Novel Virus Detection Using the Twist Comprehensive Viral Research Panel

ABSTRACT

Infections from novel viral species and strains present a serious and recurring threat to global public health. Responding to these threats critically depends on the ability to identify and develop tests for the viral agent, which is a significant challenge given the diverse and rapidly-evolving sequence space of human viruses. The problem is further compounded by novel infections acquired from animal viruses, which may have little sequence similarity to any known human-infective species. In theory, next-generation sequencing (NGS) approaches can detect completely novel viral agents. However, their application is limited by contamination from host reads, requiring significant sequencing depth to obtain full coverage over the viral template. Here, we describe the Twist Comprehensive Viral Research Panel, a hybridization-based target enrichment solution capable of detecting highly divergent viral strains. Our results highlight the Twist Comprehensive Viral Research Panel as a versatile solution for novel virus discovery and surveillance.

INTRODUCTION

The rate of viral epidemics in recent history has risen with the expansion of the human population and its interaction with (and modification of) natural habitats (Woolhouse & Gowtage-Sequeria, 2005; Chomel, Belotto, & Meslin, 2007). The specter of novel pandemic viruses demands improved tools for the discovery and surveillance of emerging and re-emerging viral pathogens.

Detection of emerging viruses is problematic due to the high rates of mutation, recombination, and reassortment their genomes undergo (Holland et al., 1982). Common diagnostics assays such as quantitative real-time polymerase chain reaction (qRT-PCR) rely on a priori knowledge of the infectious agent for successful detection. By contrast, hypothesis-free detection methods such as next-generation sequencing (NGS) can be used to identify viral agents that diverge substantially from reference strains. Although the unbiased approach of NGS offers clear diagnostic benefits, high levels of contaminating human host nucleic acids in patient samples effectively reduces the method's analytical sensitivity, presenting a major roadblock in its application to novel virus discovery (Chiu 2015). Thus, a viral enrichment strategy would improve the analytical sensitivity of NGS for novel virus discovery.

Our solution is the Twist Comprehensive Viral Research Panel, an NGS approach using hybridization-based target enrichment that facilitates the discovery and characterization of emerging viral pathogens. We applied this capture system to synthetic controls derived from viral strains that emerged after the design of the Twist Comprehensive Viral Research Panel, including a strain of Rousettus bat coronavirus GCCDC1 (RoBat-CoV GCCDC1) discovered recently in Singapore (Paskey et al., 2020) and strains of novel H1N1 influenza isolated from a June 2020 outbreak in pigs (Sun et al., 2020). We also tested the enrichment efficiency of the panel by generating synthetic controls with known levels of single-base substitutions. In this application note, we demonstrate the:

1. High sensitivity detection (lower than 1,000 viral copies) of known viral species after ribosomal depletion.
2. Enrichment and complete characterization of novel viral sequences obtained from novel RoBat-CoV GCCDC1 and H1N1 swine influenza.
3. Mismatch tolerance of the hybrid capture system using a set of synthetic, engineered H1N1 hemagglutinin (HA) fragments.
4. Identification of a novel H1N1 swine influenza sequence in a simulated infection without *a priori* knowledge of the strain, using the One Codex platform.

RESULTS

The Twist Comprehensive Viral Research Panel contains 1,052,421 unique, 120 base pair (bp) biotinylated probes for hybridization capture of viral sequences. In designing the Twist Comprehensive Viral Research Panel, we incorporated diversity from 3,153 human-infective viruses and 15,488 different strains (including single-stranded [ss] DNA, double-stranded [ds] DNA, dsRNA, and ssRNA viruses) to enable broad-based enrichment of viral sequences for detection and characterization with NGS. This exploratory panel's performance was evaluated against natural and artificial sources of sequence variation to determine its ability to detect emerging viruses.

The Twist Comprehensive Viral Research Panel was first tested on two RNA controls derived from the reference sequences of SARS-CoV-2 (PN 102024) and human rhinovirus 89 (PN 103006) as a coinfected sample. After rRNA Depletion using the NEBNext rRNA Depletion Kit v2 (NEB #E7400, #E7405), TruSeq libraries at viral titers of 1,000 and 1,000,000 copies in human carrier RNA background were generated using the Twist Total Nucleic Acids Library Preparation Kit for Viral Pathogen Detection and Characterization protocol. Hybrid capture was then performed using the Twist Comprehensive Viral Research Panel and the Twist Standard Target Enrichment workflow, as described in the Materials & Methods. At high titers (1,000,000 copies), >99.9% coverage over both viral templates was obtained, with more than 15% of sequencing reads mapping to the viral template. At low titers (1,000 copies), over 90% of the viral template sequence could still be recovered, with approximately 0.1% of reads mapping

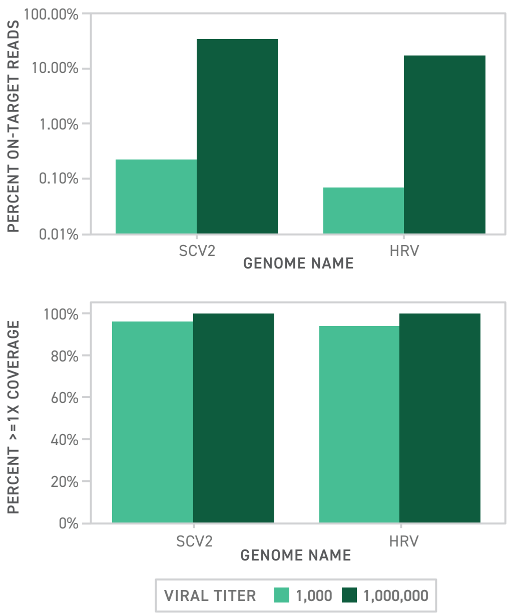


Figure 1: Percent on-target rate (top) and coverage at >=1x (bottom) for synthetic SARS-Cov-2 (SCV2) and human rhinovirus (HRV) RNA controls captured after ribosomal depletion as a coinfected sample. Captures are shown at viral titers of 1,000 and 1,000,000.

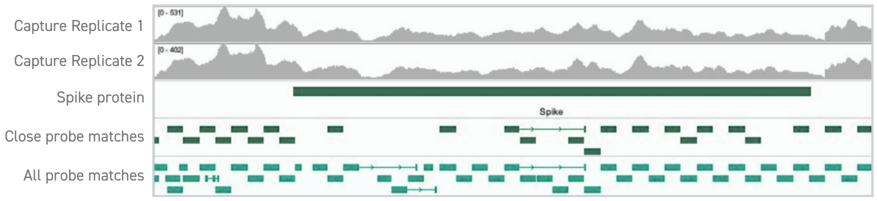


Figure 2: Genome browser views illustrating read density surrounding the spike coding region of the RoBat-CoV GCCDC1 genome. "Close probe matches" contained at least 110 nucleotide matches and fewer than five mismatches.

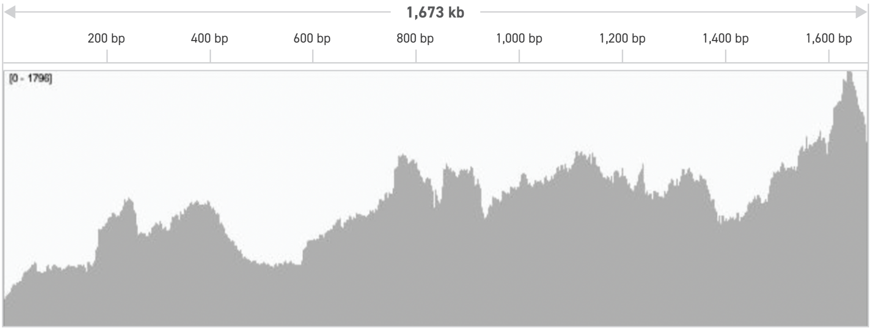


Figure 3: Genome browser views of read density from an HA variant (Genbank accession number: MN416597) derived from H1N1 influenza. At most bases, the HA variant was covered to high depth (100% of bases >200x coverage).

to the viral template (**Figure 1**). These results highlight the ability of the Twist Comprehensive Viral Panel to detect viruses with high sensitivity in spite of the large target space, which might be expected to produce large amounts of off-target capture.

Having established efficient capture of known sequences, the ability of the Twist Comprehensive Viral Research Panel to capture viral sequences not explicitly included in the panel design was tested. To this end, the complete genome of the recently described RoBat-CoV GCCDC1 Singapore strain (Paskey et al. 2020; Genbank accession number: MT350598), a novel betacoronavirus infecting rousettus bats, was synthesized. Capture was performed as described in the Materials & Methods, with 1,000,000 copies of the viral genome and without ribosomal depletion.

Figure 2 illustrates high efficiency capture of the spike coding region of the RoBat-CoV GCCDC1 Singapore genome, without prior ribosomal RNA depletion. This is particularly noteworthy because this region lacks complete probe coverage in the panel design (**Figure 2**). In total, almost the entire genome (99.8%) was covered to at least 1x depth. These results can be attributed to a panel design that effectively targets a highly diverse space of existing coronaviruses, a hybrid capture system that can tolerate large stretches of mismatches, and the ability of capture probes to capture fragments extending beyond the target space of the probe.

Next, the ability of the Twist Comprehensive Viral Research Panel to capture novel divergent strains of species that are known to infect humans was tested using recently described H1N1 swine influenza strains that were shown to have human infective potential. Recent surveillance efforts identified novel human-infective H1N1 strains related to the 2009 pandemic virus in swine workers (Sun et al. 2020). From sequences described by Sun and colleagues (2020), we synthesized four HA and neuraminidase (NA) segments each. The HA and NA segments were selected because they display high sequence divergence across H1N1 viruses (Kosik et al, 2019). Following hybridization capture with the Twist Comprehensive Viral Research Panel, all eight sequences were covered to 1x or higher at 100% genome coverage (**Table 1**). **Figure 3** shows the read density for an HA segment from influenza strain A/swine/Hebei/0116/2017(H1N1) (GenBank Accession number: MN416597). These data highlight the ability of the Twist Comprehensive Viral Research Panel to capture recently evolved viral sequences fully.

To better define the mismatch tolerance of the Twist Comprehensive Viral Research Panel, a series of HA segments were synthesized to contain predefined levels of artificial variation ranging from 5% to 30% from the HA segment of pandemic influenza strain A/California/07/2009(H1N1) (Genbank accession number: NC_026433). As shown in **Figure 4**, full coverage at 1x with 1M total mapped reads was obtained for HA segments containing up to 10% variation from the reference sequence. Total capture decreased in samples with greater than 10% mismatches. Nevertheless, 70% and 55% of the template in HA segments containing 15% and 20% mismatches were captured, respectively, at a depth of 1x with 1M total mapped reads

(**Figure 4**). Combined with the RoBat-CoV GCCDC1 Singapore capture data and the 2009 H1N1 strain data, these findings indicate that the Twist Comprehensive Viral Research Panel possesses the high mismatch tolerance necessary to fully capture divergent viral sequences.

The above figures highlight the performance of the Twist Comprehensive Viral Research Panel in capturing known human-infective viruses, novel viral strains with zoonotic potential, and heavily mutated strains of known viruses. In all of these cases, the analysis was performed by straightforward sequence alignments to the exact strains used, to avoid any bias from the bioinformatics analysis that might confound our analysis of capture efficiency. However, as this approach relies on prior knowledge of the viral sequence, it is not feasible for analyzing sequencing results for isolates containing truly novel or unknown viruses. To test whether these sequences could be detected without *a priori* knowledge, we additionally performed an analysis of one of the novel Swine Flu strains through k-mer searches to a taxonomic database using the One Codex platform. In **Figure 5**, the results of this analysis are shown—in spite of the novel sequence content, One Codex unambiguously identified the sample as Influenza A/H1N1, and even identified that these sequences likely derived from a Swine flu strain.

SUMMARY

NGS enables the unbiased identification of infectious species, including unknown species and highly divergent strains; however, the high prevalence of human host nucleic acids in diagnostic samples reduces the analytical sensitivity of this approach. When combined with Twist Target Enrichment workflow and NGS, the Twist Comprehensive Viral Research Panel provides a versatile solution for broad-based detection and surveillance of uncharacterized viruses, enabling viral sequence detection beyond those incorporated explicitly in the panel design.

MATERIAL & METHODS

TruSeq libraries were generated using Twist synthetic RNA viral controls and the Twist Total Nucleic Acids Library Preparation Kit for Viral Pathogen Detection and Characterization protocol. Briefly, RNA viral controls were spiked into a background of 50 ng human reference RNA (Agilent). The resulting samples were converted to cDNA using ProtoScript II First Strand cDNA Synthesis Kit (E6560S) and New England Biolab's Random Primer 6 (S1230S). The NEBNext Ultra II Non-Directional RNA Second Strand Synthesis kit (E6111S) was subsequently used to convert single-stranded cDNA to dsDNA. Illumina TruSeq-compatible libraries were then generated using the Twist Library Preparation Kit with Enzymatic Fragmentation (PN 101059 and 100401) and Unique Dual Indices (UDI) (PN 101307). Libraries were ultimately generated at a viral titer of 10^6.

Hybridization capture was performed using the Twist Comprehensive Viral Research Panel (PNs 103545, 103547, 103548)

SYNTHETIC VIRUS	HA ACCESSION NUMBER	NA ACCESSION NUMBER	PERCENT >=1X COVERAGE ON HA	PERCENT >=1X COVERAGE ON NA
1	MN416597	MN416693	100%	100%
2	KP735714	MN416700	100%	100%
3	MN416610	MN416718	100%	100%
4	MN416620	MN416735	100%	100%

Table 1: Detection of novel NA and HA variants from novel H1N1 viruses (Sun et al., 2020).

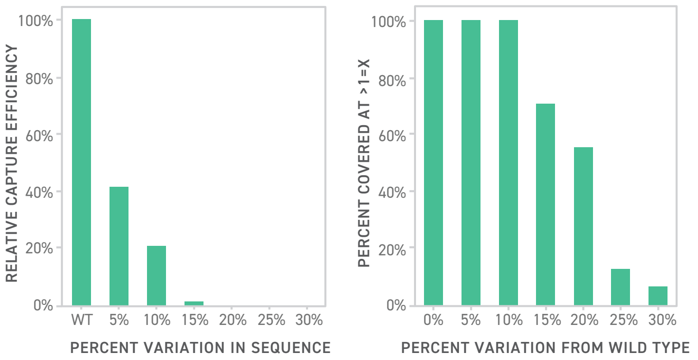
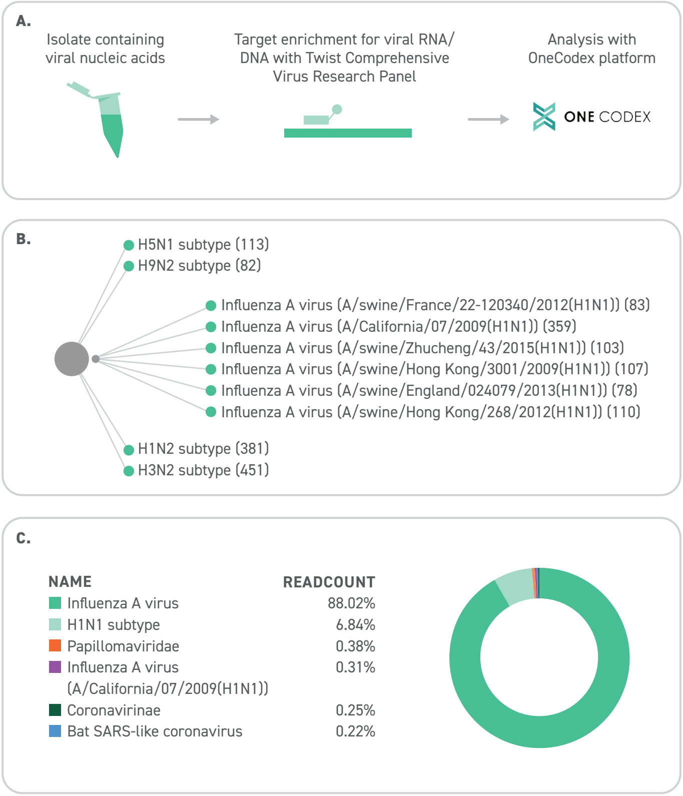


Figure 4: Percentage of variation from wild type affects coverage and capture efficiency.


A.
Isolate containing viral nucleic acids
Target enrichment for viral RNA/ DNA with Twist Comprehensive Virus Research Panel
Analysis with OneCodex platform
ONE CODEX
B.
H5N1 subtype (113)
H9N2 subtype (82)
Influenza A virus (A/swine/France/22-120340/2012(H1N1)) (83)
Influenza A virus (A/California/07/2009(H1N1)) (359)
Influenza A virus (A/swine/Zhucheng/43/2015(H1N1)) (103)
Influenza A virus (A/swine/Hong Kong/3001/2009(H1N1)) (107)
Influenza A virus (A/swine/England/024079/2013(H1N1)) (78)
Influenza A virus (A/swine/Hong Kong/268/2012(H1N1)) (110)
H1N2 subtype (381)
H3N2 subtype (451)
C.

NAME	READCOUNT
🟩 Influenza A virus	88.02%
🟩 H1N1 subtype	6.84%
🟧 Papillomaviridae	0.38%
🟪 Influenza A virus (A/California/07/2009(H1N1))	0.31%
🟩 Coronavirinae	0.25%
🟦 Bat SARS-like coronavirus	0.22%

Figure 5: (A) End-to-end workflow for target enrichment and analysis of viral samples with the Twist Comprehensive Viral Research Panel and the One Codex platform. (B) Taxonomic tree display from One Codex analysis of enriched novel Swine Flu strains. (C) Summary of all species detected by One Codex platform.

and the Twist Standard Target Enrichment workflow. 500 ng of library was used in each 16-hour hybridization capture reaction. Following enrichment, libraries were sequenced with 2x75 bp paired-end reads on the Illumina NextSeq platform, using a NextSeq500/550 High Output kit. Alignment was performed with BWA against a custom genome index comprising the human genome (build hg38) concatenated with reference sequences for each virus in the panel. All data were downsampled to 1M mapped reads per sample unless otherwise noted.

REFERENCES

Chiu CY. (2013). Viral pathogen discovery. Current opinion in microbiology, 16(4), 468–478. https://doi.org/10.1016/j.mib.2013.05.001

Chomel BB, Belotto A, & Meslin FX. (2007). Wildlife, exotic pets, and emerging zoonoses. Emerging infectious diseases, 13(1), 6–11. https://doi.org/10.3201/eid1301.060480

Holland J et al. (1982). Rapid evolution of RNA genomes. Science (New York, N.Y.), 215(4540), 1577–1585. https://doi.org/10.1126/science.7041255

Kosik I & Yewdell JW. (2019). Influenza Hemagglutinin and Neuraminidase: Yin Yang Proteins Coevolving to Thwart Immunity. Viruses, 11(4), 346. https://doi.org/10.3390/v11040346

Lam TT et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature, 583(7815), 282–285. https://doi.org/10.1038/s41586-020-2169-0

Mena I et al. (2016). Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. eLife, 5, e16777. https://doi.org/10.7554/eLife.16777

Paskey AC et al. (2020). Detection of Recombinant Rousettus Bat Coronavirus GCCDC1 in Lesser Dawn Bats (Eonycteris spelaea) in Singapore. Viruses, 12(5), 539. https://doi.org/10.3390/v12050539

Sun H et al. (2020). Prevalent Eurasian avian-like H1N1 swine influenza virus with 2009 pandemic viral genes facilitating human infection. Proceedings of the National Academy of Sciences of the United States of America, 117(29), 17204–17210. https://doi.org/10.1073/pnas.1921186117

Woolhouse ME & Gowtage-Sequeria S. (2005). Host range and emerging and reemerging pathogens. Emerging infectious diseases, 11(12), 1842–1847. https://doi.org/10.3201/eid1112.050997

These products are for research use only, and subject to additional use restrictions as set forth in Twist's Supply Terms and Conditions: www.twistbioscience.com/supply-terms-and-conditions