A Target Enrichment Approach for Identifying Viral Infections

Michael Bocek, Leonardo Arbiza, Kristin Butcher, Siyuan Chen, Brenton Graham, David Lin, Rebecca Nugent, Christina Thompson, Zachary Zwirko

TWIST BIOSCIENCE

1. Abstract

Nucleic acid tests are a highly sensitive and informative approach for studying viral infections, and can be rapidly adapted to detect novel pathogens such as SARS-CoV-2, the agent that causes the ongoing COVID-19 pandemic. While amplification-based tests like RT-PCR provide rapid and sensitive detection, they do not provide a specific sequence for the viral agent involved. Additionally, these tests are generally targeted towards a single pathogenic agent. Detection by next generation sequencing (NGS) approaches is an attractive alternative that addresses these issues. However, in many samples the viral genome is present at prohibitively low levels compared to the host genome, requiring deep sequencing (Chiu 2015). This raises the cost of sequencing and makes it difficult to achieve sufficient depth to confidently detect viral infection and characterize the viral sequence. Here, we present an approach that combines target enrichment (TE) by hybridization to DNA probes to enrich for viral nucleic acids, followed by NGS.

As viral genomes can be present as either RNA or DNA, and can be single- or double-stranded, we developed a library preparation method that operates efficiently on all four types of nucleic acids. Using Twist's target enrichment panels, including the Twist SARS-CoV-2 Research Panel and the Twist Respiratory Virus Research Panel, to enrich content from these libraries, we show enrichment of synthetic viral standards from all four classes of nucleic acids against a human RNA reference background. Our approach is highly sensitive, detecting as few as ten copies of the SARS-CoV-2 genome with over 100,000-fold enrichment compared to human RNA using the Twist SARS-CoV-2 Research Panel. With 99.8% of the SARS-CoV-2 viral template covered by reads to 100x depth at 1000 viral copies, we show that enrichment allows for high-quality base calls of different viral strains. Extending our results beyond single infections, we simultaneously detect multiple viral agents with excellent coverage in a model of co-infection with the Twist Respiratory Virus Research Panel. Finally, we demonstrate that hybridization time can be reduced to 30 minutes using Twist's Fast Hybridization system, and that 8 samples can be multiplexed into a single capture with comparable efficiency. Our results establish Twist's target enrichment systems as a sensitive and informative approach for detecting viral agents.

1. Nucleic acid agnostic library preparation

Unlike organismal genomes, which are uniformly stored as either circular or linear double stranded DNA (dsDNA), viral genomes exist in a variety of nucleic acid topologies. Viruses, including clinically important human-infective species, can store their genomes as any combination of single-stranded or doubled-stranded DNA or RNA. In order to ensure that our capture system could be adapted to the entire range of human-infective viruses, we developed a library synthesis strategy that could accommodate ssDNA, dsDNA, ssRNA and dsRNA.

In tests of our capture system using the same sequence synthesized as ssRNA, ssDNA, dsRNA or dsDNA, we find that all 4 types of nucleic acid can be effectively captured **(Figure 1)**, although the capture efficiency does depend on the type of nucleic acid used. Nonetheless, we obtained nearly complete coverage of each synthetic template (with the exception of the non-sequenceable poly-A tail of the virus) **(Figure 2)**. Base calls were highly accurate, across each template, with only two sites among all of the templates rising to the 20% cutoff that we used for minor allele fraction. In each case, these sites had low coverage (<20x) which suggests that these errors represent random sequencing error rather than a specific bias on the part of the capture system.

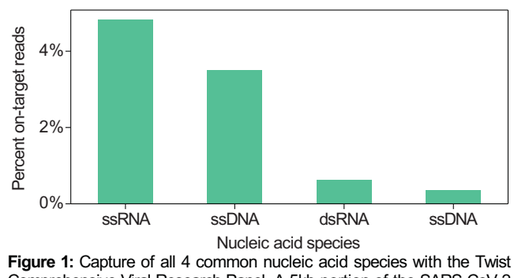

Figure 1: Capture of all 4 common nucleic acid species with the Twist Comprehensive Viral Research Panel. A 5kb portion of the SARS-CoV-2 genome was synthesized as either double- or single-stranded RNA or DNA, spiked in at 10^6 titer into a background of 50ng of human reference RNA. After library preparation and capture with the Twist Comprehensive Viral Research Panel, we assessed the on-target rate for each nucleic acid species as a measure of the capture efficiency.

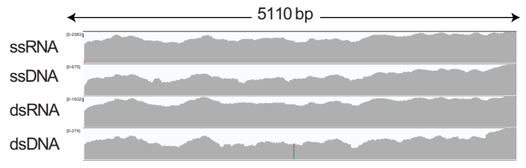

Figure 2: Genome browser view (log scale) of the above capture showing coverage across the synthetic template. SNPs are highlighted with color differences.

2. High efficiency capture of SARS-CoV-2

In the ongoing COVID-19 pandemic, next-generation sequencing has proved to be an invaluable tool for epidemiologists studying the spread of the disease. By characterizing and tracking variants in the SARS-CoV-2 genome, researchers can gain deep insight into the global patterns of transmission, unambiguously identify transmission events, and even infer aspects of virus-host interactions (for example, by identifying co-infections between different strains within a single individual.) In order to discover new strains, and unambiguously distinguish between every existing strain of SARS-CoV-2, a complete sequence must be obtained for the virus. However, in many cases this is challenging, as the viral RNA genome is present at very low copy numbers compared to host RNAs.

To allow for full characterization of the SARS-Cov-2 genome even at low copy numbers, we developed a target-enrichment system that allowed for highly sensitive targeted sequencing of SARS-CoV-2. By mixing synthetic viral RNA into human reference RNA at defined copy numbers, we found the capture system could detect <10 copies of the virus, and could completely characterize the viral genome with as few as 1000 copies. While total enrichment depended on titer, at low titers we obtained nearly million-fold enrichment above the expected input levels (**Figure 3**). Our synthesized strain contained three single-base substitutions compared to the viral reference genome, as well as one 10-base pair deletion. At 1,000 copies, all of these events were clearly detected (**Figure 4**).

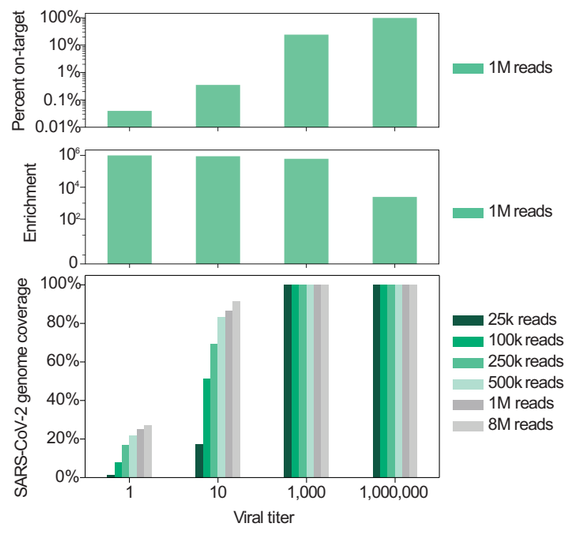

Figure 3: Percent-on-target (top), enrichment (center) and total genome coverage (bottom) for the SARS-CoV-2 template at different viral titers. Total genome coverage shown at numbers of mapped reads ranging from 25 thousand to 8 million. All other metrics calculated at 1 million total mapped reads.

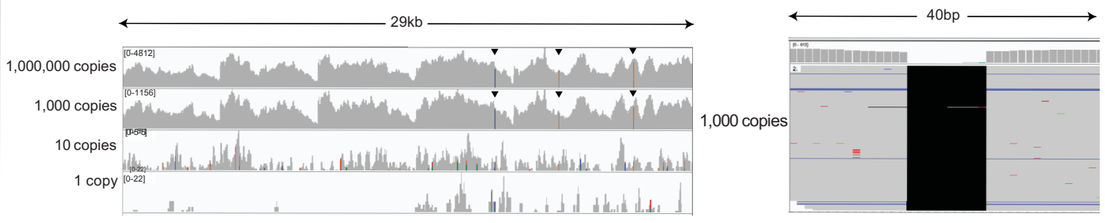

Figure 4: (Left) Genome browser view of SARS-CoV-2 capture at different viral titers. Differences from the reference template occurring at >20% MAF are highlighted in coverage plots. The locations of 3 expected SNPs in the synthetic RNA for strain MT007544.1 are marked with triangles. (Right) Genome browser view of total coverage (top) and individual alignments (bottom) for the 1,000 copy capture at a 10bp deletion in strain MT007544.1 compared to the reference. Alignments containing the deletion marked in black.

Materials and Methods

Twist synthetic RNA and DNA viral controls were spiked into a background of 50ng human reference RNA (Agilent) with viral copy numbers ranging from 100 to 1,000,000 copies per sample. Co-infections were simulated by spiking multiple synthetic viral controls into one sample. A negative control consisting solely of human reference RNA was processed in parallel. The samples were then converted to single-stranded cDNA through random priming using NEB's Random Primer 6 (S1230S), ProtoScript II First Strand cDNA Synthesis Kit (E6560S). Single-stranded cDNA was then converted to dsDNA using the NEBNext Ultra II Non-Directional RNA Second Strand Synthesis kit (E6111S). The samples were converted to Illumina TruSeq-compatible libraries using Twist Library Preparation Kit using Enzymatic Fragmentation (PN 101059 and 100401) and Unique Dual Indices (UDI) (PN 101307).

Enrichment was performed with the indicated capture panel in single-plex capture reactions using a 16-hour hybridization. For multiplexing experiments, 8 libraries were pooled (187.5 ng each) for a total of 1500ng. Enriched libraries were sequenced with 2x75bp 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.

3. Human respiratory viruses

Viruses cause hundreds of millions of respiratory system infections in the United States every year, but it is not always easy to distinguish the many potential causative agents. Particularly in the midst of the ongoing COVID-19 pandemic, it is important to have tests that can distinguish between multiple viruses that may ultimately cause similar symptoms.

To address this, we developed a small targeted panel that covered 31 viruses that were commonly known to cause serious respiratory infections (**Figure 5**). The design was performed to target the reference genomes of strains known to cause symptoms of respiratory infections in immunocompetent adults. For families of viruses with exceptionally high diversity, like influenza and rhinoviruses, probes were designed to target a metagenome using CATCH (Metsky *et al* 2019). Using the Twist Respiratory Virus Research Panel, we captured 16 synthetic standards from different genome families targeted by the virus. At 10^6 titer in a background of 50ng human RNA, we obtained higher than 70 percent on-target reads and over 99% of each genome covered to at least 30x depth (**Figure 6**). Capture was also highly sensitive at low viral titers, with substantial capture even at low (100 copy) titer. Additionally, we found the system was amenable to multiplexed capture, substantially reducing the effort and preparation time for large numbers of samples (**Figure 7**). Finally, we profiled the ability of the panel to simultaneously detect two different viral species in models of co-infections, showing nearly complete coverage over both species represented (**Figure 8**).

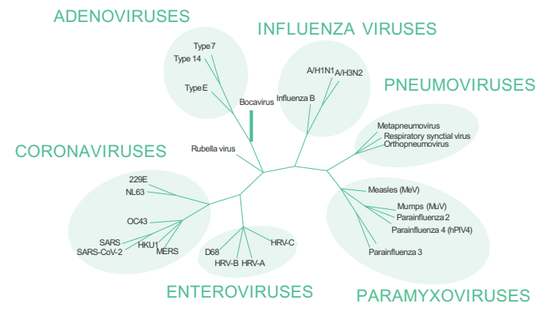

Figure 5: Taxanomic tree of all species covered in capture panel

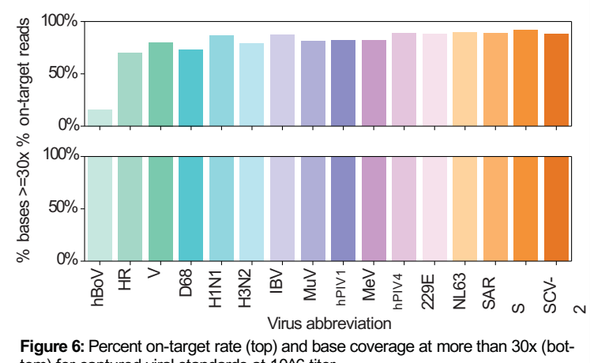

Figure 6: Percent on-target rate (top) and base coverage at more than 30x (bottom) for captured viral standards at 10^6 titer.

Figure 8: Genome browser view of simultaneous capture of a SARS-CoV-2 and rhinovirus simulated co-infection. Both strains were spiked into 50ng of human reference RNA at 10^4 titer. Coverage over each genome is shown.

Abbreviations:
hBOV: Human bocavirus, **HRV**: Human rhinovirus 89, **D68**: Enterovirus D68, **H1N1**: Influenza A/H1N1, **H3N2**: Influenza A/H3N2, **IBV**: Influenza B, **MuV**: Mumps virus, **hPIV1**: Human parainfluenzavirus 1, **MeV**: Measles virus, **hPIV4**: Human parainfluenzavirus 4, **229E**: Coronavirus 2293, **NL63**: Coronavirus NL63, **SARS**: SARS coronavirus, **SCV-2**: SARS-CoV-2, **MERS**: MERS coronavirus, **OC43**: Coronavirus OC43

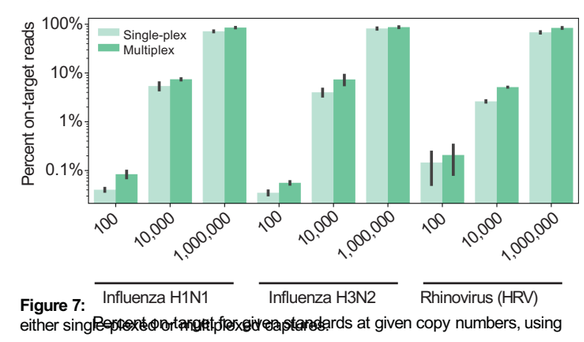

Figure 7: either single-plex or multiplexed captures.

4. Emerging threats and novel pathogens

A major benefit of NGS solutions for infectious disease is the ability to characterize novel pathogens without any prior knowledge of their sequence. However, in many cases this characterization is complicated by contamination from host-reads. This is particularly important for low-titer samples.

To address these situations, we have developed the Twist Comprehensive Viral Research Panel, a target enrichment solution that broadly targets the human virome, including closely-related viruses that infect non-human species. To test the effectiveness of this panel on novel genomes, we performed captures on three sets of synthetic viral standards. First, we tested the system's tolerance to random mismatches by engineering random mutations into the reference Influenza H1N1 (2009) hemagglutinin (HA) segment at fixed percentages. Even at 15% sequence variation, we recovered over half of the template (**Figure 9**). We next evaluated the panel's performance on a novel strain, capturing two segments from an outbreak of novel swine flu with pandemic potential recently isolated in China (Sun *et al* 2020). Even though this strain was not in the design space of the panel, we recovered entire sequences for each segment with excellent coverage (**Figure 10**). Finally, the coronaviruses responsible for SARS, MERS and COVID-19 are all thought to have emerged from animal betacoronaviruses. We thus captured the sequence of a novel bat betacoronavirus (Rousettus Bat Coronavirus GCCDC1) that was published after the panel design was finalized (Paskey *et al* 2020). Although portions of the virus sequence were not covered by probes with high homology, we nonetheless captured over 99% of the novel sequence with this panel, demonstrating its utility for zoonotic agents (**Figure 11**).

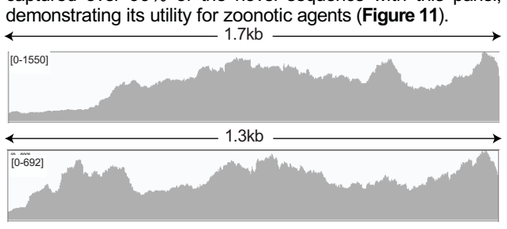

Figure 10: Capture of sequences derived from a novel swine flu with pandemic potential discovered in China in June 2020. Coverage plots show a representative NA segment (top) and HA segment (bottom).

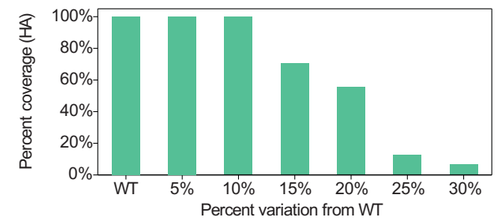

Figure 9: Capture of influenza H1N1 strain containing random variants in HA segment. After normalizing to the baseline capture of the WT HA segment compared to all other WT segments, we assessed the total amount of capture at each given level of sequence variation (top). We also measured the fraction of each template with at least 1x coverage at each level of variation (bottom.)

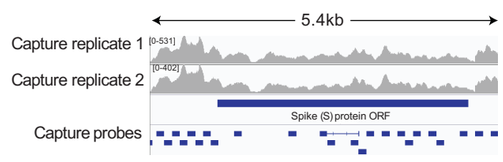

Figure 11: Capture of a novel strain of an animal betacoronavirus (Rousettus Bat Coronavirus GCCDC1), published in May 2020. A 5.4 kb portion of the sequence containing the rapidly mutating spike (S) protein is shown, along with locations of probes from the Twist Comprehensive Viral Research Panel with substantial homology

References

Chiu CY. Viral pathogen discovery. *Curr Opin Microbiol*. 2013 Aug;16(4):468-78. doi: 10.1016/j.mib.2013.05.001. Epub 2013 May 29.
Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, *et al.* Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nat Biotechnol*. 2019 Feb;37(2):160-168. doi: 10.1038/s41587-018-0006-x. Epub 2019 Feb 4. PMID: 30718881; PMCID: PMC6587591.
Paskey AC, Ng JHJ, Rice GK, Chia WN, Philipson CW, *et al.* Detection of Recombinant Rousettus Bat Coronavirus GCCDC1 in Lesser Dawn Bats (Eonycteris spelaea) in Singapore. *Viruses*. 2020 May 14;12(5):539. doi: 10.3390/v12050539.
Sun H, Xiao Y, Liu J, Wang D, Li F, *et al.* Prevalent Eurasian avian-like H1N1 swine influenza virus with 2009 pandemic viral genes facilitating human infection. *Proc Natl Acad Sci U S A*. 2020 Jul 21;117(29):17204-17210. doi: 10.1073/pnas.1921186117. Epub 2020 Jun 29.

Financial disclosure

No funding to disclose the following