CoV-2 (SARS-CoV-2/SB3-TYAGNC) was propagated in Vero E6 cells and virus stocks were quantified and sequenced as previously mentioned (Banerjee et al., 2020). Virus stocks were maintained at −80°C. Work with SARS-CoV-2 was performed in a containment level 3 laboratory and all protocols were approved by the McMaster Presidential Biosafety Advisory Committee.

HCoV-NL63 (NR-470; BEI) was propagated on VeroE6 cells and viral titers were quantified using the 50% Tissue Culture Infectious Dose (TCID50) method. TCID50 values were determined using the Reed-Muench method (Reed and Muench, 1938). Viral stocks were stored at -80°C.

Bacteria

Seven bacterial strains were used in this work *B. multivorans*(17616; ATCC), *Klebsiella pneumoniae* N25C9, *Staphylococcus aureus* IIDRC0017, *Streptococcus constellatus* C1050, *Streptococcus intermedius* B196, *Streptococcus pneumoniae* R6, and *Streptococcus sanguinis* GC83. All strains with the exception of *Burkholderia multivorans* ATCC17616 were provided by Michael G Surette.

Frozen bacterial strains were independently cultured for 48 hours on agar.

METHOD DETAILS

HUBDesign pipeline

The Hierarchical Unique Bait Design (HUBDesign) pipeline aims to identify oligonucleotides (probes) which will specifically hybridize with nucleic acids from any member of a clade, and to do this for as many clades as possible within a given set of organisms. The probes designed are intended to enrich loci which are common to members of a clade, while being unique to that clade. This does not necessarily allow for whole genome enrichment, however by having probes at multiple hierarchical levels conserved genomic regions can be targeted by higher level probes while more variable regions are targeted by probes specific to a species or genome. The design of these hierarchical and unique probes is achieved through three design phases: clustering, identifying, and filtering.

In the clustering phase, sequences from the input organisms are grouped together and collapsed into representative sequences which are then used in the subsequent phase (Figure 1.2). To avoid computationally expensive all-vs-all comparisons, HUBDesign requires annotated genomes which allows rapid identification of gene families for clustering. The grouping of similar sequences serves as the source of hierarchy information and reduces computational effort. This reduction makes it possible to rapidly design probes for inputs ranging from dozens of viruses to thousands of bacteria. This implementation of the HUBDesign pipeline performs clustering on gene sequences as annotated using Prokka (Seemann, 2014). For each annotated gene family, sequences are aligned using MAFFT (Katoh and Standley, 2013), and then a neighbour-joining (Saitou and Nei, 1987) tree is generated based on the uncorrected edit distance. The uncorrected distance is used rather than the evolutionary distance as the actual difference between sequences is more important in a probe design context. Clusters are generated from the tree by selecting sub-trees which have a maximum root-to-tip divergence less than the maximum amount of divergence which still allows hybridization between a probe and target. A representative consensus sequence is generated for each gene cluster, and these sequences are assigned to the lowest common ancestor (LCA) of each organism represented (Figure 1.3). The representative sequence may not be appropriate to use if the represented organisms do not make up a significant portion of all descendants of the LCA. Laterally transferred elements are cases where a shared sequence is not useful in identifying a group of organisms. If an element is horizontally transferred to a distantly related organism, the representative sequence would be assigned to a node further from the tips of the tree. To prevent the use of these non-identifying sequences, a penetrance threshold is set. The penetrance for a representative sequence is calculated after its cluster has been assigned to the LCA of the genomes represented. The proportion of the LCA node's descendants which actually possess a member of the cluster is the penetrance of the representative sequence. All representative sequences assigned to a given node which pass the penetrance threshold are concatenated into a pseudo-genome for that node. The individual representative sequences are buffered to prevent selecting probes which straddle non-adjacent sequences. (Figure 1.4).

In the identification phase of the pipeline, pseudo-genomes are provided to a modified version of the program Basic OligoNucleotide Design (BOND) (Ilie et al., 2013) (Figure 1.5). BOND was originally designed for the rapid identification of a single unique oligonucleotide for each gene on a chromosome. Unique is defined as sharing no more than 15 consecutive identities and no more than 75% overall identity with any other oligo in the input. The entire program was modified to handle larger inputs allowing for the identification of multiple unique oligos for each genome in a set of genomes. The modified version (SA_BOND) is strand aware and tolerant of sequences which are repeated within a single genome. In terms of memory, this is the most computationally expensive phase in the pipeline. This is also the step where the specificity of the probes is improved far beyond a naïve tiling strategy. All probes are unique to the taxa they were designed for, allowing a probe targeting a taxon higher in the tree to capture all that node's descendants without also capturing unrelated organisms.

In the filtration phase, oligonucleotides are removed which hybridize to off-target or known background sequences like the human genome or transcriptome (Figure 1.6). This implementation of the pipeline utilizes BLAST (Altschul et al., 1990) to identify and exclude candidate oligos with significant hits against background. The thresholds for this can be set based on how conservative one wishes to be. Remaining candidate oligos which are overlapping are collapsed into contiguous regions, and then low-complexity intervals are excluded using sdust (Morgulis et al., 2006). The last step of the filtering phase selects the final set of oligonucleotides from the candidates in a manner which attempts to reduce bias between organisms. The goal is to have the number of probes targeting each organism to be as close as possible across the organisms (Figure 1.7). Probe count balancing is achieved by varying tiling density such that oligos targeting over-represented organisms are tiled less densely than oligos targeting under-represented organisms. It has been shown that higher tiling density can improve the capture efficiency of probes (Bertone et al., 2006). Varying tiling density in this way is a trade-off between the number of unique targets and the efficiency with which those targets are captured. The hierarchical nature of the probes constrains the ability to balance across organisms as probes can target multiple organisms which have different levels of coverage. HUBDesign takes and iterative approach to this problem. Tiling density is performed on probe regions which are composed of candidate probes with contiguous start positions and which all targeting the same taxon. The entire probe region is said to target that taxon, or all the descendent genomes if the taxon is at an internal node in the tree. On each iteration of the procedure a target number of probes per genome is set based on evenly dividing probes across organisms. Tiling strategies are determined for each probe region based on length, leaf taxa targeted, and a minimum tiling density specified by the user. If it would not be possible to bring the number of probes for an organism down to the target level even if all probe regions targeting that organism were tiled at minimum density, then all of those probe regions are assigned to be minimally tiled. If instead it is not possible to bring the number of probes for an organism up to the target even by tiling at maximum density (probes spaced apart by only 1 bp), then all probe regions targeting that organism are assigned to be maximally tiled. If a probe region targets both an under-targeted and an over-targeted organism, it will still be maximally tiled to ensure probes are available for the under-targeted organism. On subsequent iterations the target number of probes per organism is updated to reflect that the extremes are accounted for and should be excluded from consideration. Organisms with below target numbers of probes leave more probes available for the other organisms, and those with above target leave fewer. The new target is set by dividing the available probes evenly across the remaining organisms. Tiling classes are then reassigned based on this new target and iteration continues until tiling classes cease to change. Tiling density of the probe regions not assigned to the extremes are set in the order of constraint: Probe regions which target organisms which are targeted by the fewest probe regions are processed first. Within each organism, probe regions that target the most organisms are processed first. The tiling density is set as the weighted average of the minimum and maximum tiling density, weighted towards the max when there are few potential probes for the organism, or if the current number of probes is far from the target. Processing probe regions in this order allows coarse adjustments to the balance from the most constrained regions, and fine tuning from the least constrained.

Coronavirus probe design

In this section we describe the implementation of HUBDesign used to design probes for 56 coronaviruses taken from RefSeq (O'Leary et al., 2016). The set of viruses covers the *Alpha*-, *Beta*-, *Gamma*-, and *Deltacoronavirus* genera and includes the four major seasonally circulating human coronaviruses, as well as SARS-CoV-2 and viruses responsible for earlier novel coronavirus outbreaks. (Table S1).

In the clustering phase, distances were calculated using the distmat tool from EMBOSS (Rice et al., 2000), and neighbour joining trees constructed using the neighbour tool from the phylip package (Felsenstein, 1989). Clusters were generated from sub-trees with a maximum root-to-tip divergence of 15%. This threshold was selected as probe sequences which diverge from their targets by less than this are most likely to successfully hybridize (Mason et al., 2011; Delsuc et al., 2016). Each cluster was assigned to the LCA in a dendrogram based on the lineage recorded in NCBI's taxonomy database for each genome (Schoch et al., 2020). A penetrance threshold of 50% was used, therefore only representative sequences which were based on at least half of the descendants of the LCA were included in the pseudo-genome for the LCA. We observed that 90% of all clusters had at least 50% coverage, and this value ensured that pseudo-genomes representing all input taxa were constructed.

Candidate probes were identified using SA_BOND to search for all unique oligonucleotides of length 75 across all pseudo-genomes.

BLASTn was used to find and exclude any candidate probe which matched the human genome (GRCh38 (Schneider et al., 2017)). Matches were considered significant if they had at least 75% identity, were at least 30 bp long, and had an e-value less than 0.01. Low-complexity regions were excluded using sdust (Morgulis et al., 2006) with the default parameters: a 64 bp window, and a score threshold of 20. The latter approximately corresponds to a sequence where 80% of the nucleotide triples in the window are the same. The final probe set was selected with a target minimum tiling density of 5x and a maximum of 13500 probes. This is the smallest number of probes for which the most balanced distribution of probes across targets still allows for tiling densities to vary between 5x and the maximum (1bp spacing). Any fewer and the most balanced configuration has all probe regions tiled at either of these extremes. Probes were balanced by treating all hierarchical levels independently, which maximized the number of probes for taxa represented at only one hierarchical level.

Sepsis probe design

The sepsis probe set was produced with an earlier version of HUBDesign. The input database contained 1926 bacterial genomes across 81 species and 35 genera (Table S2). All genomes were acquired from the PATRIC database (Davis et al., 2020).

Gene clusters were generated and assigned based on nominal taxonomy. Each genus was independently and recursively processed. All genes of the same name with a minimum penetrance of 95% in the particular genus were aligned using MAFFT (Katoh and Standley, 2013). Then the conservation was calculated, and a consensus sequence was generated. A minimum of 85% conservation was required. Clusters meeting the penetrance and conservation thresholds were added to the genus's pseudo-genome, while those which failed to meet the criteria were broken up into clusters based on species. Each of these was realigned and tested against the thresholds once more. Only clusters which passed at the genus or species level were included in any pseudo-genome.

Up to 100 candidate 100bp probe regions per pseudo-genome were identified using SA_BOND. All 75 bp sub-sequences of these probe regions were considered candidate probes, and a maximum number of tiled probes were produced.

BLASTn was used to find and exclude any candidate probes which matched the human genome (GRCh38 (Schneider et al., 2017)). Matches were considered significant if they had at least 75% identity and were at least 20 bp long. All remaining contiguous probe regions after filtering were tiled with probes which were spaced apart by 5 bp. As 100 bp regions were identified this resulted in most loci being targeted by 5 probes.

Comparative probe design

As a baseline comparison, naïve tiling was performed on the genomes of the 56 reference coronavirus genomes, and on the 1926 sepsis pathogens. Probes were identified by selecting each 75bp subsequence of each genome, spaced apart by 15bp, and retaining only unique sequences.

A recently described computational method for probe design is the CATCH python package (Metsky et al., 2019). It can also be configured and applied to the task of finding probes, meant for identification. CATCH was run on the 56 reference coronavirus genomes. We selected two sets of hybridization parameters to account for the fact that CATCH and HUDesign use opposing approaches to probe selection. HUBDesign identifies candidate probes via elimination. A probe is only considered if it is unlikely to hybridize to another target in the dataset. As a result, liberal hybridization parameters make HUBDesign more strict and the resulting probes more specific. The opposite is true for CATCH. It identifies targets to which each probe will likely hybridize and selects an optimal set of probes with desired coverage. The two sets of hybridization parameters provided to CATCH were a strict set which did not allow mismatches, and a more permissive set allowing up to 18 mismatches but requiring an island of exact matches 15bp long. The permissive parameters are similar to those used by BOND (Ilie et al., 2013) to eliminate non-specific probes. Probes which were 75bp long and spaced out by 5bp were designed using the identify flag, targeting 20% coverage of the target genomes. The human genome (GRCh38 (Schneider et al., 2017)) was provided as a blacklist sequence.

Viral RNA extraction

For SARS-CoV-2 infections, 2×10^5 Calu-3 cells were seeded in each well of a 6-well plate. A clinical isolate of SARS-CoV-2 (SARS-CoV-2/SB3-TYAGNC (Banerjee et al., 2020)) was used to infect Calu-3 cells at a multiplicity of infection of 0.01. Cells were harvested 48 hours post infection and total RNA was extracted from infected Calu-3 cells using the QIAamp viral RNA Mini kit (Qiagen) according to the protocol outlined by Banerjee *et al.* (Banerjee et al., 2020).

Viral RNA from HCoV-NL63-infected VeroE6 cells was extracted using the Qiagen RNeasy kit with minor modifications. Briefly, 100 μL of supernatant was mixed with an equal volume of RLT lysis buffer and 25 μL of 20 mg/mL proteinase K (Invitrogen). Samples were vortexed and incubated at 56°C for 15 minutes. Following, 200 μL of 70% ethanol was added to the solution and RNA was eluted as outlined in the RNeasy manufacturer's protocol.

Human total RNA was extracted from peripheral blood mononuclear cells using the RNeasy extraction kit (Qiagen) according to manufacturer's protocols.

Coronavirus sample preparation

The copy number of both viral extracts was determined using the Luna Universal Probe One-Step RT-qPCR Kit (NEB) and the primers in Table S3, while human RNA was quantified using Qubit RNA HS Assay Kit (Thermo Fisher). All three sets of RNA were separately treated with DNase I (NEB), and the human ribosomal RNA depletion kit (NEB) according to manufacturer's protocols. Following this, samples were thermally fragmented to roughly similar fragment size distributions.

Four mock samples and a negative control were prepared by combining RNA extracts from the two viruses and total human RNA background. The samples were prepared at a low level of two hundred RNA copies and a higher level of twenty thousand RNA copies in a 1:1 ratio of both viruses (EL and EH). Two additional samples were prepared with two thousand RNA copies of one virus and two hundred thousand RNA copies of the other. In the LoHi sample SARS-CoV-2 was the low-level virus, and HCoV-NL63 was in the HiLo sample. All samples including the negative control contained one hundred nanograms total human RNA.

While qPCR copy numbers were used to prepare the mock pools, these copy numbers do not represent full genome copies present at that level, only that the PCR amplicon is present at that estimated copy number. Variation occurs as sub-genomic RNAs are created during the coronavirus life cycle (Kim et al., 2020; Fehr and Perlman, 2015). Different regions of the genome are therefore likely to be better represented than others, with the expectation that the 3′ regions of the genome will have the highest copy numbers (Figure S1).

To generate a confident baseline to account for this, an additional sample containing a mixture of the viruses was shotgun sequenced. The sample nominally contained 2.68 million copies of the HCoV-NL63 genome and 5.1 million copies of the SARS-CoV-2 genome.

All pooled samples were prepared in triplicate. First strand synthesis was performed using Superscript III Reverse Transcriptase (Thermo Fisher) according to manufacturer's protocols with 250ng random hexamers. This reaction was purified using AMPure XP beads (Beckman Coulter) at a 1.8X ratio to sample. Single-stranded libraries were then prepared using the SRSLY Nanoplus kit from Claret Bioscience. Attachment of indexing adapters was performed according to the protocol outlined by Kircher *et al.* (Kircher et al., 2012), after which the indexed libraries were purified using MinElute spin columns (Qiagen). Each replicate was split in two where one half was enriched prior to sequencing and the other shotgun sequenced.

Probes were synthesized through Ann Arbor Biosciences myBaits custom DNA-Seq program. Enriched samples were processed according to the Ann Arbor Biosciences myBaits targeted enrichment protocol version 4.01 (Arbor Biosciences, 2018) with a 72-hour capture step. After enrichment, both shotgun and enriched libraries were quantified alongside the Illumina PhiX standard, before being pooled to equimolar quantities and undergoing a gel-based size selection for fragments between 150-500 bp. Pools were then sequenced on an Illumina Hiseq 2x90 flow cell.

Bacterial sample preparation

Three mock pools were prepared in triplicate which contained human blood spiked with the seven bacterial strains listed above. With the exception of *S. sanguinis*, the genomes of the spiked bacteria were in the dataset used to design the probes. The three pools were at Low, Medium, and High concentrations (10^1, 10^3, and 10^6 CFU/mL) of each bacteria. One additional pool at each concentration was also prepared where all the bacteria were included but the blood was omitted. Three negative controls were prepared with water and five negative controls were prepared with blood only.

Cultures of each of the strains to be spiked were suspended and diluted to the above concentrations in 0.85% saline. At each of the Low, Medium, and High concentrations the matching CFU counts for each strain were pooled together and pelleted. Fresh human blood was drawn from a healthy donor into tubes containing EDTA as an anticoagulant. Bacterial pellets for each pool were resuspended in the blood, or sterile saline for the no blood control.

DNA was extracted using the High Pure Viral Nucleic Acid Extraction large volume kit from Roche, according to their version 7 protocol for a 1mL sample. Pools were sonicated to ~200bp using a Covaris S220 focused ultrasonicator. Pools were divided such that each pool would have one shotgun library and two enriched libraries. Samples for both shotgun and enrichment were processed for library preparation using the NEBNext Ultra II DNA Library Prep Kit for Illumina (CN E7645) according to manufacturer's specifications. Probes were synthesized through Ann Arbor Biosciences myBaits custom DNA-Seq program. Samples to be enriched were processed using the Arbor Biosciences myBaits v5.0 kit with the high sensitivity protocol, according to manufacturer's specifications with a 63°C hybridization temperature.

After enrichment, all samples including those prepared for shotgun sequencing were quantified using KAPA SYBR FAST Bio-Rad iCycler Master Mix (Sigma Aldrich, CN KK4608), run alongside PhiX Control Standards (Illumina, CN FC-110-3001). Based on these concentrations, samples were pooled to equimolar amounts, then concentrated using a Minelute PCR purification column (Qiagen, CN 28006). This concentrated pool was then size-selected using NuSieve GTG Agarose (Lonza, CN 50081) in a 3% 1X TAE gel. Only molecules with a total length between 200bp and 500bp were excised from the gel. The final pool was purified from the gel using a Minelute Gel Extraction column (Qiagen, CN 28604), and eluted in 20μL. This final pool was sequenced on an Illumina HiSeq 2x90 flow cell.

Sample analysis

After demultiplexing, samples were trimmed and merged using FastP (Chen et al., 2018). Any orphaned reads were treated as single ended reads going forward. Reads were string deduplicated using prinseq (Schmieder and Edwards, 2011) as identical reads are much more likely to be PCR duplicates generated during sample preparation than identical templates generated during fragmentation. String deduplication was selected over mapping-based deduplication as a balance between removing PCR duplicates and retaining true biological duplicates from high-copy numbers. Reads for were then filtered against the GRCh38 (Schneider et al., 2017) version of the human transcriptome (coronaviruses) or genome (sepsis) using BWA (Li and Durbin, 2009). Reads from coronavirus libraries were also mapped against amplicon sequences in Table S4, to assess possible aerosolized contamination from previous PCR reactions performed in our and neighbouring labs.

Reads were mapped to contiguous probe regions, flanked by up to 350 bp of upstream and downstream sequence. Reads mapping at this step were assigned to a probe and that probe's associated taxa. Non-mapped reads were competitively mapped to the genomes of the spiked organisms: SARS-CoV-2 and HCoV-NL63 coronavirus or *B. multivorans*, *K. pneumoniae*, *S. aureus*, *S. constellatus*, *S. intermedius*, *S. pneumoniae*, and *S. sanguinis*. If any read overlapped a region targeted by a probe, the read was assigned to that probe. Unassigned reads were mapped to the set of other genomes used to design the respective probe sets. SAMtools (Li et al., 2009) was used at each filtering step, and to calculate depth of coverage.

Conversion from nominal copy number to mass of viral RNA was calculated for both viruses using the high copy number shotgun baseline sample. For both viruses, the genome was broken into non-overlapping regions of the same length as the PCR amplicon generated during copy number quantification. The ratio of the depth of coverage in each region to the reference region then was used to calculate the copy number of each region across the genome. Using a conversion of 320 g/mol/nt for ssRNA, and the length of the reference region for in each virus, the mass in each genomic region was calculated and integrated across the genome to estimate the total viral RNA given the nominal copy number.

To assess potential off-target enrichment of the human background, regions of interest in the human reference were identified as any region with read counts above the 95th percentile assuming read counts at each position are Poisson distributed with a mean equal to the average read depth across the reference. Reads from enriched libraries which mapped to these regions were additionally mapped to the probes and target genomes as above to determine by which, if any, probes these reads were captured.

QUANTIFICATION AND STATISTICAL ANALYSIS

Logistic regression was used to assess the effect of enrichment. The probability of any given read in a library mapping to a target genome was used as the regressand with enrichment input concentrations as regressors. Fold enrichment for this analysis was the fold change in the odds of observing an on target read, as determined by the value of coefficient estimate. For the CoV dataset, the mass of input RNA for each virus and the interaction between were used as continuous predictors, while whether the samples were enriched was used as a categorical predictor. For the Sepsis dataset, the number of on target reads was corrected based on the proportion of on target reads observed in the blank samples. This correction was done on a by sample basis by subtracting a number of reads equal to the number of reads observed in the blanks adjusted by the ratio in library size between each sample and the blanks. With this adjusted read count, the proportion of on target reads was used as the regressand, with spike level, enrichment status, and their interaction as categorical variables. Significance of enrichment was determined using the Wald test on the regression coefficient estimate for the enrichment parameter.

Linear regression on the with the number of doublings in fold enrichment as the regressand and various parameters of the baits as regressors was used to assess the performance of baits targeting different genomic regions. For both the CoV and Sepsis datasets continuous predictors were transformed as necessary to meet the assumptions of linearity between the predictors and the response.

Two linear regressions were performed for the CoV dataset. The first considered all genomic regions, while the latter only considered genomic regions targeted by baits. When considering all regions separate coefficients were estimated for targeted and untargeted regions for the following predictors: the deviation in GC content from the mean GC content across the genome; the proportion of reads observed in a region in the shotgun baseline sample; and the number of doublings in mass of RNA for both viruses. These predictors were untransformed. All regions were assumed to be in the SARS-CoV-2 genome, with a coefficient estimating the effect of actually being from HCoV-NL63. When considering only genomic regions, the density of probes (number of probes per kilobase), and the mean divergence of probes from their target were also included as predictors.

Linear regression performed for sepsis dataset with GC content, probe divergence, probe density, the baseline levels in the shotgun as continuous regressors. Whether the sample was in a blood or water background was used as a categorical predictor. Only data from the High concentration samples and positive control were used as most probe regions had no data in the shotgun samples at lower concentrations.