Viral_Probe_Designer.py

A bioinformatic approach to designing a small but comprehensive oligonucleotide pool for viral sequence capture

1.1 Abstract

Background: Enrichment of viral sequences from complex samples using probes can very specifically improve length and depth of coverage. When the target species exhibits a high degree of diversity, however, designing a cost-effective panel of probes that can specifically enrich targeted sequences with low levels of redundancy becomes a challenge.

Results: We designed a pipeline written in python 3 that takes as input a series of reference sequences and outputs a minimal number of probes to reflect the diversity of the given sequences. We compared the performance of our tool to other published strategies for design of oligonucleotide probes in enrichment of viral sequences using a set of 212 hepatitis C virus genomes across 13 subtypes.

Conclusion: Though it is slower and has a limit to the size of datasets it can feasibly process, the pipeline presented here can propose panels of probes that are smaller and denser in their diversity than other tools of its kind. The tool is highly customizable, with built-in quality control and visual analysis of the probes and design process. The tool is publicly available at:

https://github.com/ThorhaugeDK/Viral_Probe_Designer/

1.2 Background

When conducting research on viral sequences from complex, biological samples, targeted enrichment can increase the amount of relevant information generated, lower the costs of data acquisition, and facilitate the study of low quality samples that would have otherwise failed [1]. There are two main competing

strategies in targeted enrichment: the PCR amplicon approach and hybridization capture. Both rely on the sequence homology of either PCR primers or magnetic probes to the target sequence for specific enrichment. The efficiency and comprehensiveness of a method relies on the specificity of these oligonucleotides to the target region across the established and unknown diversities that the target sequence can exhibit. While the PCR amplicon strategy remains a popular choice for target enrichment of highly diverse viruses [2,3], the method has limitations some of which can be largely circumvented by enriching with a competently designed panel of oligonucleotide probes. By increasing the length of probes beyond what is possible for primers, their level of similarity to the target sequence can be lowered without compromising their specificity [4]. This allows a probe to reliably enrich their target sequence with a degree of leniency not permitted by amplicons. PCR amplicon approaches also routinely require multiple rounds of nested PCRs and a perfect 3' match, and may require sequence data up front to inform choice of primers [2,5]. A large panel of probes can be designed against the total expected diversity within a target virus and deployed simultaneously for ultra-wide and -deep sequencing of highly variable species without prior knowledge of strain or subtype of infection. This method has not only been used for distinct highly variable viruses [4,6,7], but to characterize entire viromes from complex samples [8–10].

Many probe-panels are designed simply by slicing reference sequences into 80-120 k-mers and creating multiple sets of overlapping, "tiled" probes. For datasets of considerable size, this relatively naïve approach will inevitably result in a redundancy in probes covering equivalent, conserved regions and prohibitively expensive panels of probes due to their sheer size. Different strategies have been employed to reduce the size of probe panels ranging from trimming the input sequences down to a dense but diverse set of references [11], using consensus sequences of input references as input [12], and downsizing large panels by programmatically removing probes at a certain threshold of redundancy [8].

Many of the existing tools for probe design are purpose-built for use with eukaryotic or multiple species, or specific assays such as fluorescence in situ hybridization or restriction-site associated DNA sequencing [13–16]. Here, we introduce a new software package for designing small, efficient panels of probes for hepatitis C virus (HCV), but it could be easily adapted for other hypervariable viruses. The tool is highly flexible with a number of optional arguments that can be used to adjust degree of coverage and final probe count.

1.3 Implementation

The tool is launched using a master python script that manages the flow of data from a series of subscripts across three main steps: The cleaning and generation of initial probes, the probe design itself, and a final quality control step. These steps are visualized in Figure 5.1. In summary: the tool will propose a set of probes that cover a set of input references at a certain threshold of similarity by iterating over each genome in sequence and filling gaps of probe-coverage in the references with the corresponding sequences of the input genomes. The pipeline is written in python 3, uses commands sent to a linux shell and requires the following python 3 modules: time, datetime, os, argparse, Bio.SeqIO, random, matplotlib.pyplot, numpy, statistics, subprocess, textwrap, sys, math, and requires a working MUSCLE command line installation. Though it uses a series of scripts for its entire workflow, it is executed using a single command and outputs progress updates to the shell through which it was launched, as well as a log file.

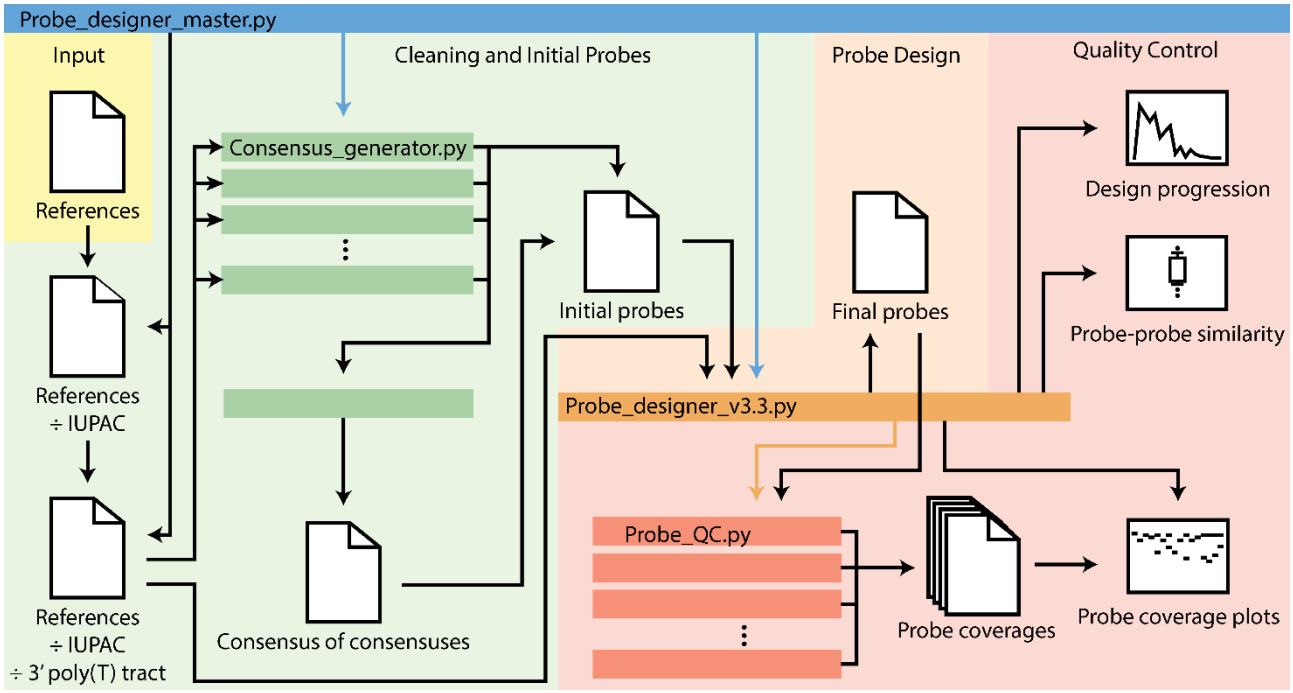


Figure Error! No text of specified style in document.**.1 Overview of the probe-designing pipeline** Horizontal bars represent the four different scripts and the order in which they are executed.

1.3.1 Input

Input is provided as a single fasta file that contains (reference) genomes of high quality whose diversity should be represented in a probe panel for enrichment. Each genome should be well curated to permit only sequences of which the researcher is most confident. The headers should have a prefix of virus genotype and subtype followed by an underscore (e.g. “>1a_EF407457” (HCV) or “>MB_K03455” (HIV)). If a subtype is unknown, it can be left out. For a cheaper, denser set of efficient probes, care should be taken to choose only sequences that can be expected to reflect the study cohort to be sequenced. Every genome of a new subtype or genotype can dramatically add to the final probe count with diminishing returns for each subsequent genome of the same type. Using an optional parameter on launch can exclude certain genotypes or subtypes as targets for probe design, but still include them for quality control. This allows a researcher for instance to design probes only against HCV4a sequences, but check the coverage of the resulting probes against other subtypes of HCV4. A list of the parameters can be found in Supplementary Figure 5.2.

1.3.2 Cleaning and Initial Probes

Before input (reference) genomes are used as targets for probe design, degenerate IUPAC code is resolved to one of its constituent bases. Additionally, the 3’ poly-(A) motif is common in both mRNA and viral genomes and thus must be trimmed away to prevent capture of low-complexity non-HCV sequences [6]. The 3’ UTR poly-(A) tract is identified and removed if one of the two following conditions are satisfied: Reading through the last 400 nucleotides of the sequence from 5’ to 3’:

1) there has been ten T’s in a row OR

2) there has been two T’s in a row and >=13 of the previous 20 nucleotides are T’s

The pipeline reports on its findings for every (reference) genome and prints it to terminal as well as to the log file for manual inspection. Below an example for two input (reference) genomes can be seen.

```
No suspicion of polyT region for reference: 1a_HQ850279
5'...AAATCAATAGGGTGGCCGCATGCCTCAGAAAACTTGGGGTCCCGCCCTTGCGAGCTT
GGAGACACCGGGCCCGAAGCGTCCGCGCTAAGCTCCTGTCCAGAGGAGGCAGGGCTGCCATA
TGTGGCAAGTACCTCTTCAATTGGGCAGTGAGAACAAAGCTCAAACTCACTCCAATAGCGGC
```

```
CGCTGGCCAGCTGGACTTGTCCGGCTGGTTCACGGCTGGCTACAGCGGGGGAGACATTTATC
ACAGCGTGTCTCGTGCCCGGCCCCGCTGGTTTTGGTTTTGCCTACTCCTGCTTGCTG '3

Suspicion of polyT region found and excluded for reference:
1a_EF407419
5'...GGCTACAGCGGGGGAGACATTTATCACAGCGTGTCTCATGCCCGGCCCCGCTGGTTC
TGGTTTTGCCTACTCCTGCTCGCTGCAGGGGTAGGCATCTACCTCCTCCCCAACCGATGAAG
GTTGGGGTAAACACTCCGGCCTCTTAGGCCA ▼ TTTCCTGTTTTTTTTTTTTTTTTTT
'3
```

Once cleaning is finalized, an initial set of probes is generated. Without this set, the first probes would simply be 120 nucleotide end-to-end segments of the genome it first encounters since no pre-existing probe exists to cover its genome anywhere. That set might not be the most efficient in covering the given diversity within an entire genotype, subtype, or all strains collectively and might lead to an unnecessarily high number of probes. Instead, a consensus of input (reference) genomes is used to produce the initial probes. The script can be launched with several options to customize what is included for the generation of this consensus sequence, and if multiple sets of consensus sequences for each genotype should be employed. Whichever options are chosen, the script aligns the relevant sequences using MUSCLE [17], generates a consensus sequence, and produces a set of 120 nucleotide end-to-end segments as initial probes.

1.3.3 Probe design

In an iterative fashion, each (reference) genome in the order they are listed in the input fasta file, is scanned in 120 nucleotide windows with a step size of one from 5’ to 3’ and has each window compared to the similarity of every probe designed thus far. In this manner, regions where no existing probe has a similarity above a given threshold are identified. Wherever these gaps exceed half the length of a probe, the corresponding sequence in the (reference) genome is used as a template for a number of new probes sufficient to bridge the gap (see Figure 5.2). The script then continues to the next (reference) genome in the input fasta file and keeps performing this series of actions, producing an increasing number of probes until every input (reference) genome has been analyzed and probes designed against it if necessary. This

panel of probes is written to a fasta file with the header of each containing the ID of the (reference) genome against which it was designed (e.g. “>Probe_17_1b_D90208”).

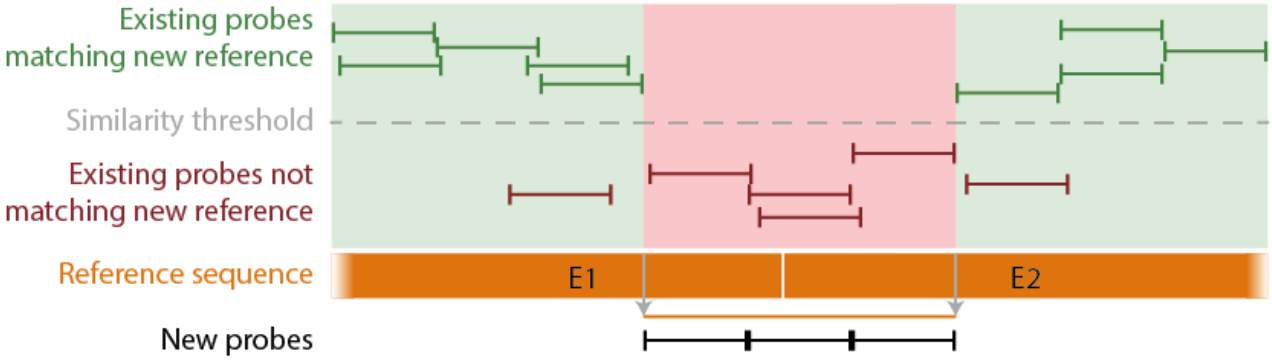


Figure Error! No text of specified style in document.**.2 Close-up visualization of probe design**
The figure is a simplified example showing where probes designed this far into the process map to a new input sequence. The red box shows a region where no probe matched the reference sequence above the sequence similarity threshold. This region is used as a template to design new probes that are added to the pool for the next input sequence.

1.3.4 Quality Control

After designing the panel of probes, the script will proceed to a quality control step in which three main analyses and visualizations are produced. First, a graph showing both the number of new probes generated per input (reference) genome as well as the cumulative total probes is produced and can assist the researcher in evaluating which strains might not be worthwhile to include in probe design (see Supplementary figure 5.3). Second, to evaluate the efficiency of probe design and as an extra measure that probe design performs as expected, the best matching alignment for each probe to every other probe is recorded and graphed as a boxplot (see Supplementary figure 5.4). Provided that only a single consensus sequence was used as an initial probe set, no probe should have a similarity to any other probe above the sequence similarity threshold. Third, each reference is iterated through once more, this time finding the best matching position for each probe in the final probe pool. A graph is produced for every input sequence showing the position and similarity for every probe with regions of non-coverage highlighted (see Supplementary figure 5.5). This is especially useful when evaluating the affinity of the probes towards genotypes and subtypes that were not included as targets for probe design but still included in the input references.

1.4 Results and Discussion

1.4.1 Sample dataset and evaluation

We compiled a set 165 full HCV genomes from our own cohort and added 47 sequences of the represented HCV subtypes from the Los Alamos and ICTV databases for a final dataset containing 212 sequences across 13 subtypes (1a: 69, 1b: 66, 3a: 21, 4d: 17, 2a: 6, 4a: 6, 4v: 6, 4k: 6, 4r: 4, 2c: 4, 4q: 3, 4c: 2, 2i: 2). The initial 120-nucleotide probes were generated from the consensus of all (reference) genomes with new probes generated at a sequence similarity threshold of 80% resulting in a final panel of 768 probes (arguments -c, -cs). Generating these probes took nine hours and thirty-three minutes with a further ten hours and forty-one minutes for automated evaluation. Two example probe-reference alignments for this run can be seen in Supplementary figure 5.5.

1.4.2 Comparison to other tools

The default protocol for many commercial providers of oligonucleotide probes like IDT, is to use 120-mer end-to-end slices with a tiling factor of 1X or 2X. For the dataset used here, this would yield approximately 17,000 or 34,000 probes resulting in a large level of redundancy, especially in conserved areas and in excessively expensive sequencing assays. Many methods to consolidate a more cost-effective panel of probes from the total potential probes generated with more naïve approaches have been suggested, but are often developed more specifically for data incompatible with research on hypervariable viral taxa. One notable exception is ProbeTools [18] which subdivides the input references into k-mers before clustering them based on a given similarity threshold using VSEARCH cluster_fast [19]. The centroid sequence of each cluster is then ranked by the size of the cluster they represent and the candidates that can account for the most k-mers are accepted as probes, a strategy that can result in a bias towards overrepresented sequences. ProbeTools then iterates through this process until 90% of the input references have more than 90% of their lengths covered by at least one probe. On our dataset with the closest equivalent parameters, ProbeTools produced 3,100 probes (see Table 5.1). We tested CATCH [20], another tool, in a similar fashion. CATCH works by generating an excessively large candidate probe panel based on input sequences and then reducing the panel by solving the set cover problem. In their original

publication the authors generated approximately 14,000 probes for 300 HCV genomes, but on our dataset with the closest equivalent parameters a more manageable 1,866 probes was generated. While there is a marginal increase in 0X coverage for our tool presented here (0.87% vs. 0.00% and 0.02%), the other tools generated 2.4 and 4.0 times more probes with an associated increase in potentially redundant depths of coverage. Maps of probe-reference alignments for a single sample of subtype HCV2a using the different tools are shown in Supplementary figure 5.7.

Since tools that attempt to collapse either the diversity of a large number of input sequences before generating probes or collapse the diversity of a large number of probes themselves can be parallelized more readily, these tools can reach faster speeds than the methodology applied here. By accounting for the diversity not before or after generating probes, but during, the tool developed here is never naïve to the diversity already accounted for and is likely to generate a denser set of diverse probes that still meet the criteria provided. This methodical approach, however, will increase in runtime exponentially with more distinct input sequences and depending on the virus diversity and genome length is likely computationally unfeasible for datasets larger than approximately 1,000 sequences.

	No. probes	Median 0X positions	Median >=10X positions	Median >=20X positions	Run time
Our tool	768	84 (0.87%)	125 (1.30%)	0 (0.00%)	9 hours, 33 minutes
CATCH	1,866	0 (0.00%)	1,262 (13.10%)	251 (2.60%)	14 minutes
ProbeTools	3,100	2 (0.02%)	7,576 (78.55%)	2,078 (21.55%)	11 minutes

Table Error! No text of specified style in document.**.1 Comparison of different probe designing tools for hypervariable viral taxa**
The percentages listed are relative to the median genome size of 9,645. All tools were applied to the same dataset of 212 full-genome HCV sequences across 13 subtypes. All tools were asked to create 120 nt probes using default parameters except for the threshold of similarity. Our tool uses an 80% sequence similarity threshold for generating new probes. For ProbeTools, the sequence identity threshold for k-mer clustering and probe-target alignments were lowered from the default 90% to 80%. The mismatches parameter in CATCH was increased to 24 for an 80% sequence similarity comparable to the other tools.

In benchmarking and optimization of the tool, we found that lowering the accepted threshold of probe-to-sequence similarity reduces the number of final probes in the pool considerably. For the dataset above our tool produced a panel of 455, 768 and 1,200 probes for a sequence similarity threshold of 75%, 80%, and 85%, respectively. Ultimately, the sequence similarity threshold should be established based on the expected diversity of the viral taxa, available data, the research question, and financial limitations. Despite

many algorithms such as CATCH, ProbeTools, and solutions that predate them often using stricter criteria for probe design [12,21], we settled on 80% as it has been suggested specifically in the context of HCV as an experimentally determined threshold for effective enrichment [6].

1.5 Conclusion

We present a new tool to produce a panel of probes for enrichment of highly variable viral taxa. The pipeline is flexible, easy to use, performs a quality assessment step upon completion, and generates figures to help evaluation of the proposed probe panel. The tool is publicly available at:

https://github.com/ThorhaugeDK/Viral_Probe_Designer/

1.6 Supplementary Materials

1.6.1 Launch parameters

The probe design tool supports a number of options to tailor the final probe count and affinities. Table 5.2 gives an overview of the parameters and their function.

Supplementary table Error! No text of specified style in document.**.2 List of launch options for probe design algorithm**

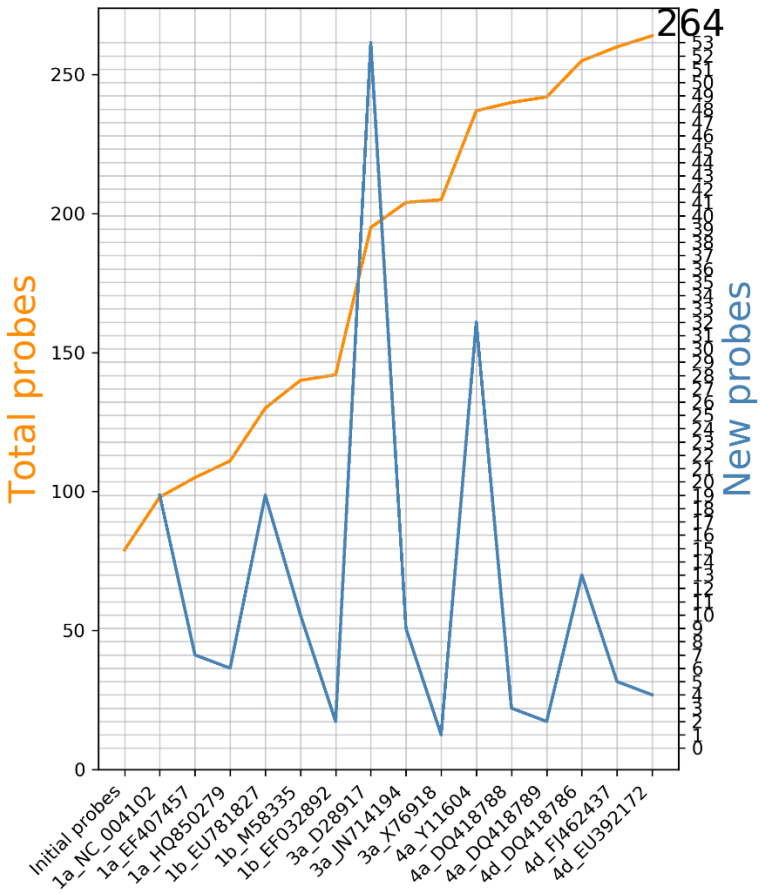
Cmd	Alternative	Function
	references_path	The full path to the references. This is mandatory.
-s	--similarity [float, default=0.80]	Insert a value between 0 and 1 of the similarity acceptance threshold between probe and reference
-t	--threads [int, default=1]	Select the number of threads to be used. Speeds up QC analysis dramatically
-c	--consensus	The consensus sequence of all samples is used to generate initial probes
-cc	--consensus_of_consensus	The consensus of all subtype consensuses is used to generate initial probes
-cg	--consensus_of_genotypes	Each consensus of a genotype is used to generate initial probes
-cs	--consensus_of_subtypes	Add consensus sequences for subtypes to probe design
-st	--subtype threshold [int, default=0]	Insert a value for a minimum number of sequences of a given subtype it takes to justify generating and including a subtype consensus for probe design
-ot	--only_typed	Only sequences with a known geno- and subtype will be used in probe design
-eg	--exclude_genotype [list]	list of space-separated genotypes to be excluded for the purposes of probe design, but still tested against designed probes. If -cg is on a consensus for these is still used to generate initial probes.

As an example, the command:

```
python Probe_designer_master.py input_references.fasta -c -s 0.8 -t 10 -eg 5 6
7 8 -cs -ot -st 10
```

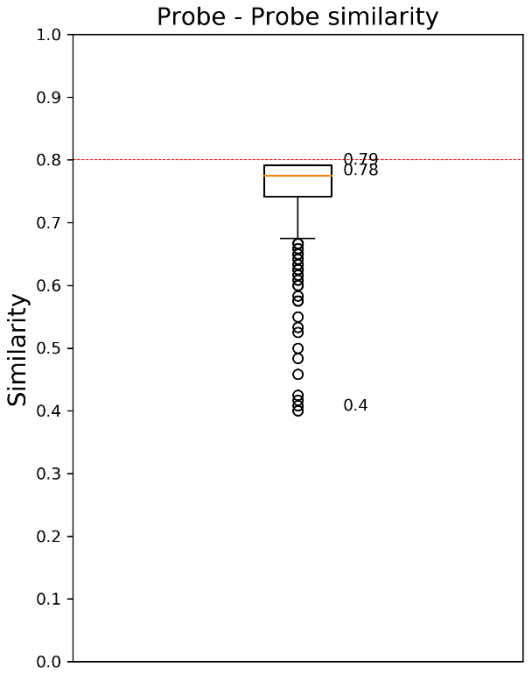
will generate generate an initial set of probes by slicing the consensus sequence made from all input references into 120 nt segments (`-c`). If any 120 nucleotide segment in a query genome is less than 80%

similar to the probe that it is the most similar to, a new probe will be designed using this 120 nucleotide stretch in the query genome (`-s 0.8`). Before iterating through each genome in the reference file for probe design, it will first iterate through the consensus sequences for each subtype found in the reference file, except for the subtypes of which there are less than 10 genomes (`-cs -st 10`). No probes should be designed against genomes of genotype 5, 6, 7, and 8, and only genomes with a known subtype should have probes designed against them (`-eg 5 6 7 8 -ot`). When performing the quality control, the process is sped up by using ten threads (`-t 10`).



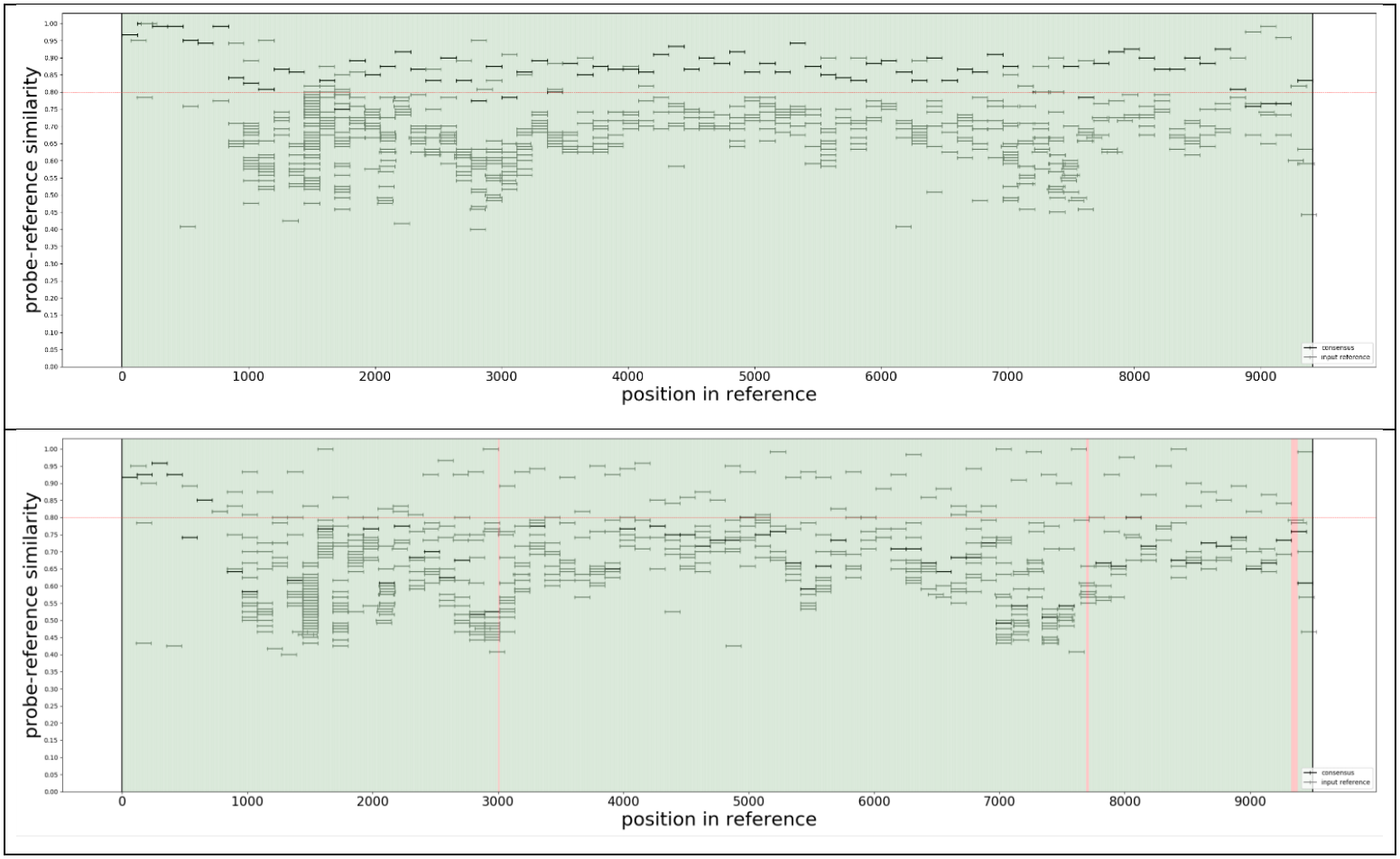
Supplementary figure Error! No text of specified style in document.**.3 Figure from quality control of probe design - Overview of design process**

Figure shows the individual number of new probes generated per reference (blue, right axis) and the running total number of probes generated so far in the process (orange, left axis). The input references are shown on the first axis: three each of HCV1a, 1b, 3a, 4a, and 4d.



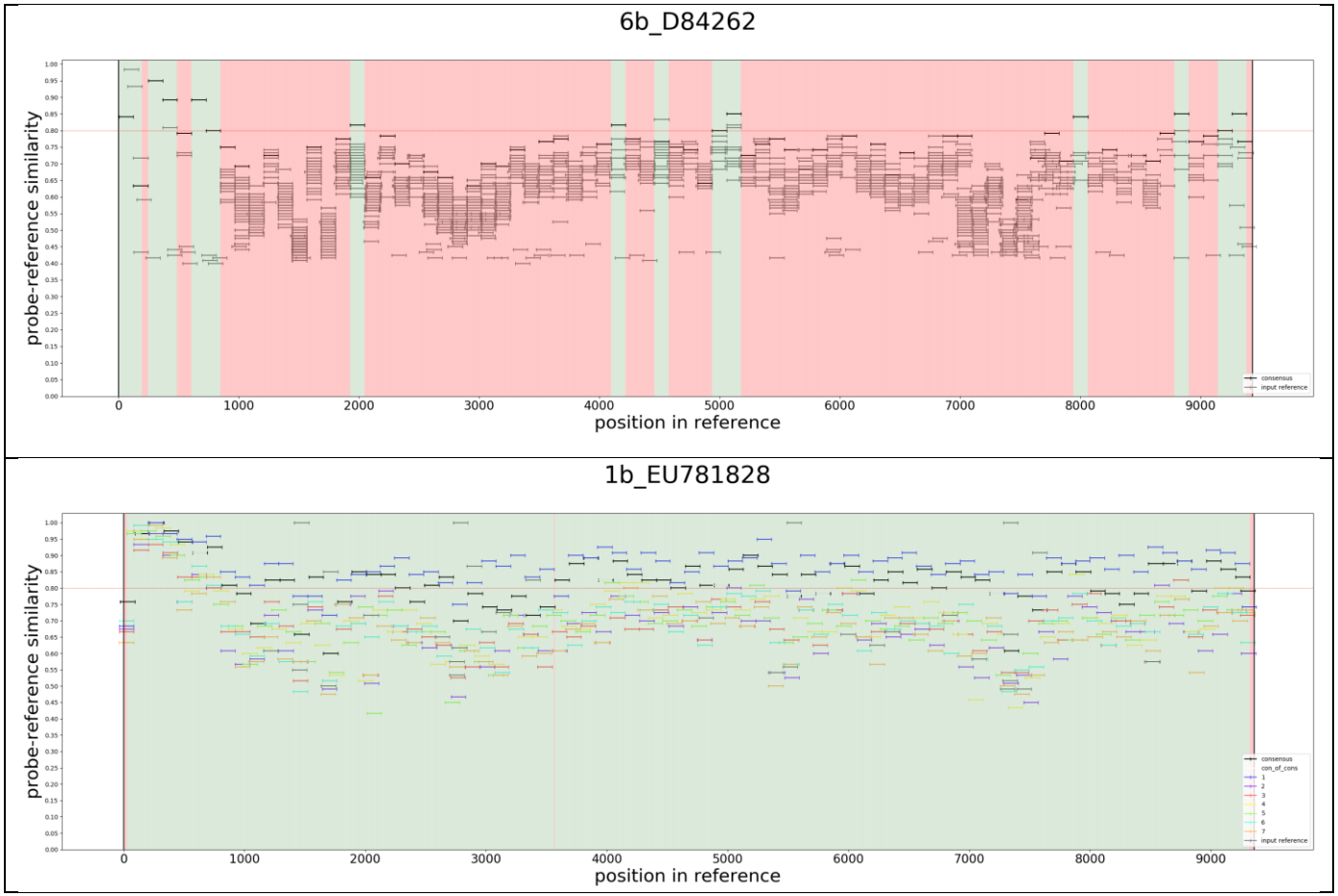
Supplementary figure Error! No text of specified style in document.**.4: Boxplot of probe-to-probe similarities**

After all probes have been designed, the similarities of each probe to every other probe is calculated and visualized. The second and third quartile of these similarities are typically very close to the similarity threshold (here 80%). The lowest similarities are probes designed against a sequence of which there is no direct equivalent region due to slight discrepancies in sequence length or the presence of indels.



Supplementary figure Error! No text of specified style in document.**.5: Figure from quality control of probe design - maps of probe-reference alignments**

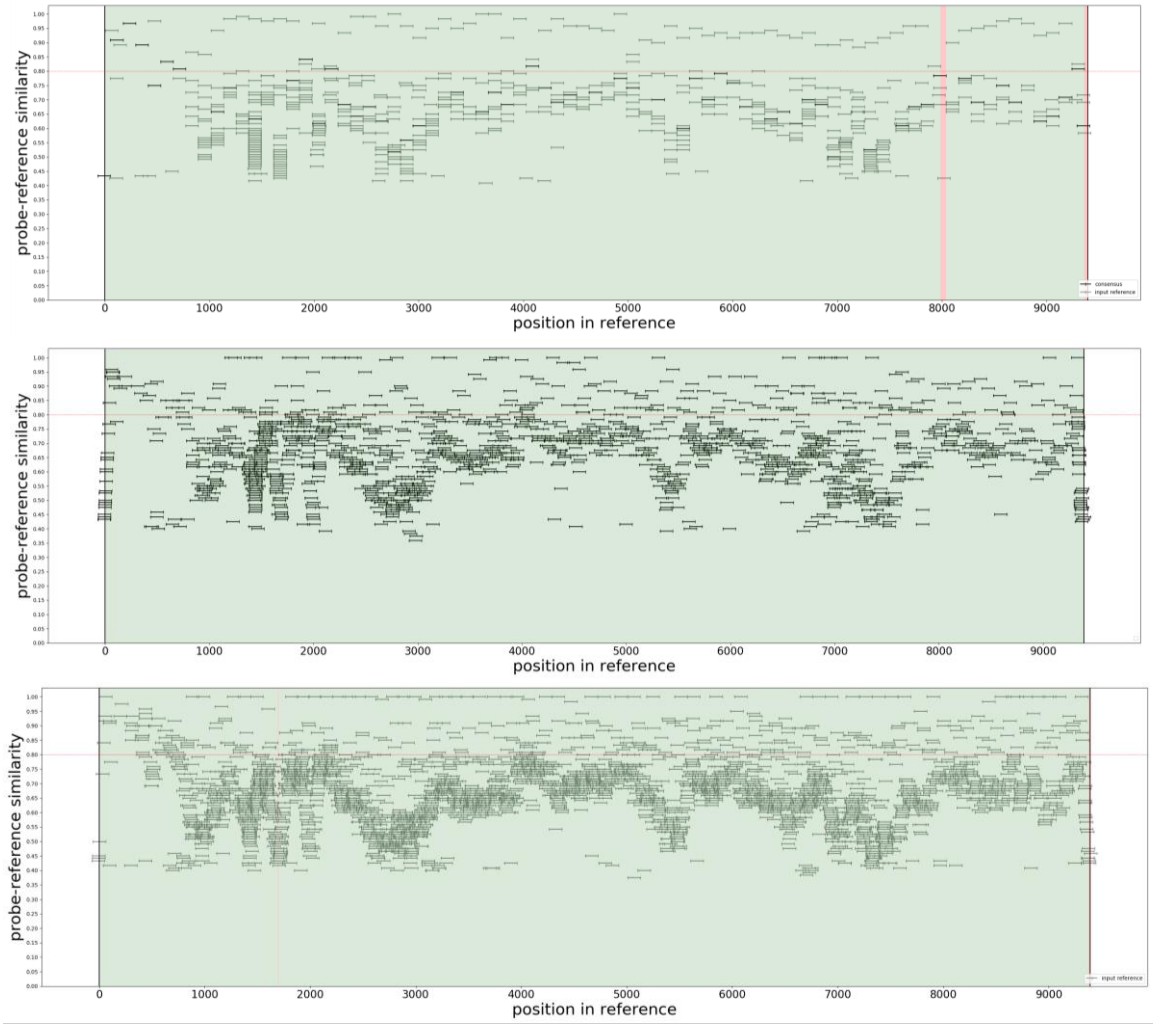
Figures are examples of probe coverage visualizations generated automatically at the end of the pipeline to help evaluate the performance of the probe design process. Figures show the best possible alignment for each probe generated during the probe designing process against two distinct input references. Darker probes show the initial probes made from the consensus sequence of all sequences. Lighter probes were generated when looping over references.



Supplementary figure Error! No text of specified style in document.**.6: Figure from quality control of probe design - probe-reference alignments with different launch options**

Top: A HCV6b reference genome with probes designed only against HCV1, 2, 3, and 4.

Bottom: A HCV1b genome where consensuses of genotypes 1-7 were used for initial probe design. The colors of probes correspond to the originating genotype of different sets of initial probes.



Supplementary Figure Error! No text of specified style in document.**.7: Comparison of probe-reference alignments for a single HCV2a reference (HQ639944) using different tools**

Top: Our tool (768 probes), middle: CATCH (1,866 probes), bottom: ProbeTools (3,100 probes)

1.7 Bibliography

[1] L. Mamanova, A.J. Coffey, C.E. Scott, I. Kozarewa, E.H. Turner, A. Kumar, E. Howard, J. Shendure, D.J. Turner, Target-enrichment strategies for next-generation sequencing, Nat. Methods 2010 72. 7 (2010) 111–118. https://doi.org/10.1038/nmeth.1419.

[2] R.A. Bull, A.A. Eltahla, C. Rodrigo, S.M. Koekkoek, M. Walker, M.R. Pirozyan, B. Betz-Stablein, A. Toepfer, M. Laird, S. Oh, C. Heiner, L. Maher, J. Schinkel, A.R. Lloyd, F. Luciani, A method for near full-length amplification and sequencing for six hepatitis C virus genotypes, BMC Genomics. 17 (2016) 1–10. https://doi.org/10.1186/S12864-016-2575-8/FIGURES/4.

[3] G.H. Kijak, E. Sanders-Buell, P. Pham, E.A. Harbolick, C. Oropeza, A.M. O'Sullivan, M. Bose, C.G. Beckett, M. Milazzo, M.L. Robb, S.A. Peel, P.T. Scott, N.L. Michael, A.W. Armstrong, J.H. Kim, D.M. Brett-Major, S. Tovanabutra, Next-generation sequencing of HIV-1 single genome amplicons, Biomol. Detect. Quantif. 17 (2019). https://doi.org/10.1016/J.BDQ.2019.01.002.

[4] J.R. Brown, S. Roy, C. Ruis, E. Yara Romero, D. Shah, R. Williams, J. Breuer, Norovirus whole-genome sequencing by SureSelect target enrichment: A robust and sensitive method, J. Clin. Microbiol. 54 (2016) 2530–2537. https://doi.org/10.1128/JCM.01052-16.

[5] C.M. Rousseau, B.A. Birditt, A.R. McKay, J.N. Stoddard, T.C. Lee, S. McLaughlin, S.W. Moore, N. Shindo, G.H. Learn, B.T. Korber, C. Brander, P.J.R. Goulder, P. Kiepiela, B.D. Walker, J.I. Mullins, Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes, J. Virol. Methods. 136 (2006) 118–125. https://doi.org/10.1016/J.JVIROMET.2006.04.009.

[6] D. Bonsall, M.A. Ansari, C. Ip, A. Trebes, A. Brown, P. Klenerman, D. Buck, STOP-HCV Consortium, P. Piazza, E. Barnes, R. Bowden, ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens., F1000Research. 4 (2015) 1062. https://doi.org/10.12688/f1000research.7111.1.

[7] J. Yamaguchi, A. Olivo, O. Laeyendecker, K. Forberg, N. Ndembi, D. Mbanya, L. Kaptue, T.C. Quinn, G.A. Cloherty, M.A. Rodgers, M.G. Berg, Universal Target Capture of HIV Sequences

From NGS Libraries, Front. Microbiol. 9 (2018). https://doi.org/10.3389/FMICB.2018.02150.

[8] T.N. Wylie, K.M. Wylie, B.N. Herter, G.A. Storch, Enhanced virome sequencing using targeted sequence capture, Genome Res. 25 (2015) 1910–1920. https://doi.org/10.1101/GR.191049.115.

[9] A.C. Paskey, K.G. Frey, G. Schroth, S. Gross, T. Hamilton, K.A. Bishop-Lilly, Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples, BMC Genomics. 20 (2019). https://doi.org/10.1186/S12864-019-5543-2.

[10] C. Wylezich, S. Calvelage, K. Schlottau, U. Ziegler, A. Pohlmann, D. Höper, M. Beer, Next-generation diagnostics: virus capture facilitates a sensitive viral diagnosis for epizootic and zoonotic pathogens including SARS-CoV-2, Microbiome. 9 (2021). https://doi.org/10.1186/S40168-020-00973-Z.

[11] Y. Xiao, J.M. Nolting, Z.M. Sheng, T. Bristol, L. Qi, A.S. Bowman, J.K. Taubenberger, Design and validation of a universal influenza virus enrichment probe set and its utility in deep sequence analysis of primary cloacal swab surveillance samples of wild birds, Virology. 524 (2018) 182. https://doi.org/10.1016/J.VIROL.2018.08.021.

[12] B.M. O'Flaherty, Y. Li, Y. Tao, C.R. Paden, K. Queen, J. Zhang, D.L. Dinwiddie, S.M. Gross, G.P. Schroth, S. Tong, Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing, Genome Res. 28 (2018) 869–877. https://doi.org/10.1101/GR.226316.117/-/DC1.

[13] T.K. Chafin, M.R. Douglas, M.E. Douglas, MrBait: universal identification and design of targeted-enrichment capture probes, Bioinformatics. 34 (2018) 4293–4296. https://doi.org/10.1093/BIOINFORMATICS/BTY548.

[14] B.J. Beliveau, J.Y. Kishi, G. Nir, H.M. Sasaki, S.K. Saka, S.C. Nguyen, C. ting Wu, P. Yin, OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E2183–E2192. https://doi.org/10.1073/PNAS.1714530115/SUPPL_FILE/PNAS.1714530115.SD02.PDF.

[15] C. Mayer, M. Sann, A. Donath, M. Meixner, L. Podsiadlowski, R.S. Peters, M. Petersen, K. Meusemann, K. Liere, J.W. Wagele, B. Misof, C. Bleidorn, M. Ohl, O. Niehuis, BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design, Mol. Biol. Evol. 33 (2016) 1875–1886. https://doi.org/10.1093/MOLBEV/MSW056.

[16] S.K. Kushwaha, L. Manoharan, T. Meerupati, K. Hedlund, D. Ahrén, MetCap: A bioinformatics probe design pipeline for large-scale targeted metagenomics, BMC Bioinformatics. 16 (2015) 1–11. https://doi.org/10.1186/S12859-015-0501-8/FIGURES/7.

[17] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797. https://doi.org/10.1093/nar/gkh340.

[18] K.S. Kuchinski, J. Duan, C. Himsworth, W. Hsiao, N.A. Prystajecky, ProbeTools: designing hybridization probes for targeted genomic sequencing of diverse and hypervariable viral taxa, BMC Genomics. 23 (2022) 1–14. https://doi.org/10.1186/S12864-022-08790-4/TABLES/2.

[19] T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics, PeerJ. 4 (2016). https://doi.org/10.7717/PEERJ.2584.

[20] H.C. Metsky, K.J. Siddle, A. Gladden-Young, J. Qu, D.K. Yang, P. Brehio, A. Goldfarb, A. Piantadosi, S. Wohl, A. Carter, A.E. Lin, K.G. Barnes, D.C. Tully, B. Corleis, S. Hennigan, G. Barbosa-Lima, Y.R. Vieira, L.M. Paul, A.L. Tan, K.F. Garcia, L.A. Parham, I. Odia, P. Eromon, O.A. Folarin, A. Goba, E. Simon-Lorière, L. Hensley, A. Balmaseda, E. Harris, D.S. Kwon, T.M. Allen, J.A. Runstadler, S. Smole, F.A. Bozza, T.M.L. Souza, S. Isern, S.F. Michael, I. Lorenzana, L. Gehrke, I. Bosch, G. Ebel, D.S. Grant, C.T. Happi, D.J. Park, A. Gnirke, P.C. Sabeti, C.B. Matranga, Capturing sequence diversity in metagenomes with comprehensive and scalable probe design, Nat. Biotechnol. 2019 372. 37 (2019) 160–168. https://doi.org/10.1038/s41587-018-0006-x.

[21] T. Briese, A. Kapoor, N. Mishra, K. Jain, A. Kumar, O.J. Jabado, W. Ian Lipkina, Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis, MBio. 6 (2015). https://doi.org/10.1128/MBIO.01491-15.