Enhanced virome sequencing through solution-based capture enrichment

Todd N Wylie, Kristine M Wylie, Brandi N Herter, et al.

Genome Res. published online September 22, 2015
Access the most recent version at doi:10.1101/gr.191049.115

Enhanced Virome Sequencing Using Targeted Sequence Capture

Todd N. Wylie[1,2,*], Kristine M. Wylie[1,2,*], Brandi N. Herter[1], and Gregory A. Storch[1]

The Department of Pediatrics[1], Washington University School of Medicine, Campus Box 8208, 660 S. Euclid Avenue, St Louis, MO 63110, USA

McDonnell Genome Institute[2], Washington University School of Medicine, 4444 Forest Park Avenue, St Louis 63108, MO, USA

[*]These authors contributed equally to this work.

Corresponding Author

Todd N. Wylie

wylie_t@kids.wustl.edu

Campus Box 8208

Washington University School of Medicine

660 S. Euclid Avenue

St. Louis, MO 63110

314-747-4069

Running Title:

Enhanced virome sequencing using sequence capture

Keywords: targeted sequence capture; infectious disease; virus; virome; microbiome; metagenomics; genomics; next-generation sequencing; deep sequencing; ViroCap

ABSTRACT

Metagenomic shotgun sequencing (MSS) is an important tool for characterizing viral populations. It is culture-independent, requires no *a priori* knowledge of the viruses in the sample, and may provide useful genomic information. However, MSS can lack sensitivity and may yield insufficient data for detailed analysis. We have created a targeted sequence capture panel, ViroCap, designed to enrich nucleic acid from DNA and RNA viruses from 34 families that infect vertebrate hosts. A computational approach condensed nearly 1 billion base pair (bp) of viral reference sequence into less than 200 million bp of unique, representative sequence suitable for targeted sequence capture. We compared the effectiveness of detecting viruses in standard MSS versus MSS following targeted sequence capture. First, we analyzed two sets of samples, one derived from samples submitted to a diagnostic virology laboratory and one derived from samples collected in a study of fever in children. We detected 14 and 18 viruses in the two sets, comprising 19 genera from 10 families, with dramatic enhancement of genome representation following capture enrichment. The median fold-increases in percent viral reads post-capture were 674 and 296. Median breadth of coverage increased from 2.1% to 83.2% post-capture in the first set and from 2.0% to 75.6% in the second set. Next, we analyzed samples containing a set of diverse anellovirus sequences and demonstrated that ViroCap could be used to detect viral sequences with up to 58% variation from the references used to select capture probes. ViroCap substantially enhances MSS for a comprehensive set of viruses and has utility for research and clinical applications.

INTRODUCTION

High-throughput, massively parallel nucleotide sequence analysis has made in-depth studies of the human microbiome feasible. Thus far, most microbiome studies have focused on bacteria (Arumugam et al. 2011; Human Microbiome Project Consortium 2012; Turnbaugh et al. 2009; Gajer et al. 2012), although some include fungi (Cleland et al. 2014; Paulino et al. 2006; Findley et al. 2013; Willger et al. 2014) and viruses (Wylie et al. 2014; 2012; De Vlaminck et al. 2013; Minot et al. 2011; Reyes et al. 2010). Viruses are particularly understudied, in part due to the challenges of assessing their presence in clinical samples. Viruses as a group have highly variable genomes, with no gene shared among all viruses that can be surveyed by an amplicon-based sequencing strategy. Therefore, studies of viruses based on nucleotide sequencing require a metagenomic approach. Metagenomic shotgun sequencing (MSS) is a relatively unbiased, culture-independent method in which nucleic acid extracted from a sample is sequenced. Sequence reads are classified based on similarity to reference genomes. This approach allows comprehensive study of the viral component of the microbiome (the virome) and has led to the discovery of novel viruses (reviewed in (Chiu 2013)) and the characterization of viruses present in healthy and sick people (Wylie et al. 2012; 2014; Reyes et al. 2010; Holtz et al. 2014; Oh et al. 2014; Lysholm et al. 2012; Minot et al. 2011; Young et al. 2014). When adequate numbers of sequence reads are generated, viruses can be characterized with regard to taxonomy and the presence of genes associated with virulence and resistance to antiviral drugs.

A limitation of MSS as employed to date for virus detection is that the amount and proportion of viral nucleic acid in samples from humans may be very low, and in these cases few viral sequences are generated. In our experience using MSS, we have detected fewer than 10 viral sequences per 25 million sequence reads generated for a virus that was detected in a sample by a molecular assay (Wylie et al. 2012). In other instances we have failed to detect viruses

known to be present based on molecular assays (Wylie et al. 2012). These difficulties may reflect the small genome size of some viruses and/or low levels of virus in the sample. This can be a particular problem for studies of the virome of healthy, asymptomatic individuals (Wylie et al. 2014; 2012), in whom virus levels may be low. In efforts to increase the sequence yield, purification or enrichment procedures have been employed, including low-speed centrifugation and/or filtration to remove bacterial and host cells, sample treatment with nucleases to digest nucleic acid not protected within virions (Allander et al. 2001), or concentration of viral particles by high-speed gradient centrifugation (reviewed in (Duhaime and Sullivan 2012)). Each of these procedures may bias against detection of some viruses (Young et al. 2014; Breitbart and Rohwer 2005).

An alternative method for enrichment of viral sequences in a metagenomic sample prior to sequencing is targeted sequence capture, a well-established approach for targeted enrichment of specific nucleic acids. Targeted sequence capture has been used extensively to assess the human exome as well as specific gene targets (Lovett et al. 1991; Okou et al. 2007; Hodges et al. 2007; Albert et al. 2007). Sequence capture has also been applied to the study of specific viruses (Duncavage et al. 2011; Depledge et al. 2011; Koehler JW et al. 2014). Our aim was to develop a comprehensive viral targeted sequence capture panel that could be used to (1) assess all viruses known to infect vertebrate cells and (2) detect divergent viruses. To this end, we created ViroCap, a targeted sequence capture panel that enhances the detection of a comprehensive set of viruses with vertebrate hosts. Here we describe the first application of ViroCap to enrich a broad range of viruses from human clinical samples.

RESULTS

ViroCap includes targets from 34 viral families, comprising 190 annotated viral genera and 337 species (**Fig. 1**). Included viruses represent all DNA and RNA viruses with sequenced genomes

105 from vertebrate hosts, except human endogenous retroviruses, which were excluded due to
106 their prevalence within the human genome. Nearly 1 billion base pairs (bp) of viral genome
107 sequences were condensed into less than 200 million bp of targets (**Supplemental Table S1**)
108 using *k*-mer and clustering analyses to define a unique set of reference sequences, as
109 described in Methods.
110
111 **Analysis of Clinical and Research Samples with ViroCap**
112 We evaluated the effectiveness of detecting DNA and RNA viruses in MSS data compared with
113 ViroCap targeted sequence capture data in two sets of human samples. In experiment 1, the
114 sample set consisted of clinical samples that had been found to be positive by molecular tests in
115 the Diagnostic Virology Laboratory at St. Louis Children's Hospital. Nucleic acid extracts
116 available in the Virology Laboratory were pooled, and a sequencing library was prepared from
117 this pooled nucleic acid (see Methods). In experiment 2, 8 patient samples from a research
118 study of young children with fever (Colvin et al. 2012; Wylie et al. 2012) were selected for use in
119 the present study because each had been found to be positive for one or more viruses when
120 tested by batteries of PCR assays used in that study. Individual sequencing libraries were
121 prepared from each of the 8 samples as described in Methods and pooled for sequencing.
122 Experiments 1 and 2 were analyzed in separate sequencing runs. In each experiment,
123 sequencing libraries were divided and the same library was sequenced without targeted
124 sequence capture (pre-capture) and following targeted sequence capture using ViroCap (post-
125 capture).
126
127 In experiment 1, we detected 10 viruses in the pre-capture MSS data (**Table 1**). After targeted
128 sequence capture using the same sequencing library, we detected the same 10 viruses plus 4
129 additional viruses. Targeted sequence capture resulted in dramatic improvements in all
130 sequence coverage metrics (**Table 1 and Supplemental Table S2**), including number and

percent viral reads, breadth and depth of coverage, and coverage gaps. In experiment 1, the median increase in percent viral reads was 674 (range >13 to 9335), and the median breadth of coverage increased from 2.1% (range 0 to 89.8%) to 83.2% (range 0.8 to 100%). Illustrative examples are shown in **Figure 2 A-D**.

In experiment 2, 11 viruses were detected in the pre-capture MSS data (**Table 2**). After targeted sequence capture with ViroCap using the same sequencing libraries, we detected those 11 viruses plus 7 additional viruses. Thus, in the two experiments together, the number of viruses detected went from 21 to 32, a 52% increase. All of the viruses detected in both experiments were confirmed by PCR assays except for a torque teno virus in the clinical pool, which was not evaluated by PCR (**Table 1, Supplemental Tables S8 and 9**). Viruses detected encompassed 19 genera from 10 families (**Supplemental Figure S1**)**.** In experiment 2, we again found that targeted sequence capture resulted in dramatic improvements in sequencing parameters. In experiment 2, the median fold increase in percent viral reads was 296 (range >56 to 2,722), and the median breadth of coverage increased from 2.0% (range 0 to 99.9%) to 75.6% (range 13.5 to 100%). Illustrative examples are shown in **Figure 2 E-H**.

Using targeted sequence capture, >80% breadth of coverage of the viral genomes was obtained for 16 of 32 viruses, including diverse DNA and RNA genomes of sizes ranging from 5 to 161 kilobases (**Tables 1 and 2 and Supplemental Tables S2 and S3)**. Greater than 90% breadth of coverage was obtained for 12 of 32 viruses, and 8 viruses had 100% coverage. Pre-capture, the median gap size in genome coverage was 1704 bp (range 4 to 152,261 bp) and post-capture the median gap size was 82 bp (range 0 to 13,734) (**Supplemental Tables S2 and S3**). High genome representation was obtained for multiple viruses in the same capture reaction, as experiments 1 and 2 were each single, independent capture reactions encompassing multiple samples (see Methods).

157

158 **Targeted sequence capture identifies divergent viral sequences**

159 To determine whether or not divergent sequences could be identified using targeted sequence
160 capture, we tested ViroCap on samples containing anelloviruses, a highly divergent group of
161 ssDNA viruses which have a common genome structure, but may have up to 30 to 50%
162 nucleotide sequence diversity among separate species (Ninomiya et al. 2007; de Villiers et al.
163 2011). We selected anellovirus-positive samples that we had previously characterized using
164 multi-strand displacement amplification followed by high-throughput sequencing. After
165 assembling the pre-capture sequences to generate contiguous sequences (contigs), we
166 identified anellovirus contigs greater than 1 kilobase in length. The contigs had varying degrees
167 of similarity to the reference genomes used in the ViroCap panel based on BLAST alignments,
168 ranging from 58% to 98% nucleotide sequence identity for the top high-scoring segment pair
169 (HSP) alignment (**Fig. 3A**, **Supplemental Table S4**). All of the contigs assembled using the pre-
170 capture sequence data were also detected post-capture. The contig with 58% identity to the
171 reference database was missing 13% of its length post-capture (**Fig. 3A**). The contig with the
172 next lowest percent identity to the reference database (62%) was fully sequenced (i.e. 100%
173 breadth-of-coverage) (**Figs. 3A and B**). Figure 3B illustrates the nucleotide sequence
174 matches/mismatches between the contig and the most similar reference genome in the
175 sequences used for the ViroCap design. These results demonstrate that targeted sequence
176 capture using the ViroCap panel allows us to identify variant virus sequences having as low as
177 58% nucleotide sequence identity.

178

179 **Specificity of targeted sequence capture**

180 In order to determine whether ViroCap systematically enriched off-target sequences, we
181 compared the filtering and classification statistics of the non-viral sequences in the pre-capture
182 MSS and targeted sequence capture data (**Supplemental Table S5, Supplemental Figures**

S2 and S3). If our probes were specific, we would not observe any systematic enrichment of specific human chromosomes or bacterial genomes post-capture. However, we anticipated a small amount of variation because the targeted sequence capture library had been through more sample handling in the form of incubations, dilutions, and amplifications. We found that the proportions of the non-viral sequences were strongly correlated (Pearson's correlation value: r=0.9881 to 0.9996) (**Supplemental Table S5 and Supplemental Figure S2**). A slightly higher percentage (mean 5.8%, median 5%, range 0 to 10.7%) of reads aligned to non-viral reference genomes in the post-capture data compared with pre-capture in all but one of the samples. However, the distribution of sequences among reference genomes did not show a systematic bias. This can be seen in the conserved distribution of sequences among human chromosomes (**Supplemental Figure S3**).

DISCUSSION

We designed the ViroCap panel to enhance the sensitivity of metagenomic shotgun sequencing for comprehensive detection of known vertebrate viruses as well as detection of divergent viruses that have nucleotide sequence similarities to known viruses. Here we have demonstrated that targeted sequence capture using ViroCap dramatically increases the amount of viral sequence obtained from human samples compared to conventional MSS, greatly enhancing the resolution of genomic characterization and increasing the number of viruses detected by more than 50%. Enhancement was demonstrated for DNA and RNA viruses from multiple diverse families. The increased sensitivity will be valuable in multiple research applications including descriptions of the human virome and will also improve the potential for MSS as a diagnostic tool in human and animal health.

The dramatic enrichment of viral nucleic acids present within the targeted sequence capture libraries offers important advantages. First, as we demonstrate, MSS with ViroCap can be used

to generate complete or nearly complete genome sequences directly from clinical samples, including those with very low proportions of viral nucleic acid, without culturing the viruses. Availability of extensive sequence data provides the opportunity to distinguish among closely related virus subtypes or even among viral strains, which might not be distinguished by other types of assays. In the data set presented here, we demonstrated the ability to type rhinoviruses, and distinguish between human herpesvirus 6B and 6A, adenovirus types A and C, and polyomaviruses JC and BK. Notably, influenza A virus was identified pre-capture but could only by typed as an H3N2 virus post-capture. Elsewhere, we used ViroCap to sequence the enterovirus D68 genome directly from clinical samples (Wylie et al. 2015), and in that work, the extensive sequence data that we obtained allowed us carry out detailed comparative analysis of closely related strains that differed at a limited number of nucleotide positions. Second, the use of ViroCap can reduce the depth of sequencing needed to detect viruses in samples. Because targeted sequence capture results in a large increase in the percentage of sequencing reads that are viral (**Fig. 2, Tables 2 and 3, Supplemental Tables 2-4**), ViroCap achieves better viral coverage while requiring the generation of fewer total sequence reads. This increased efficiency has the potential to lower sequencing costs.

An important feature of ViroCap is the tiling of capture probes across genomes, including highly conserved regions which may allow detection of genomic fragments of divergent viruses that share little overall sequence homology with known viruses. We illustrated such capability using anelloviruses containing divergent nucleotide sequence (**Fig. 3**). In addition, the inclusion of Genome Neighbor targets enhanced our design not only by expanding beyond the tiled RefSeq viruses, but also by adding sensitivity for genomic regions where RefSeq capture probes alone might not have captured divergent strains (see Methods). ViroCap cannot detect viruses that do not share any nucleotide sequence similarity to known viruses; however, we note that because the enrichment of viral nucleic acids occurs after sequence library construction, the uncaptured

235 portion of the library could subsequently be sequenced for additional attempts at pathogen
236 discovery. Furthermore, the ViroCap panel is extensible and will be updated periodically with
237 new viral sequences as they are added to RefSeq and the Genome Neighbors databases.
238 Updates will be publicly available through our GitHub repository (see Methods: Data access).
239
240 There were a few genomes (<10) in the NCBI reference databases that had been cloned into
241 bacterial vectors prior to sequencing, and the deposited viral genome sequences contained
242 bacterial vector sequence. We were not aware of this prior to probe design, so ViroCap includes
243 capture probes that target these sequences. This resulted in enrichment of some sequences (on
244 average 1.1% of total non-viral reads) that were subsequently recognized by our analysis
245 pipeline as bacterial based on nucleotide sequence alignment. In subsequent versions of
246 ViroCap, we will filter out these bacterial vector sequences.
247
248 In the experiments reported here, we pooled sequencing libraries prior to targeted sequence
249 capture in order to reduce cost, but still achieved enhanced detection of multiple viruses of
250 varying abundance. As has been reported for strategies that involve sequencing indexed,
251 pooled libraries (Kircher et al. 2012), we observed some sample cross-contamination. This
252 cross-contamination is recognizable when a high number of viral sequences are detected in the
253 truly positive sample, while few sequences ($\leq$0.05% of the viral sequences in the truly positive
254 sample) of the same virus are detected in other samples in the pool. In a clinical setting, each
255 sample would optimally be captured and sequenced independently to reduce the possibility of
256 sample cross-contamination. However, future methodological improvements could allow pooling
257 of clinical specimens.
258
259 The success of viral targeted sequence capture is affected by the representation of the virus in
260 the sequencing library. In our sample preparation, total nucleic acid extracted from the sample

was reverse transcribed and randomly amplified prior to library construction (Wang et al. 2003), allowing detection of DNA and RNA viral genomes within the same sequencing experiment. The uneven sequence representation observed for some genomes (**Fig. 2**) is likely due in part to detection of messenger RNA, whose abundance reflects patterns of gene expression, as well as to primer biases during the reverse transcription and amplification steps. Capture hybridization may also induce bias, in that sequences that diverge from target probe sequences may be captured less efficiently than those with exact or nearly exact matches to the probe. Taken together, these data suggest that further improvement in performance of viral targeted sequence capture may be achievable by improving efficiency of reverse transcription, amplification, and library construction, while continuing to update the ViroCap panel as new, divergent genome sequences become available.

Methods other than genome sequencing have been used for virus characterization and discovery, including Virochip, a microarray-based method for detection/genotyping of viral pathogens (Chen at al. 2011; Wang et al. 2002), and PathoChip, a microarray designed to detect viruses and other microbial pathogens (Baldwin DA et al. 2014). While designed to detect known viruses by means of microarray probe spotting, this technology has also shown the ability to detect emerging viruses (Yu et al. 2012; Wang et al. 2003). The primary difference between the designs of these microarrays and ViroCap targeted sequence capture is that the latter approach targets complete viral genomes while the microarrays target smaller, discrete genomic regions. The results obtained from each approach also differ significantly. The microarray approach detects the presence of a virus but does not directly provide sequence information. In contrast MSS enhanced by ViroCap targeted sequence capture provides sequence data, sometimes covering the entire genome.

In conclusion, ViroCap greatly enhances the sensitivity of MSS for nucleotide sequence-based virus detection. To our knowledge, ViroCap represents the first effort to apply a targeted sequence capture approach to the detection of a comprehensive set of viruses. Its research applications are far reaching, allowing a new, higher resolution view of eukaryotic DNA and RNA viruses in the microbiome. ViroCap should also help realize the potential of MSS as a clinical diagnostic tool that can simultaneously detect viruses as well as provide immediate characterization including taxonomic assignment, strain typing, virulence characteristics, and anti-viral drug resistance genotyping. ViroCap could also be modified into a tool for broader pathogen identification, which might include a comprehensive set of human pathogens: genes from viruses, bacteria (e.g. toxin genes, antibiotic resistance genes), fungi, protists, and other microbes.

METHODS

Taxonomy selection. At the time we designed the ViroCap panel, NCBI GenBank had available for download a total of ~1 Gb of sequence representing 440 viral families, well beyond the 200 megabase (Mb) of target space supported by the custom SeqCap EZ library format (NimbleGen, Madison, WI, USA). Therefore, we developed the following approach for selecting representative targeted sequence capture probes. Because we were interested in studying viral diseases of humans, we excluded bacteriophages and endogenous human retroviruses. We also specifically did not include references from the following NCBI viral reference genome database host categories: algae, archaea, bacteria, diatom, environment, fungi, invertebrates, plants, and protozoa. After filtering, our target list contained reference sequences from the following host categories: human, vertebrates, and “unknown”. This list included viruses that could have both vertebrate and invertebrate hosts, such as vertebrate viruses with insect vectors. Based on these broad viral-host categories, we downloaded all of the associated viral reference sequences in each chosen category from NCBI (accessed February 3, 2014). These

sequences comprise the core reference database from which our capture library is designed. Our capture library includes targets from 34 viral families composed of 190 annotated viral genera and 337 species (**Fig. 1, Supplemental Tables S6 and 7**). Sources of viral sequences include complete representation of the viral genomes from NCBI's Reference Sequence (RefSeq) collection, complementary representation of unique regions from Genome Neighbor targets, selected representation of NCBI Influenza Virus Resource sequences, and the entirety of the probe space represented on the Virochip microarray (Yu G et al. 2012), GEO accession number GPL15905. The methods used to consolidate these database sequences follow.

RefSeq. NCBI's RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) genome collection is a database of taxonomically diverse entries representing comprehensive, well-annotated genome sequences (Pruitt et al. 2014; Tatusova et al. 2014). As RefSeq entries are the most complete sequence representatives in terms of annotation and metadata consistency, we targeted selected viral RefSeqs by tiling of targeted sequence capture probes across the entire length of each RefSeq's genome, with the intention of capturing the entire viral genome. For our capture library, RefSeq nucleotide FASTA sequences were downloaded for desired viral-host categories (human; vertebrates; vertebrates, human; vertebrates, invertebrates; vertebrates, invertebrates, human; invertebrates, vertebrates; unknown) using both the online NCBI taxonomy viewer (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=virus&taxid=10239) as well as the RefSeq specific FTP site (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral). Entries were merged to avoid redundancy. RefSeq targets were pooled with the other sequence candidates (see: **Design consolidation**). A total of 1,456 RefSeq FASTA entries (26.9 Mb) representing 190 viral genera were completely tiled for inclusion in the ViroCap library, accounting for 13.5% of the total capture library's target-space.

Genome Neighbors. While RefSeq entries are single, canonical species representations, other complete or partial viral sequences also exist in DDBJ/EMBL/GenBank. In the case of viral sequences, there is extensive redundancy in these databases due to the large number of similar viral strains, isolates, and mutants. Therefore, non-RefSeq (e.g. DDBJ, EMBL, GenBank) nucleotide sequences of complete viral genomes that belong to the same species as a RefSeq are classified as Genome Neighbors for that reference sequence, provided that they match all of the criteria that were used to select complete genomic sequences (Bao et al. 2004). At the time of our ViroCap panel design, Genome Neighbors (sequences downloaded from Entrez Genome link "Other genomes for species"; accessed February 3, 2014) in total represented an additional 56,314 entries and 507.1 Mb of sequence, more than 2.5 times our SeqCap EZ capture target sequence space limit. Therefore, an alternative target selection approach was employed to add diversity to our RefSeq selections by selecting unique, complementary Genome Neighbor sequences.

RefSeq and Genome Neighbor sequence association. We began the process of variant sequence selection by identifying conserved regions in Genome Neighbors already represented by completely tiled RefSeq capture probes. First, we associated our viral RefSeq selections with corresponding Genome Neighbors. This was performed by downloading Genome Neighbor annotation files from NCBI (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=virus&taxid=10239) and associating the information with our RefSeq annotation files, by means of *ad hoc* Perl parsing/coupling scripts (results in **Supplemental Tables S6 and S7**). Once associated, the parent RefSeq sequences could be compared to related Genome Neighbor sequences to determine conserved and divergent nucleotide regions. Each viral RefSeq entry was individually reviewed, along with associated Genome Neighbor entries. FASTA sequences were collected for each RefSeq entry and its related Genome Neighbors for subsequent *k*-mer analysis.

***K*-mer analysis.** Each of the Genome Neighbor sequences was split into 100 bp *k*-mers by means of an exhaustive 1 bp sliding window algorithm, as depicted in **Supplemental Figure S4**. The resultant output thus included all possible 100 bp sequences based on the combined Genome Neighbor sequence space. As our SeqCap EZ targeted sequence capture probe lengths are 100 bp, the sequences generated by the sliding window algorithm represent the total number of possible probe combinations based on the aggregate of Genome Neighbor sequences. Based on our conservative expectation of hybridization/homology at the capture probe level, we then clustered all of the Genome Neighbor 100-mers back to the parent RefSeq sequence at >=90% sequence identity using length-sorted FASTA entries and the UCLUST (Edgar 2010) package (version 1.1.579; parameters: --rev --id .90). Given that all of our candidate sequences were 100 bp in length, and all RefSeq entries are >100 bp, UCLUST always used the longer RefSeq as the first seed (*centroid*) in which to attempt folding of other sequences. As the parent RefSeq had complete probe tiling in our design, any Genome Neighbor 100-mer with >= 90% identity was considered already represented in our capture library, and therefore discarded. Genome Neighbor 100-mers with less than 90% identity were chosen for inclusion in the capture library. As the sliding window approach produces 100-mers that overlapped one another, we merged overlapping 100-mers based on their Genome Neighbor genomic coordinates into single contiguous spans using BEDTools (Quinlan and Hall 2010) functions.

Genome Neighbor sub-sequences. Resultant sub-sequences were excised as FASTA entries from corresponding Genome Neighbor references using WU-BLAST's (http://blast.wustl.edu) xdget application and added to the ViroCap panel. These supplementary entries are easily identifiable in our final target design, as the FASTA headers for the entries list the original parent sequence ID with the excised span indicated in curly braces (e.g.

gi|1249624|emb|A28090.1| HPV42 (partial) genomic sequence {SQ 2444-2644}). In this manner, for each RefSeq species, we generated Genome Neighbor sub-sequences from 100 bp to 21 kb in length to add to our capture panel.

These processing steps reduced the aggregate input Genome Neighbors targeted sequence space from 507.1 Mb to 153.2 Mb (**Supplemental Table S1**), and these sequences were pooled with our other targeted capture sequence targets (see: **Design consolidation**). A total of 130,808 partial Genome Neighbor FASTA entries (153.2 Mb) were added for capture in our ViroCap library, accounting for 77.1% of the total capture library's target-space.

Influenza Virus Resource. We obtained reference sequences from NCBI's Influenza Virus Resource database (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html), which contains sequence data from the NIAID Influenza Genome Sequencing Project as well as from GenBank. At the time of our capture panel design, the NCBI Influenza Virus Resource contained 305,524 influenza entries, representing 458.1 Mb of sequence. This is 17 times the size of our viral RefSeq selections and 3 times the size of our collapsed Genome Neighbor targets. Our selected RefSeqs included 29 influenza RefSeq entries (each influenza virus segment is represented as a separate entry), targeted in its entirety. These sequences served as the core of influenza reference genomes against which all other influenza sequences were compared. We directly clustered the long influenza sequences using length-sorted FASTA and the UCLUST package (version 1.1.579; parameters: --rev --id .90). In UCLUST, a cluster is defined by one sequence, known as the *centroid* or representative sequence. To lessen the computational burden and ensure that our core influenza RefSeq genomes were always the longest first seeds (*centroids*) in UCLUST's clustering process, we artificially concatenated the 29 parent RefSeq sequences into one linear sequence representation, and then split this representation into 6 segments ranging in size from 18-26 kb. UCLUST preferentially seeded

with the long RefSeq construct segments when clustering, ensuring that clustering was first attempted within the longer, canonical references. ULCUST was run with a requirement of >=90% sequence identity to fold into a parent influenza RefSeq segment. Therefore, only sequences that (1) had <90% identity to influenza RefSeqs and (2) were subsequent centroids in non-RefSeq clusters were chosen for inclusion in our capture panel. This process reduced the aggregate input Influenza Resource Database reference sequence from 458.1 Mb to 15.7 Mb (**Supplemental Table S1**). Finally, supplementary influenza targets were pooled with the other sequence candidates (see: **Design consolidation**). A total of 9,759 influenza FASTA entries (15.7 Mb) were added for targeted sequence capture in our ViroCap library, accounting for 7.9% of the total capture library's target-space.

Virochip microarray. Considering the biologically important short sequence signatures represented on the Virochip panel (Yu G et al. 2012), as well as the comparatively small footprint, we subsumed these sequences within our targeted sequence capture panel design. The probe sequences for the microarray are publicly available at NCBI's Gene Expression Omnibus (GEO) repository (Edgar et al. 2002). We downloaded this information for Platform GPL15905 (Viro5AG-60k) as a text file (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL15905. This platform included more than 60,000 oligonucleotides of length 60-70 bp, corresponding to 3.1 Mb of probes (**Supplemental Table S1**). Virochip targets were pooled with the other sequence candidates (see: **Design consolidation**). Upon review, 1.3 Mb of the probes were already directly represented by RefSeq, Genome Neighbor, and Influenza Viral Resource sequences during capture library design and synthesis. Therefore, the remaining 25,749 (60-70 bp) Virochip FASTA entries of 1.8 Mb total size were added to ViroCap, accounting for <1% of the total targeted sequence capture panel.

Design consolidation. All of our selected candidate target sequences from RefSeq, Genome Neighbors, Influenza Virus Resource, and the Virochip microarray were combined into a single FASTA sequence file. Human endogenous retroviruses were removed from inclusion by means of a two-part process: 1) entries were filtered by taxonomic annotation indicating human endogenous retrovirus identity; 2) the remaining entries were BLAST-aligned to the GRCh37-lite build of the human reference genome to remove sequences with high percent identity (>=75%) at the 100 bp probe level. Finally, sequences were hard-masked (i.e. bases converted to N's) in low complexity regions using the DUST (R. Tatusov and D.J. Lipman, unpublished) software module. The final ViroCap targeted sequence capture panel consists of 185,835 FASTA sequences totaling 198.9 Mb (see: **Taxonomy selection**).

NimbleGen Sequence Capture design. Our consolidated target sequences were submitted to Roche NimbleGen (Madison, WI, USA) on March 31, 2014 for capture library design and synthesis. As our final ViroCap design required 198.9 Mb, manufacturing was implemented under the custom NimbleGen SeqCap EZ Developer Library format, which has a maximum capture space of 200 Mb of non-human sequence. NimbleGen's Sequence Capture design offered up to 2.1 million of 50-105mer sequence probes. It was at the discretion of NimbleGen, based on proprietary algorithms, to redistribute probes for better capture uniformity, redundancy, and comprehensive target base coverage. NimbleGen provided us with a proposed capture design accompanied by coordinate (GFF, BED) files and associated sequence coverage metrics on April 14, 2014. The design set contained probe representation generated by first masking all but one exact copy of each 100-mer in our original FASTA file, tiling the unmasked regions, screening the resulting probes against the (hg19) human genome, and finally selecting only those probes that had no matches in the human genome as determined by the SSAHA (Ning et al. 2001) algorithm. NimbleGen provides two metrics for assessing in silico targeted sequence capture design coverage: 1) target bases covered with 0-bp-offset are

467 determined by counting target bases directly represented in probe sequences; 2) target bases
468 covered with 100-bp-offset are determined by counting all target bases within 100 bp of a probe.
469 The capture design provided 95.9% 0-bp-offset coverage and 99.6% 100-bp-offset coverage of
470 our initial 198.9 Mb target request. We approved the design on April 17, 2014 for capture library
471 synthesis and received our first 12 SeqCap EZ Library reactions for in-house Illumina
472 sequencing and analysis on April 28, 2014.
473
474 **Human subjects approval and sample selection.** Samples were collected under protocols
475 approved by the Human Research Protection Office at Washington University School of
476 Medicine (IRB protocol numbers 201106177, 201102561, and 201102045). Samples were
477 selected to represent a broad range of viruses that are commonly encountered in the clinical
478 laboratory and in our research studies. Viruses were identified in samples based on clinical
479 laboratory test results in the Diagnostic Virology Laboratory at St. Louis Children's Hospital or
480 by PCR assays and sequencing results carried out in previous studies (Colvin et al. 2012;
481 McElvania TeKippe et al. 2012; Wylie et al. 2012). Specimens of nasopharyngeal secretions,
482 plasma, and stool were included.
483
484 **Sequencing.** Total nucleic acid was extracted from clinical samples using the EasyMag
485 NucliSENS instrument (bioMerieux, Marcy l'Etoile, France). Samples were processed in one of
486 two ways. In experiment 1, nucleic acids from clinical specimens from the Diagnostic Virology
487 Laboratory were combined, and the resulting pooled nucleic acid was used as input for a single
488 sequencing library (constructed as described below). These samples are designated with a
489 sample identification prefix of "P" in the various tables and figures. Alternatively, in experiment
490 2, individual sequencing libraries were made from each set of 8 different specimens prior to
491 combining the libraries for sequencing. These samples are designated with a sample
492 identification prefix of "S" in the various tables.

For sequencing libraries, DNA and RNA viruses were assessed in the same assay as described (Wang et al. 2003). Specifically, the RNA in the total nucleic acid was reverse transcribed with reverse transcriptase (Promega, Fitchburg, Wisconsin) and random nonomers tagged with a conserved sequence (5' GTTTCCCAGTCACGATA 3') to be used for subsequent amplification (Integrated DNA Technologies, Coralville, Iowa), and second strand synthesis was carried out with Sequenase V2.0 DNA polymerase (Affymetrix, Santa Clara, California). DNA and RNA were subsequently amplified with Accuprime Taq (Life Technologies, Grand Island, New York) using the conserved sequence on the ends of the random primers, and the DNA/cDNA mixture was sheared using the Qsonica Q800R instrument (Qsonica, Newtown, Connecticut) to generate fragments with an average length of 500 basepairs. Dual-indexed sequencing libraries were constructed using the KAPA Low Throughput Library Construction Kit (KAPA Biosystems, Wilmington, Massachusetts).

For the anellovirus samples, DNA was amplified with the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Pittsburgh, Pennsylvania); RNA was not assessed. DNA was sheared and libraries were constructed from each sample as described above. Sequencing libraries were pooled (**Tables 2 and 3, column "Sample ID"**).

In each case, the libraries were divided, and part was directly sequenced (pre-capture) and part was subjected to targeted sequence capture with the custom ViroCap probes prior to sequencing (post-capture). Targeted sequence capture was carried out according to the manufacturer's specifications. We carried out 10, 10, and 16 cycles of post-capture linker-mediated PCR for experiments 1 (pooled clinical samples), 2 (individual samples from the research study) and 3 (anellovirus samples), respectively, prior to sequencing. The number of cycles was empirically determined to be the minimum number needed to obtain a 5 nanomolar

dilution of library material for qPCR and loading. Libraries were sequenced on the Illumina HiSeq 2000 or HiSeq 2500 instrument, generating 100 bp, paired-end reads.

Sequence analysis. Viral sequences were identified based on nucleotide and translated protein sequence alignment against reference genomes. The pipeline is adapted from previously described methods (Wylie et al. 2014), except that nucleotide alignments were carried out using BWA-MEM with default settings (Li and Durbin 2009). Because many similar genomes are included in the reference database, we used the initial alignment statistics for each sample to choose a single reference from each species to calculate and report coverage statistics. References were chosen based on having the highest number of reference bases covered. Sequences were re-aligned to the selected references with BWA-MEM for calculation of coverage statistics and comparison of samples pre- and post-capture. Sequence alignments were evaluated with samtools (Li 2011), and sequence coverage was determined with RefCov (http://gmt.genome.wustl.edu/packages/refcov/) and visualized with Plot2 (http://plot2doc.micw.eu). For illustrative purposes, the genome coverage panels in **Figures 2 and 3** were normalized by removing (*deduplicating*) reads based on identical alignment start-sites using the samtools *rmdup* command. For each alignment start-site, only the highest quality read was retained for forward and reverse alignment orientations. Therefore, for the 100 bp read data shown in each coverage panel the theoretical maximum depth is 200X.

Anellovirus contigs were assembled from the pre-capture sequence data using IDBA-UD (Peng et al. 2012). Contigs were aligned against the sequence database used to design the ViroCap panel using BLAST (Altschul et al. 1997) with the following parameters to detect low-similarity sequences: -G 5 -E 2 -r 1 -q -1. The percent identity of the highest scoring HSP is reported in **Supplemental Table S3**.

DATA ACCESS

Data files associated with our ViroCap panel are publicly available through the https://github.com/WashU-PMG/ViroCap GitHub repository. Hosted files include 198.9 Mb of ViroCap target sequence in FASTA format, taxonomy information, corresponding RefSeq and Genome Neighbor associations, and NimbleGen's target design coverage metrics for 0- and 100-bp-offset intervals. MSS data used for ViroCap evaluation have been deposited (with potentially identifiable human sequences removed) in the NCBI Sequence Read Archive (PRJNA273884).

ACKNOWLEDGMENTS

We thank Richard Hotchkiss and Andrew Walton for providing samples for the anellovirus work; Maria Cannella and Richard Buller for carrying out the virus-specific PCR assays on the samples from the research study; Stephanie Bledsoe and the staff of the Virology Laboratory at St. Louis Children's Hospitals for supplying remnant specimen material; Elena Deych and William Shannon for assisting with statistical analysis; and The McDonnell Genome Institute for generating sequence data. We thank Efrem Lim and David Wang for performing the PCR validation test on the HPyV10 virus.

AUTHOR CONTRIBUTIONS

TNW and KMW developed and benchmarked the methodology. TNW developed computational tools for targeted sequence capture selection. BNH and KMW generated MSS libraries, optimized molecular methods, and validated results. TNW, KMW, and GAS performed the MSS analysis and virus review. KMW and GAS were involved with study design. GAS provided patient samples representing a broad range of viruses that are commonly encountered in the clinical laboratory and provided the samples from the research study. TNW, KMW, and GAS wrote the manuscript.

FUNDING

This study was supported by grant number R01AI097213 from the National Institute of Allergy and Infectious Diseases, awarded to GAS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have no conflicts of interest to report.

FIGURE LEGENDS

Figure 1. Taxonomic distribution of target genomes included in ViroCap. Shown are the viral classes, families, and genera included in the ViroCap targeted sequence capture panel. Taxonomic assignments were obtained from the NCBI Taxonomy Viewer (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=virus&taxid=10239).

Figure 2. Targeted sequence capture enrichment. Examples are given showing the impact of targeted sequence capture on breadth and depth of genome coverage for 8 representative viral genomes. For illustrative purposes, all of the coverage panels in this figure have been normalized by removing (*deduplicating*) reads based on identical alignment start-sites. Nucleotide positions along the reference genome are shown on the X-axis. The depth of deduplicated reads is shown on the Y-axis. The shaded portion indicates the sequence coverage (breadth and depth) for each virus. Post-capture sequence coverage is represented in the larger panels in blue; pre-capture sequence coverage is shown in the insets in red. Note that Y-axis ranges are different for each panel. At the top of each panel is shown the breadth of coverage (BoC) for the sample. The header of each panel includes breadth of coverage gain (BoC gain), sample id, and reference genome name and NCBI version number. BoC gain is calculated by subtracting the percent of the length of the reference genome that was covered by

sequence reads in pre-capture MSS from the percent of the length of the reference genome covered by post-capture sequence reads.

Figure 3. Targeted sequence capture identifies divergent sequences. (A) The percent identity of the top HSP identified from the BLAST alignment of anellovirus contig sequences to the references used to design ViroCap is plotted on the Y-axis. The X-axis represents the percent of the length of the anellovirus contig covered after targeted sequence capture. (B) This coverage plot represents the sequence coverage of a divergent anellovirus contig sequence. The figure is designed as described in the figure legend for **Figure 2**, with the following addition: the post-capture coverage plot is shaded to show regions of nucleotide sequence variation between the anellovirus contig and the most similar reference genome in the ViroCap panel. Dark shading represents areas of identical sequence, and each position with nucleotide mismatch between aligned sequences is shown in the lighter color. All of the HSPs are shown, rather than just the top HSP.

REFERENCES

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**: 903–905.

Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. 2001. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci USA* **98**: 11609–11614.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, et al. 2011. Enterotypes of the human gut microbiome. *Nature* **473**: 174–180.

Baldwin DA, Feldman M, Alwine JC, Robertson ES. 2014. Metagenomic assay for identification of microbial pathogens in tumor tissues. *MBio* **5**: e01714–14.

Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T. 2004. National center for biotechnology information viral genomes project. *J Virol* **78**: 7291–7298.

Breitbart M, Rohwer F. 2005. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *BioTechniques* **39**: 729–736.

Chen EC, Miller SA, DeRisi JL, Chiu CY. 2011. Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. *J Vis Exp* e2536–e2536.

Chiu CY. 2013. Viral pathogen discovery. *Curr Opin Microbiol* **16**: 468–478.

Cleland EJ, Bassioni A, Boase S, Dowd S, Vreugde S, Wormald P-J. 2014. The fungal microbiome in chronic rhinosinusitis: richness, diversity, postoperative changes and patient outcomes. *Int Forum Allergy Rhinol* **4**: 259–265.

Colvin JM, Muenzer JT, Jaffe DM, Smason A, Deych E, Shannon WD, Arens MQ, Buller RS, Lee WM, Weinstock EJS, et al. 2012. Detection of viruses in young children with fever without an apparent source. *Pediatrics* **130**: e1455–62.

de Villiers E-M, Borkosky SS, Kimmel R, Gunst K, Fei J-W. 2011. The diversity of torque teno viruses: in vitro replication leads to the formation of additional replication-competent subviral molecules. *J Virol* **85**: 7284–7295.

De Vlaminck I, Khush KK, Strehl C, Kohli B, Luikart H, Neff NF, Okamoto J, Snyder TM, Cornfield DN, Nicolls MR, et al. 2013. Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* **155**: 1178–1187.

Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P, Breuer J. 2011. Specific capture and whole-genome sequencing of viruses from clinical samples. ed. R. Jhaveri. *PLoS ONE* **6**: e27805.

Duhaime MB, Sullivan MB. 2012. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434**: 181–186.

Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, Abel HJ, Pfeifer JD. 2011. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J Mol Diagn* **13**: 325–333.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207–210.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, NIH Intramural Sequencing Center Comparative Sequencing Program, et al. 2013. Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498**: 367–370.

Gajer P, Brotman RM, Bai G, Sakamoto J, Schütte UME, Zhong X, Koenig SSK, Fu L, Ma ZS, Zhou X, et al. 2012. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med* **4**: 132ra52–132ra52.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**: 1522–1527.

Holtz LR, Cao S, Zhao G, Bauer IK, Denno DM, Klein EJ, Antonio M, Stine OC, Snelling TL, Kirkwood CD, et al. 2014. Geographic variation in the eukaryotic virome of human diarrhea. *Virology* **468-470**: 556–564.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207–214.

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

Koehler JW, Hall AT, Rolfe PA, Honko AN, Palacios GF, Fair JN, Muyembe J-J, Mulembekani P, Schoepp RJ, Adesokan A, et al. 2014. Development and evaluation of a panel of filovirus sequence capture probes for pathogen detection by next-generation sequencing. ed. J.H. Kuhn. *PLoS ONE* **9**: e107007.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

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

697 Lovett M, Kere J, Hinton LM. 1991. Direct selection: a method for the isolation of cDNAs
698 encoded by large genomic regions. *Proc Natl Acad Sci USA* **88**: 9628–9632.

699 Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, Lindberg AM, Persson
700 B, Allander T, Andersson B. 2012. Characterization of the viral microbiome in patients with
701 severe lower respiratory tract infections, using metagenomic sequencing. ed. S.K.
702 Highlander. *PLoS ONE* **7**: e30875.

703 McElvania TeKippe E, Wylie KM, Deych E, Sodergren E, Weinstock G, Storch GA. 2012.
704 Increased prevalence of anellovirus in pediatric patients with fever. ed. R. Jhaveri. *PLoS*
705 *ONE* **7**: e50937.

706 Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011. The
707 human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**:
708 1616–1625.

709 Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases.
710 *Genome Res* **11**: 1725–1729.

711 Ninomiya M, Takahashi M, Shimosegawa T, Okamoto H. 2007. Analysis of the entire genomes
712 of fifteen torque teno midi virus variants classifiable into a third group of genus Anellovirus.
713 *Arch Virol* **152**: 1961–1975.

714 Oh J, Byrd AL, Deming C, Conlan S, NISC Comparative Sequencing Program, Kong HH, Segre
715 JA. 2014. Biogeography and individuality shape function in the human skin metagenome.
716 *Nature* **514**: 59–64.

717 Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. 2007. Microarray-based
718 genomic selection for high-throughput resequencing. *Nat Methods* **4**: 907–909.

719 Paulino LC, Tseng C-H, Strober BE, Blaser MJ. 2006. Molecular analysis of fungal microbiota in
720 samples from healthy human skin and psoriatic lesions. *J Clin Microbiol* **44**: 2933–2941.

721 Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell
722 and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–
723 1428.

724 Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart
725 J, Landrum MJ, McGarvey KM, et al. 2014. RefSeq: an update on mammalian reference
726 sequences. *Nucleic Acids Res* **42**: D756–63.

727 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
728 features. *Bioinformatics* **26**: 841–842.

729 Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in
730 the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**: 334–338.

731 Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes

database: new representation and annotation strategy. *Nucleic Acids Res* **42**: D553–9.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. 2009. A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL. 2002. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* **99**: 15687–15692.

Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, et al. 2003. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* **1**: E2.

Willger SD, Grim SL, Dolben EL, Shipunova A, Hampton TH, Morrison HG, Filkins LM, O'Toole GA, Moulton LA, Ashare A, et al. 2014. Characterization and quantification of the fungal microbiome in serial samples from individuals with cystic fibrosis. *Microbiome* **2**: 40.

Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. 2012. Sequence analysis of the human virome in febrile and afebrile children. ed. C. Zhang. *PLoS ONE* **7**: e27735.

Wylie KM, Mihindukulasuriya KA, Zhou Y, Sodergren E, Storch GA, Weinstock GM. 2014. Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol* **12**: 71.

Wylie KM, Wylie TN, Orvedahl A, Buller RS, Herter BN, Magrini V, Wilson RK, Storch GA. 2015. Genome Sequence of Enterovirus D68 from St. Louis, Missouri, USA. *Emerging Infect Dis* **21**(1), 184-186.

Young JC, Chehoud C, Bittinger K, Bailey A, Diamond JM, Cantu E, Haas AR, Abbas A, Frye L, Christie JD, et al. 2014. Viral Metagenomics Reveal Blooms of Anelloviruses in the Respiratory Tract of Lung Transplant Recipients. *Am J Transplant* **15**: 200-209.

Yu G, Greninger AL, Isa P, Phan TG, Martínez MA, la Luz Sanchez de M, Contreras JF, Santos-Preciado JI, Parsonnet J, Miller S, et al. 2012. Discovery of a novel polyomavirus in acute diarrheal samples from children. ed. T. Ramqvist. *PLoS ONE* **7**: e49449.

Figure 1

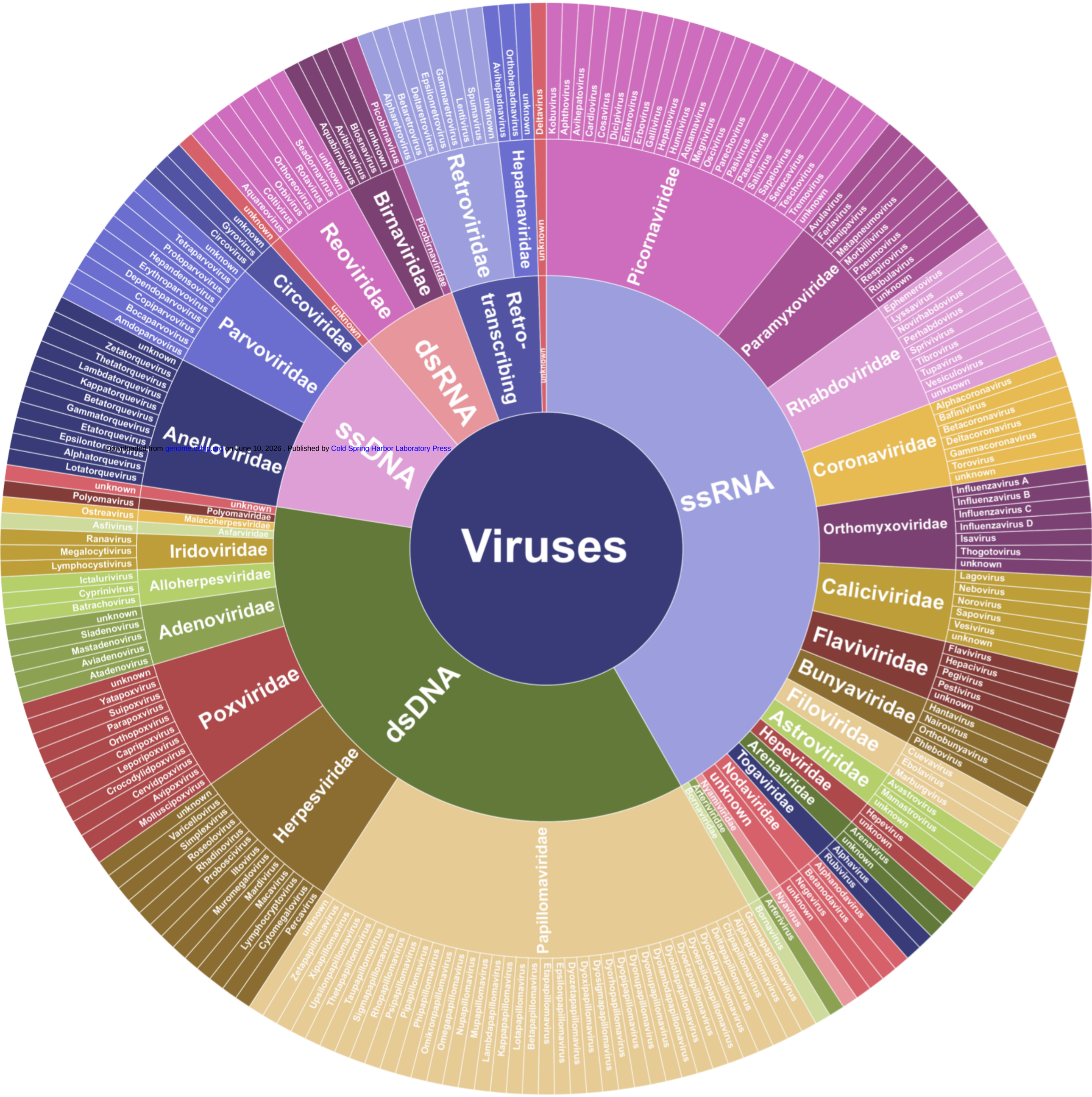


Figure 2

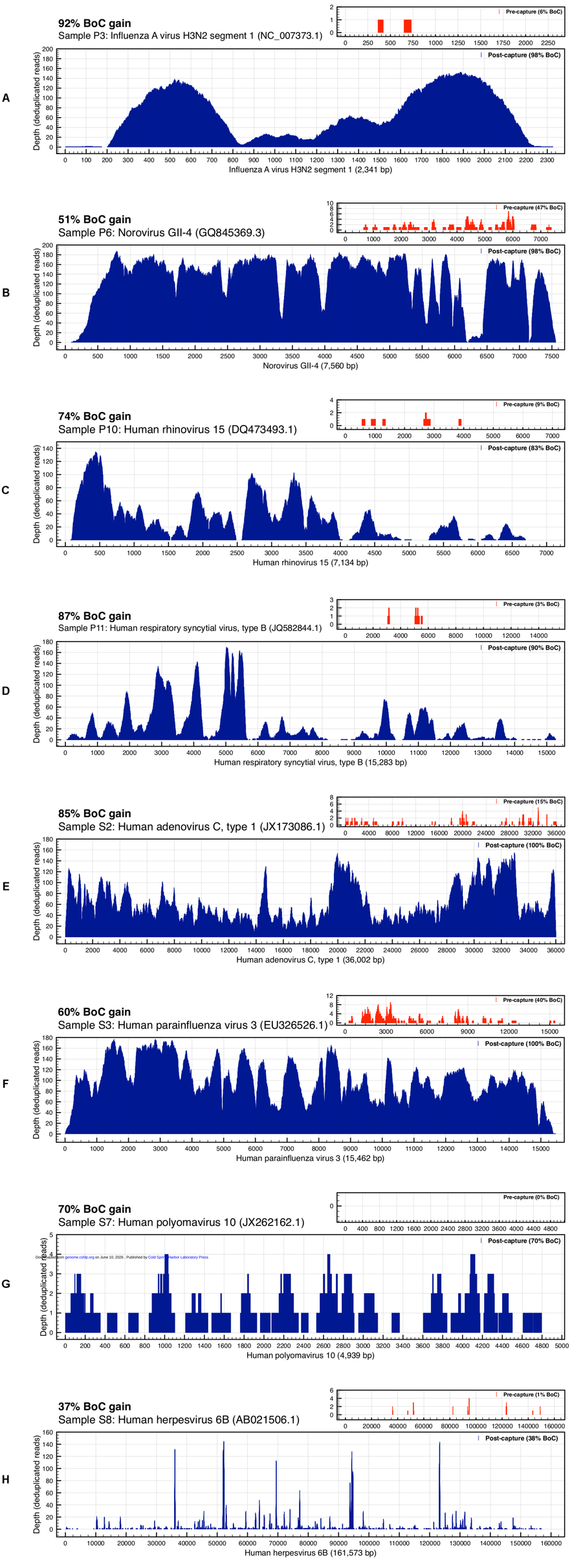
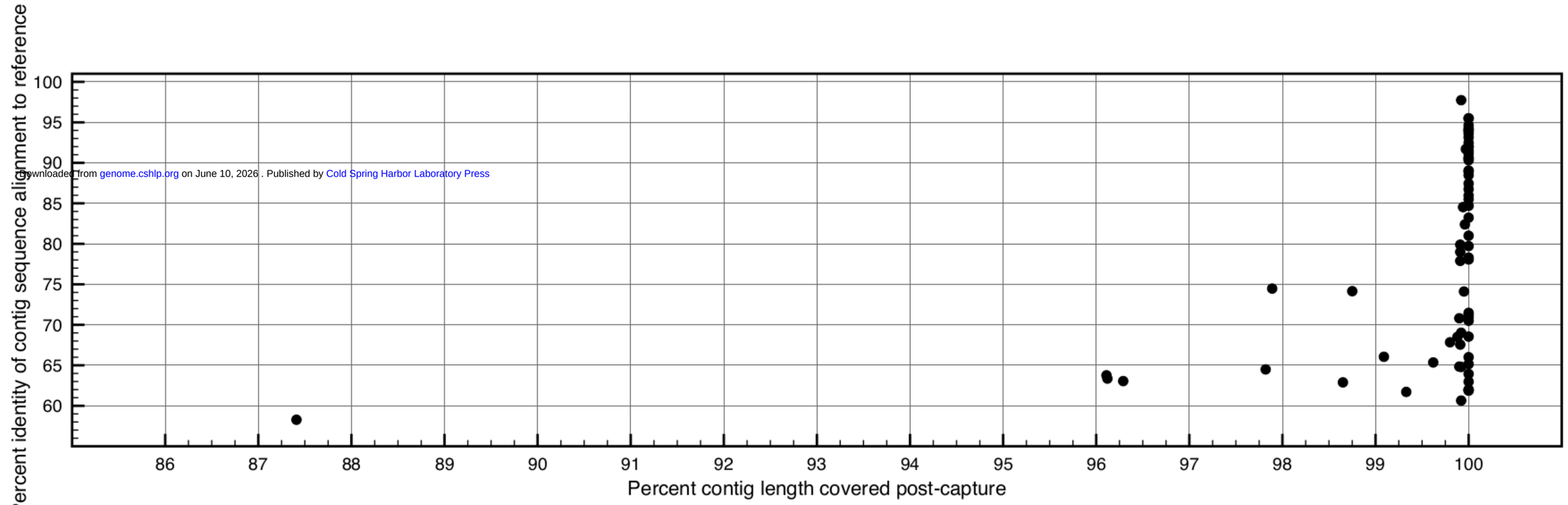

A
92% BoC gain
Sample P3: Influenza A virus H3N2 segment 1 (NC_007373.1)
Pre-capture (6% BoC)
Post-capture (98% BoC)
Depth (deduplicated reads)
Influenza A virus H3N2 segment 1 (2,341 bp)
B
51% BoC gain
Sample P6: Norovirus GII-4 (GQ845369.3)
Pre-capture (47% BoC)
Post-capture (98% BoC)
Depth (deduplicated reads)
Norovirus GII-4 (7,560 bp)
C
74% BoC gain
Sample P10: Human rhinovirus 15 (DQ473493.1)
Pre-capture (9% BoC)
Post-capture (83% BoC)
Depth (deduplicated reads)
Human rhinovirus 15 (7,134 bp)
D
87% BoC gain
Sample P11: Human respiratory syncytial virus, type B (JQ582844.1)
Pre-capture (3% BoC)
Post-capture (90% BoC)
Depth (deduplicated reads)
Human respiratory syncytial virus, type B (15,283 bp)
E
85% BoC gain
Sample S2: Human adenovirus C, type 1 (JX173086.1)
Pre-capture (15% BoC)
Post-capture (100% BoC)
Depth (deduplicated reads)
Human adenovirus C, type 1 (36,002 bp)
F
60% BoC gain
Sample S3: Human parainfluenza virus 3 (EU326526.1)
Pre-capture (40% BoC)
Post-capture (100% BoC)
Depth (deduplicated reads)
Human parainfluenza virus 3 (15,462 bp)
G
70% BoC gain
Sample S7: Human polyomavirus 10 (JX262162.1)
Pre-capture (0% BoC)
Post-capture (70% BoC)
Depth (deduplicated reads)
Human polyomavirus 10 (4,939 bp)
H
37% BoC gain
Sample S8: Human herpesvirus 6B (AB021506.1)
Pre-capture (1% BoC)
Post-capture (38% BoC)
Depth (deduplicated reads)
Human herpesvirus 6B (161,573 bp)

Figure 3

A



B

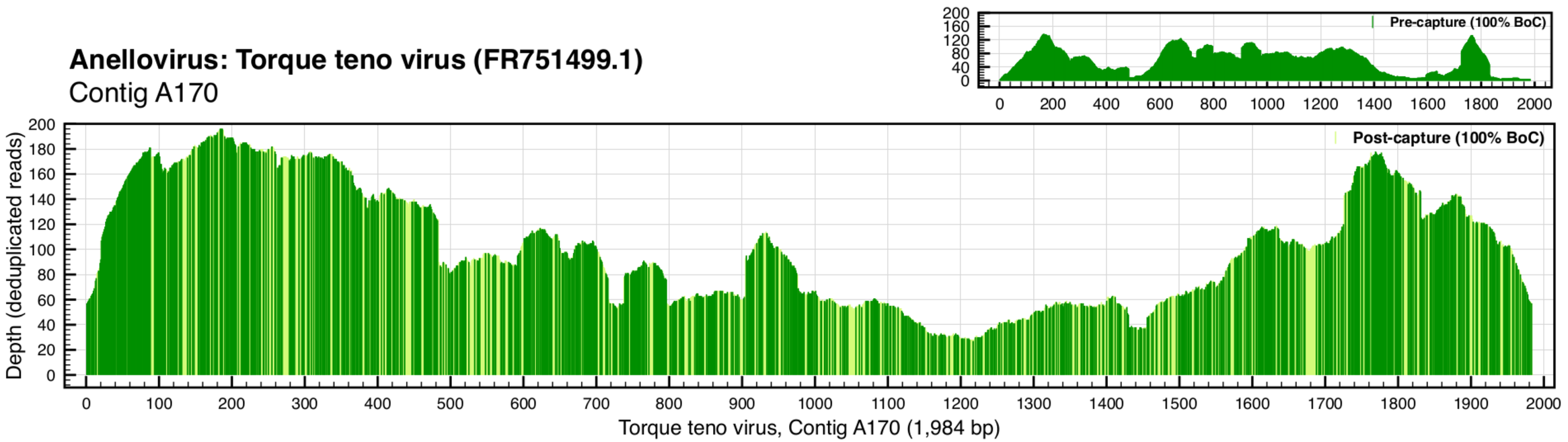


Table 1: Results of metagenomic shotgun sequencing for pooled specimens before and after viral targeted sequence capture

Virus[a]	Sample ID	Virus length (bp)	Viral sequence reads: Viral read count* Pre-capture[b]	Viral sequence reads: Viral read count* Post-capture[c]	Viral sequence reads: PVR fold increase[d]	Genome coverage: Breadth of coverage: percent Pre-capture	Genome coverage: Breadth of coverage: percent Post-capture	Genome coverage: Depth of coverage: mean (SD) Pre-capture	Genome coverage: Depth of coverage: mean (SD) Post-capture
Human adenovirus B, type 35	P1	34,794	5	8,103	1,300	0.9	83.6	<0.1 (0.1)	19.3 (52.7)
Human bocavirus 1	P2	5,299	110	15,277	111	68.6	100.0	1.8 (2.0)	251.1 (122.1)
Influenza A virus (H3N2)	P3	13,267	4	46,540	9,335	2.4	74.0	<0.1 (0.3)	263.0 (394.4)
Influenza B virus	P4	14,452	0	513	>513‡	0	9.8	0	2.6 (0.7)
Human parvovirus B19	P5	5,596	1,867	474,849	204	89.8	100.0	28.8 (17.9)	7,367.9 (4,193.0)
Norovirus GII-4	P6	7,560	72	527,656	5,880	46.5	98.4	0.8 (1.1)	5,870.4 (8,194.3)
Parechovirus 1	P7	7,339	0	13	>13‡	0	8.0	0	0.2 (0.6)
BK polyomavirus	P8	5,142	1	1,520	1,220	1.6	88.6	<0.1 (0.1)	24.9 (32.2)
JC polyomavirus	P9	5,121	5	2,760	443	8.6	98.5	0.1 (0.3)	46.3 (54.0)
Human rhinovirus 15	P10	7,134	8	9,624	965	8.8	82.9	0.1 (0.3)	115.9 (170.3)
Human respiratory syncytial virus, type B	P11	15,283	9	67,778	6,042	3.0	89.5	<0.1 (0.3)	350.1 (1,667.2)
Human herpesvirus 1	P12	152,261	0	14	>14‡	0	0.8	0	<0.1 (0.1)
Torque teno virus	P13	3,260	1	447	447	1.78	60.74	0.02 (0.13)	10.4 (23.4)
Human herpesvirus 3	P14	125,030	0	834	>834‡	0	8.33	0	0.6 (3.5)

Pre-capture: Metagenomic shotgun sequencing without targeted sequence capture.
Post-capture: Metagenomic shotgun sequencing using (ViroCap) targeted sequence capture.
* Viral reads per million sequences generated pre-capture and post-capture are statistically different (P<0.0001, Wilcoxon test).
[a] Viruses listed were incorporated into a viral pool that was subjected to MSS without and with targeted sequence capture, as described in the text.
[b] MSS of the virus pool without targeted sequence capture yielded 7,458,192 total reads.
[c] MSS of the virus pool after targeted sequence capture yielded 9,295,438 total reads.
[d] PVR (percent viral reads) fold increase: percent of post-capture viral reads divided by percent of pre-capture viral reads.
‡ PVR fold increase could not be calculated because the number of pre-capture reads was 0.

Table 2: Results of metagenomic shotgun sequencing for individual specimens before and after viral targeted sequence capture

						Viral sequence reads			Genome coverage			
		Total reads				Viral read count*			Breadth of coverage: percent		Depth of coverage: mean (SD)	
Sample type	Sample ID	Pre-capture	Post-capture	Virus(es) detected	Virus length (bp)	Pre-capture	Post-capture	PVR fold increase[a]	Pre-capture	Post-capture	Pre-capture	Post-capture
Nasopharyngeal swab	S1	4,202,474	16,080,640	Torque teno virus	3,736	0	231	>231‡	0	74.4	0	5.2 (7.6)
				Human adenovirus B, type 3	35,269	15,116	7,703,787	133	89.6	100.0	36.1 (213.8)	19,097.4 (17,285.5)
Nasopharyngeal swab	S2	5,087,234	3,713,962	TTV-like mini virus isolate	2,912	0	249	>249‡	0	37.5	0	6.4 (18.1)
				Human adenovirus C, type 1	36,002	87	116,502	1,834	15.3	100.0	0.2 (0.6)	283.1 (422.3)
Nasopharyngeal swab	S3	5,139,462	4,094,568	Torque teno mini virus	2,908	0	256	>256‡	0	22.7	0	6.4 (21.8)
				Human parainfluenza virus 3	15,462	172	262,806	1,918	39.5	99.9	1.0 (1.6)	1,462.1 (1,767.0)
Nasopharyngeal swab	S4	6,124,424	5,659,554	TTV-like mini virus isolate	2,915	1	887	960	2.8	13.5	<0.1 (0.2)	23.9 (105.7)
				Human bocavirus 1	5,543	4,419	460,617	113	95.4	100.0	68.3 (94.1)	7,294.7 (3,443.1)
Nasopharyngeal swab	S5	9,152,970	5,834,446	Torque teno virus	3,741	0	56	>56‡	0	14.9	0	1.0 (3.2)
Nasopharyngeal swab	S6	10,179,884	12,064,068	Human adenovirus B, type 3A	35,264	0	1,071	>1,071‡	0	75.3	0	2.6 (2.8)
				KI polyomavirus	5,040	1,034	411,173	336	83.4	100.0	18.4 (20.5)	7,373.2 (7,123.3)
				Human rhinovirus 80	7,138	19,232	3,120,184	137	65.0	75.9	238.2 (313.2)	39,218.6 (102,545.7)
Stool	S7	3,691,496	4,104,534	Human adenovirus C, type 1	36,006	3	9,081	2,722	0.8	99.9	<0.1 (0.1)	22.9 (25.7)
				Sapovirus	7,429	33	61,982	1,689	23.3	100.0	0.4 (0.8)	740.1 (437.6)
				Human astrovirus 1	6,816	6,641	1,140,900	155	99.9	100.0	85.1 (135.6)	14,725.6 (13,370.2)
				Human polyomavirus 10	4,939	0	81	>81‡	0	70.4	0	1.6 (1.6)
Plasma	S8	10,875,448	7,088,360	Torque teno virus	3,880	0	1,817	>1,817‡	0	38.2	0	37.0 (105.0)
				Human herpesvirus 6B	161,573	38	27,523	1,111	1.1	38.4	<0.1 (0.2)	14.2 (112.9)

Pre-capture: Metagenomic shotgun sequencing without targeted sequence capture; Post-capture: Metagenomic shotgun sequencing using (ViroCap) targeted sequence capture.

* Viral reads per million sequences generated pre-capture and post-capture are statistically different (P=0.0002, Wilcoxon test).

[a] PVR (percent viral reads) fold increase: percent of post-capture viral reads divided by percent of pre-capture viral reads.

‡ PVR fold increase could not be calculated because the number of pre-capture reads was 0.