Tiling Across Viral Genomes with 300mer Oligonucleotides to Investigate Immune Escape

Learn how 300mer oligos can be used to tile across genomes to interrogate otherwise hidden genetic functionality

ABSTRACT

Viruses exploit high-dimensional RNA structures to facilitate viral replication and impede host restriction. Here, a novel next-generation sequencing (NGS)-based screening workflow, called Fate-seq, was used to define and characterize high-dimensional RNA structures from the severe acute respiratory syndrome coronavirus (SARS-CoV) genome in an unbiased, high-throughput, and comprehensive manner. The Fate-seq workflow leverages 300mer Twist Oligos to simplify library preparation and capture genomic sequences long enough to probe RNA secondary structures.

INTRODUCTION

Viral genomes possess and produce high-dimensional RNA structures (e.g., stem-loops and pseudoknots) that facilitate viral replication and gene expression. High-dimensional RNA structures arise when an RNA molecule base-pairs with itself. Simple structures like stem-loops contain a base-paired "stem" region and an unpaired "loop" structure. More complex structures like pseudoknots form when a separate region of the same RNA molecules base-pair with the unpaired nucleotides of a stem-loop structure. Pseudoknots in particular enable a wide variety of viral replication and gene expression tasks through interactions with viral transcriptional and host translational machinery, respectively (Brierley, Pennell, & Gilbert, 2007).

Viruses also exploit high-dimensional RNAs to interfere with host restriction factors. This is exemplified by the production of subgenomic RNAs during flavivirus replication. Flavivirus genomes contain pseudoknot structures that prevent cellular exoribonucleases from fully degrading them (Pijlman et al., 2008). The resulting partially-degraded genomic fragments, called subgenomic RNA, interfere with the host's innate immune response by binding to and blocking the function of host proteins that activate interferon production (Manokaran et al., 2015). This mechanism is thought to underlie the enhanced fitness of the dengue virus strain that caused a severe 1994 outbreak in Puerto Rico (Manokaran et al., 2015).

RNA secondary structure can be predicted using computational tools. However, revealing the functional importance of higher-dimentional RNA structures requires validation and interrogation through high-throughput genome scanning workflows.

Researchers at the University of Tokyo have devised a novel strategy for scanning entire viral genomes for high-dimensional RNA structures, leveraging Oligo Pools from Twist Bioscience (Wakida et al, 2020). The NGS-based screen, called Fate-seq, relies on the stability conferred by RNA secondary structure to determine the fate of viral genome fragments in human cells: do they get degraded or not? Fate-seq capitalizes on the spacious length limits of 300mer Twist Oligo Pools to provide a straightforward workflow for capturing high-dimensional viral RNA sequences. In short, in any genome tiling application, the longer the oligo, the greater the information density per-oligo for the elucidation of genomic function. Moreover, the vanishingly low error rate of Twist Oligo synthesis supports the workflow by guaranteeing high-quality Fate-seq hits.

In this application note, Fate-seq was applied to identify high-dimensional RNAs in the SARS-CoV genome. Fate-seq led to the discovery of multiple stem-loop structures that were also present in SARS-CoV-2, the virus that causes COVID-19. Comparative analysis also identified a stem-loop structure unique to SARS-CoV-2. These findings support the use of Fate-seq to screen for viral RNA structures that may act as virulence factors.

WORKFLOW

Fate-seq follows five major steps (**Figure 1**):

1. **Oligo design:** Viral genomes are selected for screening and fragmented in silico to produce overlapping fragments that tile the genomic sequence (**Figure 2**). This information is used to design 300 nt oligos. As shown in **Figure 2**, each 300 nt fragment contains a 260 nt viral genomic sequence flanked by 20 nt primer sequences. Primer sequences enable PCR amplification of the genome tiling library after synthesis.
2. **Vector library construction:** Fragments are assembled into a plasmid library such that each fragment is placed downstream of a fluorescent reporter gene.
3. ***In vitro* transcription:** The vector library is subsequently transformed in *E. coli*, purified, linearized, and *in vitro* transcribed to produce an mRNA library.
4. **Electroporation:** The *in vitro* transcribed mRNA library is electroporated into a cell line (e.g., HeLa). A 0 hour sample is collected as a control. After a pre-defined period of time, the remaining mRNAs (i.e., those not yet degraded) are extracted for analysis.
5. **NGS analysis:** Extracted mRNAs are sequenced by NGS. Reads are aligned to reference sequences and counted. Samples taken at 0 hours should show uniform coverage of all sequences, whereas samples collected at the end of the experiment only produce high coverage of stable RNA sequences.

RESULTS

An initial Fate-seq screen was performed to scan a total of 5,924 double-stranded genomic sequences corresponding to 11,848 forward or reverse strand sequences from 26 viruses (10 DNA viruses and 16 RNA viruses; Wakida et al., 2020). Each oligo was mapped across each virus genome with a 100 nt sliding window. Following six hours *in vivo* incubation, NGS sequencing of extracted mRNAs identified 625 significantly overrepresented viral sequences.

Sequences from the SARS-CoV genome were among the most highly enriched fragments identified. To further probe high-dimensional RNA structures related to the SARS-CoV genome, a SARS-CoV Fate-Seq library containing 296 fragments spanning its 29,751 base-long genome was also generated. The resulting library was electroporated into HeLa cells. Six hours later, the remaining RNA sequences were recovered and sequenced by next-generation sequencing. Of the 296 fragments screened, 21 were significantly enriched at the end of the screen (**Figure 3A**).

A conservation ratio was calculated to determine which of these 21 enriched sequences were evolutionarily conserved across 37 viruses in the *Coronaviridae* family. **Figure 3B** shows high conservation across *Coronaviridae* in the central region (7,000 to 21,500 nt) of the SARS-CoV genome. Two Fate-seq enriched sequences (16,901–17,160 nt and 20,901–21,160 nt) mapped to this central region, demonstrating >55% sequence conservation. These two sequences—called COV001 and COV002 hereafter—exhibited very high conservation (92.7% and 84.6%, respectively) between SARS-CoV and SARS-CoV-2.

The secondary structure of COV001 and COV002 was investigated using the CentroidFold tool. This analysis revealed a notable difference in the structure of COV001 between SARS-CoV and SARS-CoV-2: whereas the latter contained a stem-loop (#1), the former did not (**Figure 4A**). COV002 contained three stem-loop regions (#2–4) but did not differ between SARS-CoV and SARS-CoV-2 (**Figure 4B**). As shown in **Figure 4C**, the SARS-CoV-2 stem-loop #1 structure could be induced in SARS-CoV COV001 by introducing two mutations (C17,114A and C17,144T). Thus, these two residues appear critical for the formation of stem-loop #1 in SARS-CoV-2. These residues are present in >99% of 1,116 sampled SARS-CoV-2 genomic sequences from human patients that existed at the time of the study's publication. Together, these results show that genome tiling and subsequent sequencing with the Fate-Seq workflow identifies highly stable RNA structures ready for further study into host immune escape.

CONCLUSIONS

High-dimensional RNA structures allow viruses to bypass host defense mechanisms. Fate-seq enables high-throughput discovery of viral genomic sequences that give rise to these high-dimensional RNA structures. Of the 21 SARS-CoV genomic sequences identified by Fate-seq here, two were selected for further study because they exhibited moderate (55%) sequence conservation across the *Coronaviridae* family and very high

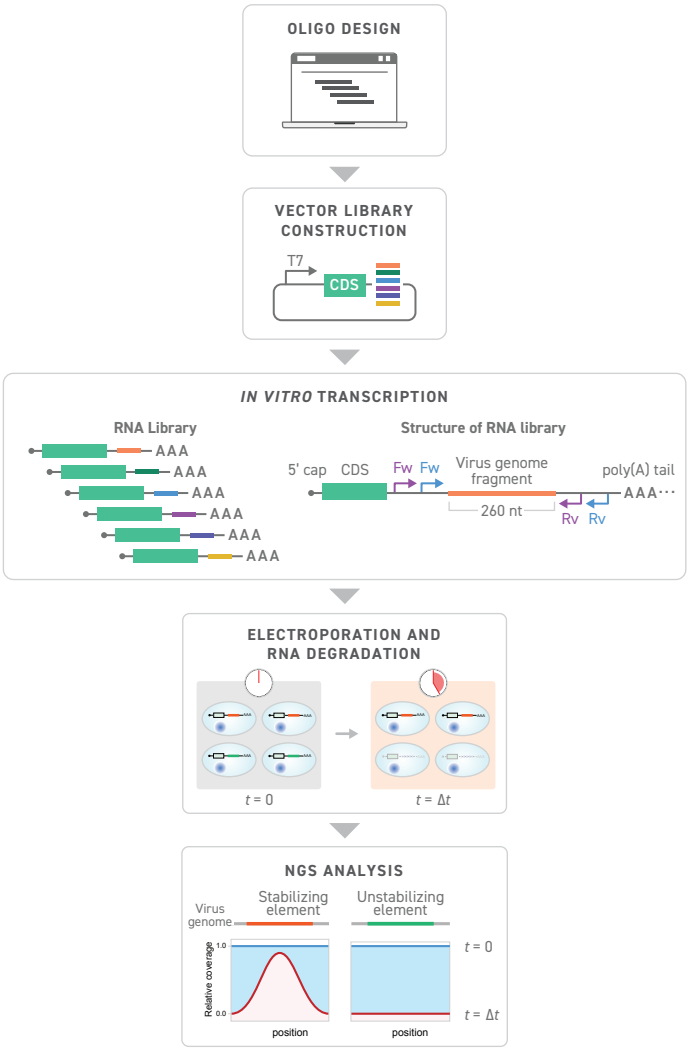

Figure 1. Fate-seq workflow. Schematic diagram of Fate-seq, a novel high-throughput workflow for discovering stable viral RNA sequences.

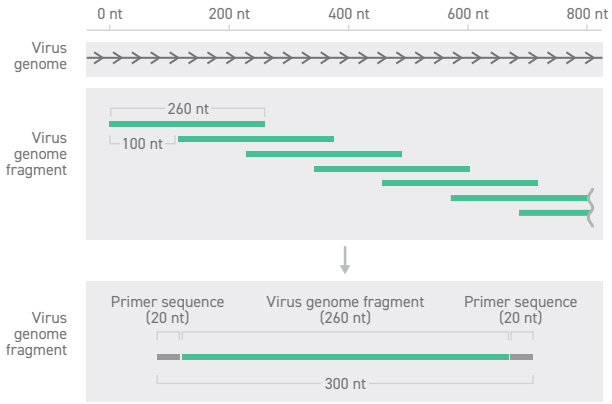

Figure 2. Fate-seq oligo design. Fate-seq libraries are designed *in silico* by fragmenting a viral genome into 260 nt sequences such that each consecutive fragment overlaps its neighboring fragments by 100 nt. Primer sequences (20 nt) are then appended to each end, resulting in sequences 300 nt in length.

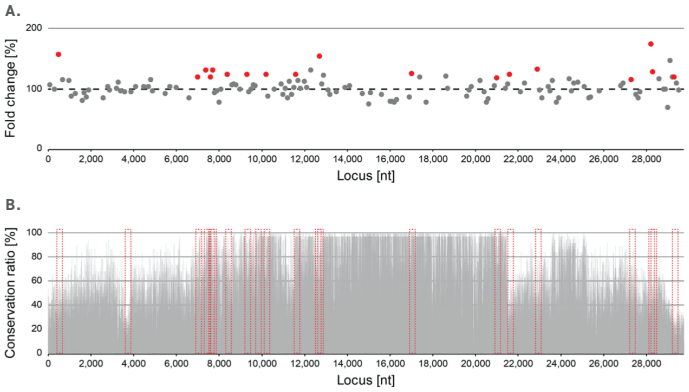

Figure 3. Discovery of stable SARS-CoV-2 RNA sequences. (**A**) Each screened genomic fragment is represented by a dot and is mapped to its position in the genome (x-axis). Red dots denote genomic sequences that were significantly more abundant in 6 hour samples (adjusted $p < 0.05$). (**B**) Each fragment was mapped to the conservation rate among *Coronaviridae* for each base in the SARS-CoV genome (gray). Significantly enriched genomic fragments from (A) are outlined in red.

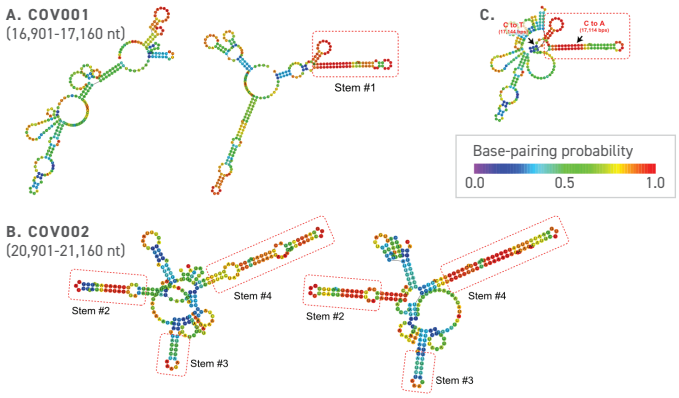

Figure 4. SARS-CoV-2 has hyperstable genomic regions not present in SARS-CoV. Secondary structures of COV001 (**A**) and COV002 (**B**) sequences in SARS-CoV (left). and SARS-CoV-2 (right). (**C**) Secondary structure of a mutated variant (C17,114A & C17,144T) of the SARS-CoV COV001 sequence showing the hyperstable stem-loop structure can be introduced with variants observed in this region of the SARS-CoV-2 genome. Colors indicate base-pairing probability.

similarity (>84.6%) between SARS-CoV and SARS-CoV-2. Although the COV002 sequences of both viruses possessed comparable secondary structures, a stem-loop identified in the COV001 sequence of SARS-CoV-2 was absent in the corresponding SARS-CoV sequence. The role of this stem-loop structure will require further investigation, however, such evidence offers an exciting glimpse into how SARS viruses evade host immunity, and offer potential new therapeutic targets.

Fate-seq takes advantage of the long length limit (300mer) of Twist Oligos to enable the identification of high-dimensional RNA structures. Because longer RNA sequences generally form more complex secondary structures, long oligos are needed to completely capture these structures. Long oligos also accommodate sequence elements (e.g., primer sequences) that simplify cloning and library preparation. Although such sequences can be added to oligos with an extra PCR step, doing so unnecessarily lengthens the workflow and can introduce sequence errors.

Fate-seq quantifies changes in the representation of viral RNA sequences between time points. Because of this, Fate-seq requires that viral sequences be accurately represented and uniformly distributed throughout the experiment. Sequence errors, whether introduced by extra PCR steps or subpar oligo synthesis, can skew the representation of viral sequences at $t = 0$ and adversely impact the analysis of a Fate-seq experiment. The exceptional fidelity and uniformity of Twist Oligo Pools ensure that Fate-seq results reflect true results and not technical artifacts.

In addition to the application described here, efforts to develop vaccines against SARS-CoV-2 could also benefit from Fate-seq. Recent advances in exogenous RNA technologies have made mRNA vaccines a reality. In the future, Fate-seq could be used to further optimize the stability of RNA therapeutics, a key challenge to their efficacy.

ACKNOWLEDGEMENTS

Twist Bioscience would like to thank Prof. Nobuyoshi Akimitsu and Dr. Kentaro Kawata for the detailed discussions around Fate-seq, genome tiling and host immune escape, and for allowing us to showcase their research in this application note.

REFERENCES

Brierley I, Pennell S, & Gilbert RJ (2007) Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nature Reviews Microbiology, 5(8), 598–610.

Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780.

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, & Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics, 23(21), 2947–2948.

Love MI, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.

Manokaran G, Finol E, Wang C, Gunaratne J, Bahl J, Ong EZ, Tan HC, Sessions OM, Ward AM, Gubler DJ, Harris E, Garcia-Blanco MA, & Ooi EE (2015) Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness. Science, 350(6257), 217–221.

Pijlman GP, Funk A, Kondratieva N, Leung J, Torres S, van der Aa L, Liu WJ, Palmenberg AC, Shi PY, Hall RA, & Khromykh AA (2008) A highly structured, nuclease-resistant, noncoding RNA produced by flaviviruses is required for pathogenicity. Cell Host & Microbe, 4(6), 579–591.

Shu Y & McCauley J (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveillance, 22(13), 30494.

Wakida H, Kawata K, Yamaji Y, Hattori E, Tsuchiya T, Wada Y, Ozaki H, & Akimitsu N (2020) Stability of RNA sequences derived from the coronavirus genome in human cells. Biochemical and biophysical research communications, 527(4), 993–999.

World Health Organization (2020) Draft landscape of COVID-19 candidate vaccines. https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines. Accessed October 25, 2020.