Influenza A virus utilizes noncanonical cap-snatching to diversify its mRNA/ncRNA

LICHAO LI,[1,3] HUI DAI,[1,3] AN-PHONG NGUYEN,[1] RONG HAI,[2] and WEIFENG GU[1]

[1]Department of Molecular, Cell, and Systems Biology, University of California, Riverside, California 92521, USA
[2]Department of Microbiology and Plant Pathology, University of California, Riverside, California 92521, USA

ABSTRACT

Influenza A virus (IAV) utilizes cap-snatching to obtain host capped small RNAs for priming viral mRNA synthesis, generating capped hybrid mRNAs for translation. Previous studies have been focusing on canonical cap-snatching, which occurs at the very 5′ end of viral mRNAs. Here we discovered noncanonical cap-snatching, which generates capped hybrid mRNAs/noncoding RNAs mapped to the region ∼300 nucleotides (nt) upstream of each mRNA 3′ end, and to the 5′ region, primarily starting at the second nt, of each virion RNAs (vRNA). Like canonical cap-snatching, noncanonical cap-snatching utilizes a base-pairing between the last nt G of host capped RNAs and a nt C of template RNAs to prime RNA synthesis. However, the nt upstream of this template C is usually A/U rather than just U; prime-realignment occurs less frequently. We also demonstrate that IAV can snatch capped IAV RNAs in addition to host RNAs. Noncanonical cap-snatching likely generates novel mRNAs with start AUG encoded in viral or host RNAs. These findings expand our understanding of cap-snatching mechanisms and suggest that IAV may utilize noncanonical cap-snatching to diversify its mRNAs/ncRNAs.

Keywords: cap-snatching; influenza virus; priming and realignment; noncoding RNA or ncRNA; novel influenza virus mRNA

INTRODUCTION

Influenza A virus (IAV) often causes epidemic and pandemic respiratory infection in humans and animals. Its genome contains eight negative strand virion RNAs (vRNA). IAV utilizes an RNA-dependent RNA polymerase (RdRP) complex to generate positive strand mRNAs and complementary RNAs (cRNA) from template vRNAs, and vRNAs from template cRNAs (Shi et al. 1995; Kobayashi et al. 1996; Shih and Krug 1996; Bouvier and Palese 2008; Guilligay et al. 2008; Sugiyama et al. 2009; Reich et al. 2014). vRNAs and cRNAs are exactly complementary and both bear 5′ triphosphate (ppp) but no poly(A) tail (Desselberger et al. 1980; Bouvier and Palese 2008). In contrast, each IAV mRNA is a hybrid RNA composed of a host capped small RNA, a.k.a., "cap or host cap," IAV-encoded sequence, and poly(A) tail obtained via the stuttering mechanism using a template poly(U) sequence encoded in each vRNA; most IAV-coded mRNA sequences are one nucleotide (nt) shorter at the 5′ end than corresponding cRNAs since IAV RdRP usually initiates mRNA synthesis using the penultimate nt, C, of template vRNAs rather than the last nt, U (Supplemental Fig. S1A; Bouloy et al. 1978; Krug et al. 1979; Luo et al. 1991; Pritlove et al. 1998; Poon et al. 1999; Zheng et al. 1999).

IAV RdRP is composed of polymerase basic protein 1 (PB1), polymerase basic protein 2 (PB2), and polymerase acidic protein (PA) (Kobayashi et al. 1996; Shih and Krug 1996; Guilligay et al. 2008; Sugiyama et al. 2009; Reich et al. 2014). To generate a hybrid mRNA using cap-snatching, (1) PB2 binds host capped RNAs; (2) PA cleaves at positions 10–15 nt from 5′ ends, obtaining host caps; and (3) IAV RdRP utilizes the last nt, usually G, of host caps (capped small RNAs) to base pair with the penultimate (−2) nt, always C, of template vRNAs to prime mRNA synthesis (Fig. 1; Bouloy et al. 1978; Krug et al. 1979; Plotch et al. 1979; Shi et al. 1995; Kobayashi et al. 1996; Fodor et al. 2002; Bouvier and Palese 2008; Guilligay et al. 2008; Dias et al. 2009; Sugiyama et al. 2009; Yuan et al. 2009; Reich et al. 2014). Although this G appears as encoded by a template C, it actually belongs to host caps (Beaton and Krug 1981; Hagen et al. 1995; Rao et al. 2003; Bouvier and Palese 2008). Unlike IAV mRNAs, vRNAs, and cRNAs are synthesized by IAV RdRP without any primer (Desselberger et al. 1980; Bouvier and Palese 2008; Pflug et al. 2017).

[3]These authors contributed equally to this work.
Corresponding author: weifeng.gu@ucr.edu
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.073866.119.

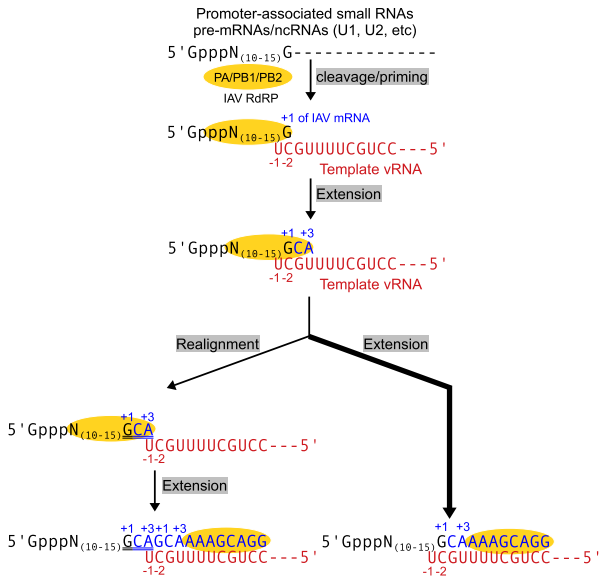

FIGURE 1. Model of canonical cap-snatching. IAV RdRP cleaves host capped RNAs preferentially 3′ of G to generate capped small RNAs (cap), utilizes this G to base pair with the penultimate nt (−2), C, of template vRNAs, and extends caps by a few nts. The extension usually continues to generate full-size mRNAs. However, a fraction of the extended caps disassociate from template vRNAs, utilize the last nt (−1), usually A, to reanneal (realign) with the last (−1) nt of template vRNAs, U, and are extended again.

Approximately 20% of IAV mRNAs contain additional nts between a host cap and virus-coded sequence, as estimated using IAV mRNAs containing U1/U2 snRNA caps (Decroly et al. 2011; Gu et al. 2015; Koppstein et al. 2015). At least three models may account for these extra nt, which appear as a repeat of the first few nts in IAV mRNA 5′ UTRs. In the most favored model "prime-*cis*-realignment" or "prime-realignment," (1) a cap is cleaved by IAV RdRP, annealed to a template vRNA using a single base-pairing between the cap last nt (G) and template penultimate nt (C), and extended up to 9 nt (predominantly 4 or less); (2) a fraction of nascent mRNAs detach from the template, reanneals with the same template using a base-pairing between the mRNA last nt (A) and template last nt (U), and then is extended again (Fig. 1; Gu et al. 2015; Koppstein et al. 2015; Te Velthuis and Oymans 2018). Consistent with this, PA usually prefer cutting host caps 3′ of G and the first extension predominantly ends with A, perfectly matching the last 2 nt, 3′UC5′, of templates, respectively (Fig. 1; Gu et al. 2015; Koppstein et al. 2015). A second model "prime-random-realignment" assumes that the first step utilizes the same annealing/extension mechanism, but in the second step, the released nascent mRNAs anneal with both *cis* (original) and *trans* (other) templates. Although there has been no evidence supporting this model, it generates similar results as the first model because the 5′ UTRs of IAV mRNAs are almost identical (Supplemental Fig. S1A). A third "prime-only" model assumes that a fraction of IAV RdRP snatches both host and viral capped RNAs. If host caps are used, viral mRNAs contain no extra sequences between host caps and virus-encoded sequences. However, if a viral mRNA is snatched, a cap composed of a host RNA and a few IAV-encoded nt ending with nt A anneals with the template last nt (U), priming mRNA synthesis. The net result is that the 5′ end of one IAV mRNA is snatched for making another IAV mRNA, generating a short IAV-encoded repeat in the latter. This model is clearly not preferred since (1) IAV selectively protects its mRNAs from cap-snatching and (2) IAV RdRP prefers to cleave 3′ of G instead of A (Datta et al. 2013; Koppstein et al. 2015).

In the era of Sanger sequencing, it had been long held that IAV snatches caps from host pre-mRNAs (Bouloy et al. 1978; Krug et al. 1979; Plotch et al. 1979, 1981; Caton and Robertson 1980; Dhar et al. 1980; Beaton and Krug 1981; Shaw and Lamb 1984). However, this conclusion may be at least incomplete since: (1) very limited sequencing data and gene annotations were available at that time; and (2) mRNAs may be preferentially selected as cap donors out of multiple genomic matches resulted from the small query size of snatched host caps. Recently, three groups including us used high-throughput sequencing to obtain a more comprehensive spectrum of host cap donors (Sikora et al. 2014, 2017; Gu et al. 2015; Koppstein et al. 2015). Both our group and Koppstein et al. independently demonstrated that noncoding RNAs (ncRNA) are the top cap donor while Sikora et al. did not examine ncRNAs in their first paper but included them later (Sikora et al. 2014, 2017). For example, we showed that U1 and U2 snRNAs alone provided ∼7% of viral caps and that ncRNAs including U1 and U2 provided at least ∼55% (Gu et al. 2015). We also used in situ hybridization to verify the high-throughput sequencing result, excluding the possibility that the U1/U2 caps were added to IAV mRNAs via cloning artifacts.

Promoter associated small RNAs (PASR) or capped small RNAs (csRNA) constitute a new ncRNA species, which is generated during Pol II-mediated transcription initiation (Seila et al. 2008; Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Nechaev et al. 2010; Gu et al. 2012). PASRs usually exhibit a bimodal distribution with sense PASRs comapped with the 5′ end of pre-mRNAs/mRNAs and antisense PASRs mapped ∼150 nt upstream in the opposite direction. Although we cannot single out sense PASRs from host mRNAs/pre-mRNAs/sense PASRs mixture to explicitly determine its contribution to IAV caps, we were able to demonstrate that antisense PASRs contributed ∼7% of IAV caps (Gu et al. 2015) and their snatching rate (IAV cap)/(IAV cap + host cap) was much higher than that of mRNAs/pre-mRNAs/sense PASRs. This raises a possibility that IAV caps which appear to be snatched from pre-mRNAs may be derived from an alternative source, that is, sense PASRs.

Most cap-snatching analyses used PCR to enrich the annotated 5′-end sequences of IAV mRNAs, and had to rely on host gene annotations to identify cap source, usually generating nonunique matches due to the tiny size of snatched caps (Sikora et al. 2014, 2017; Koppstein et al. 2015). We previously developed CapSeq, a one-pot strategy to enrich and clone the 5′-end sequences of host and viral capped RNAs in a single library. CapSeq utilizes sequential enzymatic treatments to enrich capped RNAs, including Terminator exonuclease (Terminator) for removing monophosphorylated RNAs (p-RNA), predominantly rRNAs, Calf intestinal phosphatase (CIP) for removing any residual p-RNA after the Terminator treatment, and Tobacco acid pyrophosphatase (TAP) for generating p-RNAs from capped RNAs and ppp-RNAs for ligation-dependent cloning. CapSeq simultaneously obtains a real-time profile of the 5′-end sequences of both host and IAV capped RNAs, including those not annotated, for example, PASRs, allowing us to perform more comprehensive analyses and obtaining unique matches by linking host and IAV caps obtained in the same data set.

Our previous study focused on canonical cap-snatching occurring primarily at the first nt of IAV mRNAs, defined as "mRNA +1" for convenience (Gu et al. 2015). As shown in Supplemental Figure S1A, IAV mRNAs, cRNAs, and vRNAs are read from 5′ to 3′ as +1, +2, etc., and from 3′ to 5′ as −1, −2, etc. Here we identified noncanonical cap-snatching in two regions: one as a cluster of loci ∼300 nt upstream of each mRNA 3′ end (mRNA 3′ cluster) and the other covering the +1 to +10 (primarily at +2) nt of each vRNA (vRNA 5′ region). mRNA 3′ clusters generate novel mRNAs and capped ncRNAs sharing the same strands with annotated mRNAs, while vRNA 5′ regions likely only produce ncRNAs. These novel mRNAs are usually in-frame with annotated mRNAs but encoding shorter proteins with 0-3 amino acids derived from host caps. Like canonical cap-snatching, noncanonical cap-snatching also prefers a G/C base-pairing between the last (−1) nt, usually G, of a cap and the template −2 C (vRNA 5′ regions) or an internal template C (mRNA 3′ clusters), for priming the initial synthesis of viral RNAs. We find that the nt downstream from the template C, usually A or U (collectively as W), plays important roles in selecting internal template C's for mRNA synthesis in mRNA 3′ clusters. Although *cis*-realignment, which shares the same templates with initial priming/extension (Fig. 1), utilizes the same A/U base-pairing mechanism in canonical and noncanonical regions, it occurs less frequently in noncanonical regions. Our evidence indicates that *trans*-alignment does occur, suggesting that one IAV mRNA could contain three sequences, one derived from a host cap and the other two derived from two IAV mRNAs. In conclusion, this study provides further insight into the cap-snatching mechanism and suggests that IAV may use cap-snatching to diversify its mRNAs and ncRNAs.

RESULTS

CapSeq highly enriches reads mapped to the 5′ end of capped RNAs and ppp-RNAs

We previously used CapSeq coupled with high-throughput sequencing to enrich/obtain RNA reads (a read is represented by its 5′ nt in this paper) mapped to the 5′ end of capped RNAs in noninfected samples (Gu et al. 2012, 2015; Hainer et al. 2015). However, we barely obtained any reads mapped to the 5′ end of ppp-RNAs due to lack of ppp-RNAs in the samples. In this study, we found that CapSeq efficiently cloned ppp-RNAs (vRNA) in IAV-infected A549 cells despite a CIP treatment step, which aims to dephosphorylate any phosphorylated RNAs to prevent cloning. The failure of excluding ppp-RNAs was likely due to insufficient CIP, which was overwhelmed by excessive monophosphorylated nucleotides (∼3000 folds as much as intact RNAs) and resulted from a preceding rRNA depletion step in the "one-pot" CapSeq reaction (Gu et al. 2012, 2015).

Although the CIP step did not work, we cloned few p-RNAs, a major artifact in CapSeq analyses, which was likely generated by specific RNA degradation pathways and usually mapped to the internal sites of capped RNAs and other RNAs. This depletion is achieved primarily by Terminator exonuclease (Terminator) treatment, which only destroys p-RNAs but not ppp-RNAs. For example, in the noninfected samples, CapSeq utilizes Terminator to remove 99.5% of rRNAs, the major p-RNA species in total RNA. Consistent with such an efficiency, a similar depletion of p-RNA reads mapped to mRNAs generated at least 50,000-fold enrichment for reads (capped RNAs) mapped to the 5′ end over those (p-RNAs) to any internal position of mRNAs (Gu et al. 2012, 2015).

Here in the IAV-infected samples, ∼97.9% of IAV mRNA reads were mapped to +1, the capped read position, and only ∼2.1% were mapped to internal sites; ∼92.0% of vRNA reads were mapped to +1, the ppp-RNA read position, and 8.0% were mapped to internal sites (Fig. 2). This constitutes ∼77,000- and ∼20,000-fold (97.9/2.1 × 1700 and 92/8 × 1700) enrichment for reads starting at +1 over reads starting at any internal position, calculated based on the average size, ∼1700 nt, of IAV RNAs. Based on a high depletion (99.5 ± 0.3%) of p-rRNAs in all the five samples analyzed, we believe that only a fraction of these internal mRNA or vRNA reads represent p-RNAs. Consistent with this hypothesis, we found that the majority of these internal sites on mRNAs and a significant fraction on vRNAs represent capped RNAs synthesized via noncanonical cap-snatching, as discussed below. We have not optimized the conditions to minimize the unexpected ppp-RNA byproducts, since they do not affect cap-snatching analyses. More importantly, as uncapped RNAs, they serve as negative internal controls for cloning artifacts.

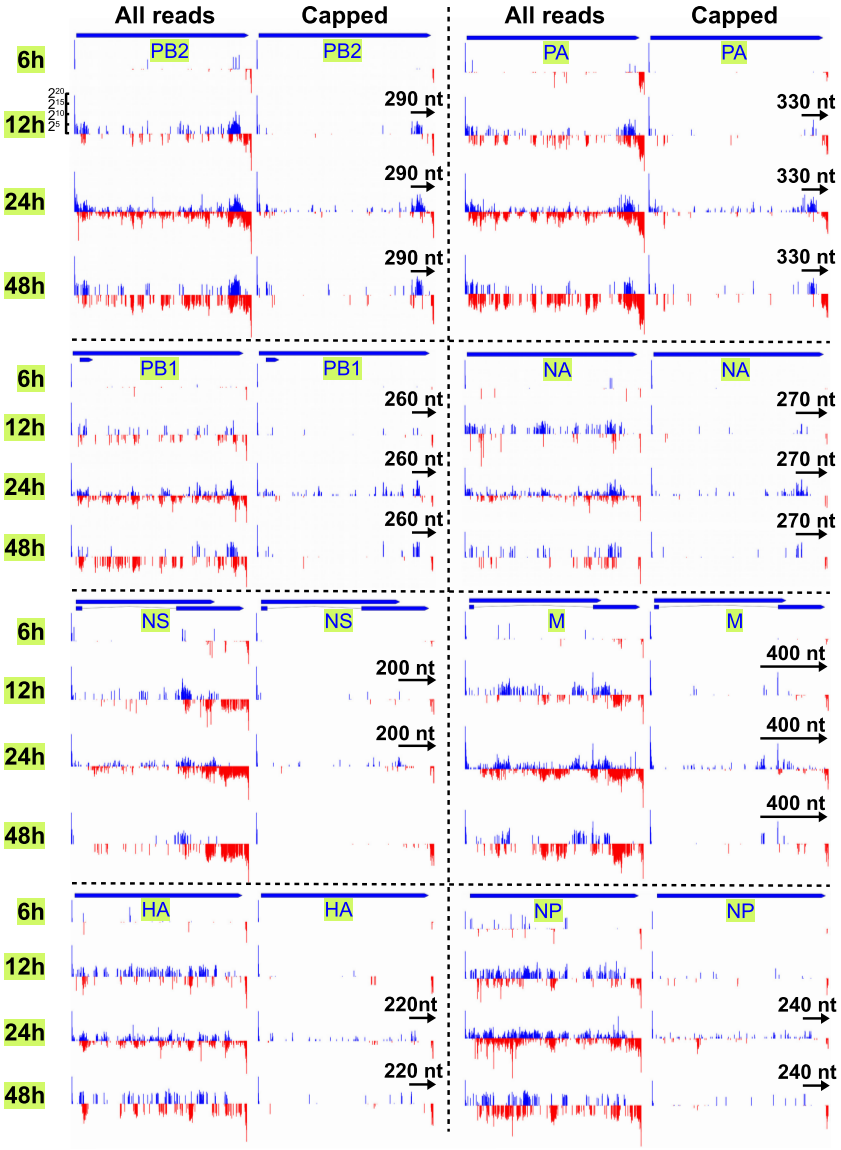

FIGURE 2. The histogram of the IAV reads. IAV RNA reads were represented by their first mapped nts with blue indicating mRNAs/cRNAs and red indicating vRNAs, and each IAV position was visualized using the combined read number (*Y*-axis) derived from the comapped reads and normalized to the total non-rRNA/tRNA host/IAV reads. For each IAV RNA strand, the blue annotations at the *top* represent coding frames, which are a little bit smaller than corresponding IAV RNA strands; the *left* part of each panel represents the histogram of all the IAV reads, and the *right* panel represents that of only capped RNA reads at 6–48 h postinfection timepoints. The black arrows indicate the distance between the left edges of mRNA 3′ clusters and cRNA 3′ ends. Each IAV RNA strand is represented with the same width (*x*-axis), generating different unit sizes. However, each panel uses the same log scale of *Y*-axis, as labeled in the *top left* panel.

Identification of noncanonical cap-snatching

As a 5′ ligation-dependent method, CapSeq allows us to obtain a directional library with cDNA sequence explicitly representing RNAs, avoiding confusion when mapping reads derived from dsRNAs. Since the RefSeq database lacked IAV 5′ and 3′ UTRs, we used the CapSeq reads and a Perl script to assemble the 5′ UTRs, generating 5′AGC(A/G)AAA GCAGG (G for PB1 and A for the rest) for cRNAs, 5′GC(A/G)AAAGC AGG for mRNAs, most of which lack 5′ A as compared to the corresponding cRNAs, and 5′AGUAGAAACA AGG for vRNAs (Supplemental Fig. S1A). Based on the reciprocal template/product relationship, the 3′ UTRs of cRNAs and vRNAs contain 5′CCUUGUUUCU ACU and 5′CCU GCUUU(U/C)GCU. As reported, the 5′ and 3′ UTRs of each vRNA and cRNA are basically inverted repeats, which are able to form imperfectly base paired dsRNA "panhandle" structures within the same molecule (Supplemental Fig. S1A; Desselberger et al. 1980).

We previously analyzed canonical cap-snatching at 12- and 24-h postinfection timepoints (Gu et al. 2015). Since cells may exhibit distinct viral RNA profiles and substrate availability/specificity following postinfection, here we added an early (6-h) and late (48-h) postinfection timepoints, generating a broader time course. Under our condition, IAV mRNA and vRNA levels were extremely low at the 6-h timepoint, accounting for ~0.2% and 0.04% of total host/viral RNAs, and increased progressively at the 12- and 24-h timepoints, indicating active transcription and replication; vRNA levels continuously increased while mRNA levels dropped dramatically at the 48-h timepoint, indicating a typical late stage timepoint (Supplemental Fig. S1B). We repeated the 6-h timepoint but only obtained similar mRNA and vRNA levels, 0.5% and 0.08%, respectively. These low levels prevent us from obtaining sufficient IAV reads via high-throughput sequencing at reasonable cost. We, therefore, focused on the 12- to 48-h timepoints. Since all the three timepoints exhibit similar conclusions and the 12- and 24-h timepoints contain higher levels of IAV mRNAs (7.2% and 14.7%), we used them to represent most analyses.

The "abnormal" frequency of CapSeq reads mapped to the internal sites of mRNAs in the IAV-infected samples prompted us to investigate if cap-snatching occurs at loci

other than IAV mRNA +1, generating mRNAs starting internally. We modified our previous method to map the CapSeq reads to the full length of IAV cRNAs and vRNAs, and extracted non-IAV 5′ portions as potential host-derived caps and 3′ portions completely matching IAV RNAs (Gu et al. 2015). Here we use the first matched nt of a read to represent the mapped site of the whole read and the part 5′ of the matched nt to serve as a potential host cap. For example, if a 40-nt read is mapped to vRNA +2 to +41, the mapped position is defined as +2 and the read does not contain a host cap or a cap of size 0; if the last 30 nt of this read is mapped to vRNA +2 to +31, the mapped position is defined as +2 and the read contains a 10-nt host cap. Any IAV site could contain both capped RNA reads, which are defined as "containing a host cap of at least 5 nt long" in this paper, and noncapped reads.

We observed a cluster of noncanonical capped reads mapped ~300 nt upstream of each mRNA 3′ end in addition to capped reads at the canonical site mRNA +1 (Fig. 2). As expected, the noninfected controls exhibited almost no reads (Supplemental Fig. S2). In the 12–48-h postinfection samples, the reads form 3′ clusters of high abundance on PA, PB1, PB2, and NA mRNAs, 3′ clusters of low abundance on M and NS, and no apparent 3′ clusters on HA and NP (Fig. 2). However, the 6-h sample only exhibited tiny 3′ clusters on PA, PB1, and PB2 and no clusters on others likely due to the extremely low IAV expression (0.24% of total host/IAV RNAs) (Fig. 2). Overall, the capped RNA reads in mRNA 3′ clusters are ~0.3% of those mapped to IAV mRNA +1. However, this ratio varies dramatically among individual mRNAs with the highest for NA (8.52%) and lowest for M (0.04%) at the 24-h timepoint (Supplemental Fig. S3A). We observed an even higher ratio (9.4%) for NA at the 48-h timepoint. In addition to mRNA +1 and 3′ clusters, other mRNA regions also contain many capped RNA reads of less abundance as well as some noncapped RNA reads (Fig. 2). However, the non-5′ regions of vRNAs contain few capped RNA reads but many noncapped RNA reads (Fig. 2). This stark contrast suggests that the noncanonical capped RNA reads mapped to IAV mRNAs were not generated by cloning artifacts, also as discussed below.

We found that the 5′ region (+1 to +10) of each vRNA strand, especially at +2, all bear a high rate of capped RNAs (capped RNAs divided by all RNAs) in all four timepoints, while reads mapped downstream barely contain host caps (Table 1). Among the nine million non-rRNA/tRNA reads in the 24-h sample, ~70.4% were mapped to host mRNAs, and ~14.7% and ~14.9% were mapped to IAV mRNAs and vRNAs, respectively. The 12-h sample contains relatively more IAV mRNAs (7.2%) than vRNAs (3.5%), both of which are much lower than those in the 24-h sample. Among the reads mapped to the first-10 nt of vRNAs, 97.7%, 1.3%, and 1.0% were assigned to the +1, +2, and the rest 8 nt (Fig. 3). Overall, ~0.7% of the total reads in this region are capped RNAs. However, the distribution pattern does not follow that of the total reads as 46.6%, 46.3%, and 7.1% were mapped to the +1, +2, and the rest 8 nt, respectively. As a consequence, the rate of capped RNA reads is much higher at +2 (26.4%) and the rest 8 nt (5.2%) than +1 (0.3%) (Fig. 3). This suggests that vRNA +1 and +2 are used as the preferred start sites for ppp-RNAs and capped RNAs, respectively. Interestingly cRNA +1 and +2 (the same as mRNA +1) are also used the same way. Therefore, not only do vRNA and cRNA share a sequence symmetry (reverse complement), they also exhibit a symmetry of transcription start site usage (Supplemental Fig. S1A). Moreover, the inverted repeat feature of vRNA or RNA ends determines that

TABLE 1. The cap-snatching rates at canonical and noncanonical loci of IAV RNAs

	mRNA +1		mRNA 3′ cluster		vRNA +2		vRNA +1		vRNA After +10	
	Rate[a]	Ratio[b]	Rate	Ratio	Rate	Ratio	Rate	Ratio	Rate	Ratio
PB2	8.3×10^{-2}	64,482	3.9×10^{-2}	30,108	2.7×10^{-2}	20,914	7.8×10^{-6}	6	0	0
PB1	4.4×10^{-2}	33,612	2.3×10^{-2}	17,817	1.1×10^{-2}	8489	0	0	0	0
PA	7.6×10^{-2}	58,743	1.0×10^{-2}	7832	5.0×10^{-3}	3889	2.7×10^{-6}	2	2	87
HA	5.7×10^{-2}	43,793	4.4×10^{-3}	3403	1.1×10^{-2}	8680	0	0	1.1×10^{-4}	0
NP	5.3×10^{-2}	41,329	3.5×10^{-3}	2711	3.8×10^{-3}	2960	0	0	0	0
NA	7.2×10^{-2}	55,878	4.7×10^{-2}	36,278	0	0	0	0	0	0
M	1.9×10^{-2}	14,295	7.8×10^{-3}	6022	1.2×10^{-2}	9107	0	0	0	0
NS	4.3×10^{-2}	33,003	2.0×10^{-2}	15,328	1.3×10^{-2}	9659	0	0	0	0
All IAV	4.4×10^{-2}	34,300	2.2×10^{-2}	17,345	1.3×10^{-2}	14,830	4.6×10^{-6}	4	8.1×10^{-6}	0
Host	1.3×10^{-6}	1	NA	NA	NA	NA	NA	NA	NA	NA

"+1, +2, and +10" refer to reference RNA sites where the matched parts of reads start.
[a]RNA containing a U1 or U2 cap/all capped and noncapped RNA.
[b]The rate of IAV/the rate of host A549 cells (1.3×10^{-6}).

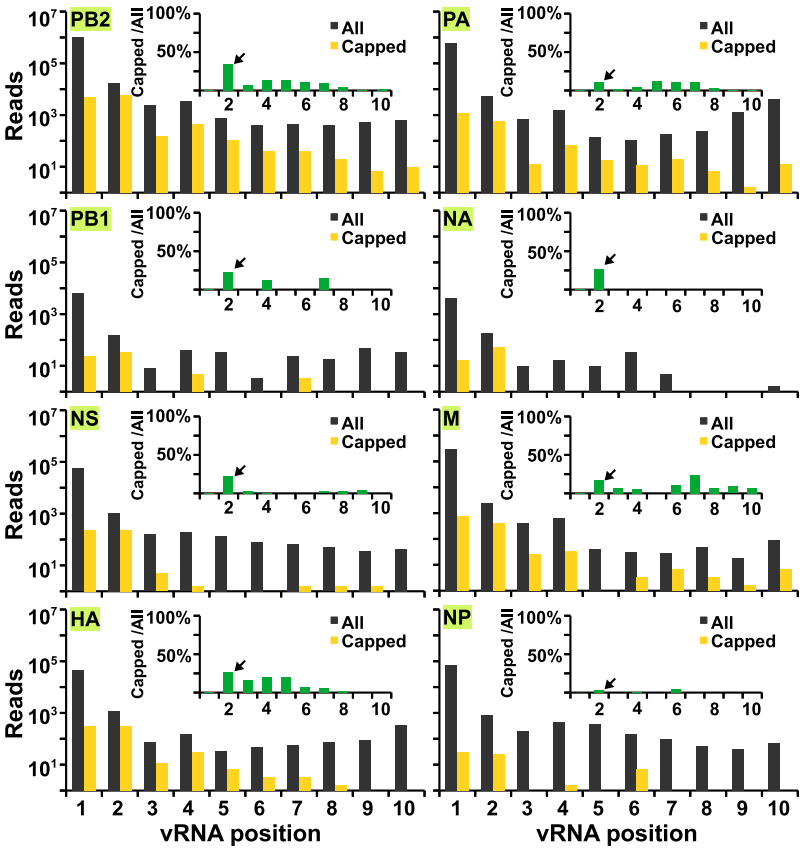

FIGURE 3. The histogram of vRNA 5′ regions. The reads are normalized, mapped, combined, and visualized using their start sites, as described in Figure 2, with yellow indicating capped RNA reads and black indicating all reads at the position +1 (the first) to +10 of vRNAs. The *insets* represent the ratio of capped RNA reads to all reads at each position with arrows indicating vRNA +2.

cRNA/vRNA +1 and +2 encode the same 5′AG sequence (Supplemental Fig. S1A).

U1 and U2 snRNAs are the top cap donors for noncanonical cap-snatching

We previously reported that U1 and U2 were the top cap donors, contributing 3.3% and 3.5% of all caps at IAV mRNA +1 (Gu et al. 2015). Here we demonstrate that they are also the top donors in noncanonical cap-snatching since U1 and U2 in the 24-h postinfection sample contributed 3.3% and 5.1% caps in mRNA 3′ clusters, and 1.6% and 5.2% in vRNA 5′ regions (Supplemental Fig. S3B). These rates are at least 10 folds as much as those of the top host mRNA donors. These canonical and noncanonical rates appear to be within the same range with variations likely caused by the low capped read number in noncanonical regions.

We observed a weak positive correlation between the levels of cap donors and IAV caps at the 24-h timepoint ($P < 1 \times 10^{-4}$ in Supplemental Fig. S3B). This weak correlation ($r = 0.26$ instead of 1 for a perfect correlation) may be caused by: (1) the low read number of IAV capped RNAs; and (2) the host cap number includes pre-RNAs, mature RNAs and PASRs, not all of which are cap-snatching substrates. ncRNAs appear to have a higher cap-snatching rate (IAV cap/[IAV cap + host cap]) than host mRNAs/pre-mRNAs/PASRs (Supplemental Fig. S3B). For example, U1 and U2 have higher cap-snatching rates than most mRNAs (Supplemental Fig. S3B). We also reached similar conclusions using the 12-h timepoint (Supplemental Fig. S3C). All these observations are consistent with our previous finding that canonical cap-snatching at mRNA +1 prefers ncRNAs, especially U1 and U2 snRNAs, suggesting that IAV RdRP utilizes the same substrate pool for canonical and noncanonical cap-snatching (Gu et al. 2015).

Verifying noncanonical cap-snatching using U1 and U2 snRNAs

The non-IAV 5′ portion of a read can be generated by cloning artifacts of low frequency rather than by cap-snatching. For example, host caps can be ligated to IAV RNAs by RNA ligases, substituting for 5′ ligation linkers. Such events are very rare since (1) the 5′ linker amount used is 50 pmol while the total host capped RNA is only ~0.06 pmol, only a small fraction of which is likely degraded to generate 11-nt caps (the average size on IAV); (2) degraded RNAs usually contain 3′ cyclic phosphate or 3′ monophosphate, incompatible with RNA ligation. Non-IAV 5′ portions can be generated by reverse transcriptase jumping along templates, generating noncontinuous IAV sequences which were dissected as non-IAV and IAV parts by our algorithm, as in the defective interfering (DI) particles (Saira et al. 2013). This model is disfavored since (1) jumping events cannot de novo generate ~11-nt U1/U2 sequences out of IAV sequences; and (2) jumping events cannot generate cap-snatching signatures including the cap size, cleavage motif, priming motif and realignment feature, as discussed below. In addition, non-IAV 5′ portions can be generated by reverse transcriptases via template-switching, in which a reverse transcriptase utilizes two templates, leading to a ligation-like behavior (Cocquet et al. 2006). Again, this mechanism occurs at a very low frequency and cannot account for cap-snatching signatures either.

To support our noncanonical cap-snatching model, we provided three negative internal controls. The host mRNAs cloned in the same reaction serve as a "perfect" negative control since it was treated exactly the same way as IAV capped RNAs. Here we found that the rate of U1/U2 cap-containing host mRNA reads, defined as "U1/U2-containing reads divided by all mRNA reads," is extremely low (1.3×10^{-6}) in the 24-h sample. In contrast, IAV mRNA +1 and noncanonical sites (mRNA 3′ clusters and vRNA 5′ regions) all contain similar rates of U1/U2 caps, ~15,000- to 30,000-fold higher than that of host mRNAs (Table 1). This general conclusion also applies to almost every individual RNA strand, including eight mRNA and seven vRNA strands (Table 1). The only exception is the NA vRNA 5′ region, which lacks enough read coverage for obtaining a U1/U2 cap rate. A second negative control is the non-5′ region of vRNAs, in which the U1/U2 cap rate is very close to that of host mRNAs, that is, the "background rate" (Table 1). A third negative control is vRNA +1, which contains ~92% of all vRNA reads. Usually the +1 or 5′ end is the hot spot for template-switching since it is the last template nt for cDNA synthesis, as we utilized this mechanism for cloning small RNAs (Gu et al. 2009). However, we only observed a rate a little bit higher than the background (Table 1). In conclusion, these negative controls serve as convincing evidence supporting that noncanonical cap-snatching is not caused by cloning artifacts. Consistent with this, the ratio of U1/U2 caps to all caps in canonical and noncanonical cap-snatching regions are almost the same, as discussed above. Moreover, both share the same cap-snatching features, as discussed below. The canonical cap-snatching results, therefore, serves as positive controls for the noncanonical ones.

We also used an only-RT/PCR-based method to confirm our results without using ligation. We used random hexamers and oligo $(dT)_{12–18}$ to generate IAV cDNA, respectively, and then amplified the cDNA using a shared reverse primer with different 10-nt forward primers 5′-attached with an 11-nt U2 5′ sequence. Three positive forward primers were picked from the sites containing a U2 cap in the 3′ region of PB2 mRNA, and three negative control primers were randomly picked from the upstream sites containing no U2 cap (Supplemental Fig. S4). We had to use a second shared reverse primer in the nested PCR reactions to achieve PCR specificity. As expected, we can easily detect the targets, as confirmed by Sanger sequencing, in the positive PCR reactions even at the low PCR cycle number while failing to detect any product in the negative controls even at the high PCR cycle number. Interestingly, we also obtained a truncated product for each positive reaction, all of which contains the same 33-nt deletion (Supplemental Fig. S4). This deletion does not affect our conclusion since it is ~60–70 nt downstream from the start sites of the capped RNA reads. Moreover, we observed the same positive results at the same PCR cycle number using the cDNA templates made with either random hexamers or oligo $(dT)_{12–18}$, suggesting at least a significant fraction of capped RNAs in IAV mRNA 3′ regions contain poly(A) tails.

Realignment occurs much less frequently in mRNA 3′ clusters and vRNA 5′ regions

We analyzed the size distribution of IAV caps including the extra nts added via the realignment mechanism. Both mRNA 3′ clusters and vRNA 5′ regions exhibit an almost symmetric bell-like size distribution of IAV caps of 7–16 nt long peaking at 11 nt (Supplemental Fig. S5A). In contrast, mRNA +1 sites exhibit a 12-nt peak with a skewed size distribution, as the slope left of the peak is much steeper than that right of the peak. This was apparently caused by the extra nts added via the realignment mechanism since we obtained a symmetric distribution peaking at 11 nt after removing the most abundant species of these nts (Supplemental Fig. S5A). There were no dramatic size changes after removal of the realigned extra nts in mRNA 3′ clusters and vRNA 5′ regions (Supplemental Fig. S5A), suggesting that realignment occurs much less frequently in noncanonical cap-snatching.

To explicitly compare realignment rates, we analyzed U1 and U2 caps on IAV RNAs. We only considered mRNA +1 and vRNA 5′ regions, since all but PB1 mRNAs and all vRNAs share the first 11 and 13 nt, respectively (Supplemental Fig. S1A), allowing us to easily figure out prime-realignment patterns. Moreover, these loci represent at least 90% cap-snatching events on the corresponding RNAs. We found that 16% and 23% of U1 and U2 caps at IAV mRNA +1 contain extra nt, and at least 15% and 22% clearly contain recognizable realignment or rerealignment patterns, suggesting that almost all the extra nt were added via realignment. In contrast, ~7% each of U1 and U2 caps in IAV vRNA 5′ regions contain extra nts, and only ~1% and ~5.6% possibly contain recognizable realignment patterns.

Identification of *trans*-realignment

In the canonical prime-*cis*-realignment model, prime and realignment steps are coupled on the same RNA templates (Decroly et al. 2011; Geerts-Dimitriadou et al. 2011b; Koppstein et al. 2015; Te Velthuis and Oymans 2018). To examine if a realignment step can utilize a different RNA template, a process defined here as "*trans*-realignment," we first analyzed the realignment patterns at mRNA +1 and extracted the most abundant "extra nt" generated via realignment. As shown in Figure 4A, the first priming step utilizes the base-pairing of the host cap −1 G and template vRNA −2 C; the cap is usually extended for 2 −4 nt, ending at "A" (first extension); the extended sequence is realigned using the base-pairing of the cap −1 A and template vRNA −1 U (realignment or second

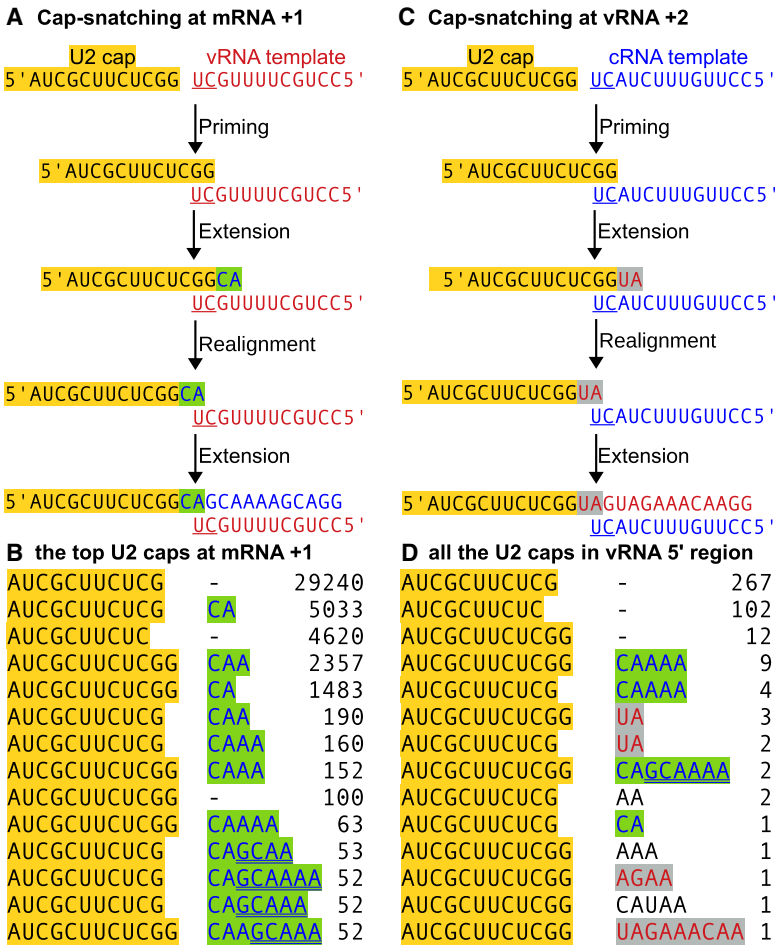

FIGURE 4. The U2 cap realignment at mRNA +1 and vRNA +2. (*A*,*C*) A U2 cap highlighted in yellow is annealed with a template vRNA (*A*) or cRNA (*C*) via the base-pairing between the cap −1 G and template −2 C in the initial priming step, extended (green in *A* or gray in *C*) but only prematurely ended with *A*, realigned with the template via the base-pairing between the extended cap −1 A and template −1 U, and then extended to a full size RNA (blue fonts for mRNAs or cRNAs and red fonts for vRNAs); (*B*) the top 14 U2 caps containing at least 5′AUCGCUUCUC at mRNA +1 with extra nts added via realignment (green), rerealignment (green double-underlined), no realignment (–); (*D*) all the U2 caps containing at least 5′AUCGCUUCUC in vRNA 5′ regions with extra nts added via *trans*-realignment (green), *cis*-realignment (gray), unknown mechanisms (black fonts), and no realignment (–).

priming); the RNA is extended again (second extension); in rare cases, a third round of priming-extension may occur. Since all IAV mRNAs share almost identical 11-nt 5′ UTRs and extra nts, usually 2–4 nt generated by realignment, are copied from 5′ UTRs, technically prime-*cis*-realignment using one vRNA template twice and prime-*trans*-realignment using two vRNA templates each once generate the same results (Fig. 4A,B).

Since the 5′ UTRs of mRNAs and vRNAs bear different sequences (Supplemental Fig. S1A), any *trans*-realignment moving a cap from one vRNA template to another cRNA template can be identified using unique sequence information. Although ~5.6% of the U2 caps in vRNA 5′ regions exhibit recognizable realignment patterns, only ~1.7% were generated by prime-*cis*-realignment using the cRNA template twice (Fig. 4C,D). The top realignment patterns bear the signature sequences (CA, CAAAA, and CAGCAAAA) derived from caps at IAV mRNA +1. Apparently these caps are used to prime mRNA synthesis with a vRNA template, extended but only prematurely terminated with nt A, and then transferred to prime capped vRNA synthesis with a cRNA template (Fig. 4B,D). We estimate that *trans*-realignment contributed ~4% of the U2 caps in the vRNA 5′ regions.

IAV utilizes a more general WG motif for priming capped RNA synthesis

Previous studies have found that cap-snatching at IAV mRNA +1 usually utilizes the base-pairing between the −1 G of a host cap and the −2 C of a template vRNA to prime mRNA synthesis (Geerts-Dimitriadou et al. 2011b; Gu et al. 2015; Koppstein et al. 2015). Therefore, although the +1 G of IAV mRNAs appears to be encoded, it is actually derived from a host cap via priming. To examine whether a specific base-pairing plays a similar role in synthesizing capped RNAs in vRNA 5′ regions and mRNA 3′ clusters, we first divided hybrid RNA reads into two parts, IAV RNA sequences and host caps. 56% and 66% of IAV-encoded parts in mRNA 3′ clusters and vRNA 5′ regions start with G (Supplemental Fig. S5B,C). We speculated that IAV may prefer cleavage sites 3′ of G on host RNAs, generating this +1 G preference on capped IAV RNAs via priming, as in the canonical cap-snatching (Gu et al. 2015). To examine this hypothesis, we mapped the snatched host caps to human genome and analyzed the nt preference surrounding the last (−1) nt of these caps. There is an obvious preference for G immediately after the −1 nt of host caps used in cap-snatching in mRNA 3′ clusters and vRNA 5′ regions (Fig. 5A,B). This suggests that IAV RdRP prefers to cleave host caps 3′ of G and utilizes this G to base pair with a template C, priming RNA synthesis. Since we assigned this priming G to the IAV RNA parts, the host caps losing this G appeared to be cleaved 5′ of G when mapped to human genome (Fig. 5A,B).

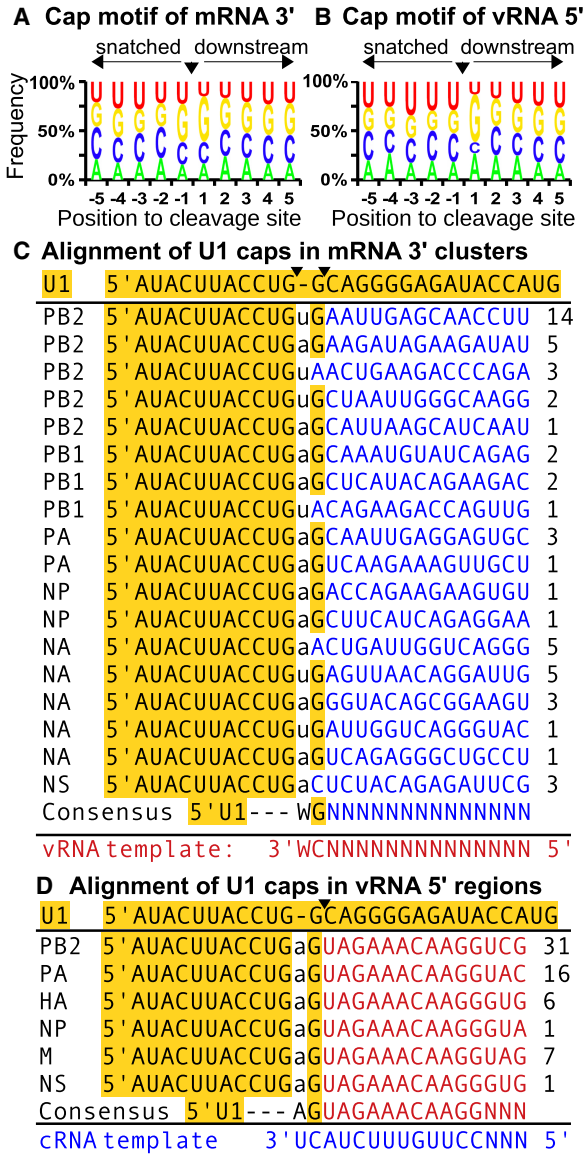

FIGURE 5. The cleavage/priming motif of noncanonical cap-snatching. (*A*,*B*) IAV capped RNA reads derived from mRNA 3′ clusters and vRNA 5′ regions were mapped to human genome and the nt frequency of each position (*x*-axis) surrounding the last nt (−1, the reference point) of the host caps was displayed (*y*-axis) with "▼" indicating the "apparent" cut site; (*C*,*D*) all the capped RNA reads containing the U1 sequence 5′AUACUUACCUG were mapped/aligned to each IAV mRNA 3′ cluster and vRNA 5′ region with yellow indicating the U1 derived sequence in U1 snRNA (*top*) and each hybrid capped RNA read, blue and red indicating IAV mRNA/cRNA and vRNA sequences, respectively, lowercase indicating virtual nts not transcribed but converted from template nts, W representing A/U, the *last* column representing the read number, and the arrows indicating the cap cleavage sites.

We then examined which C's on the template RNAs were selected for the cap-mediated priming. For synthesizing capped RNAs in mRNA 3′ clusters, multiple C's on template vRNAs were used (Fig. 5C; Supplemental Fig. S6A). For synthesizing capped RNAs in vRNA 5′ regions, cRNA templates have two C's, −2 and −4, and almost all priming events occur at −2. For example, 100% of U1 caps and 99% of U2 caps utilized the template −2 C for priming and only 0.7% of U2 caps utilized the template −4 C (Fig. 5D; Supplemental Fig. S6B). This template −2 C preference also occurred in canonical cap-snatching at mRNA +1 since mRNA synthesis was predominantly primed using the cap −1 G almost exclusively with the template −2 C instead of the −4 and −9 C's (Geerts-Dimitriadou et al. 2011a,b; Gu et al. 2015; Koppstein et al. 2015).

The priming events during IAV capped RNA synthesis prefer U or A (collectively as W) followed by G or WG. In IAV mRNA 5′ regions, at least 95% of cap-snatching utilized the −2 C of template vRNAs for initial priming events (Gu et al. 2015), generating the start nt G preceded by a virtual nt A, which is not transcribed but converted from the −1 U of template vRNAs. In other words, the last 2 nt of vRNAs, 3′UC-5′, serve as the template for making 5′AG-3′ in at least 95% of IAV mRNAs, in which G is the first nt and A is the virtual nt upstream. In vRNA 5′ regions, ~66% of capped RNAs start with G, and ~92% of the virtual nt upstream of the start nt are A's, which are converted from template cRNAs (Supplemental Fig. S5C). Based on these two observations, we concluded that the initial priming events prefer an AG motif, in which G is the first nt and A is the virtual nt. Because the 5′ ends of vRNAs and mRNAs are similar, this AG motif is deduced from a very limited sequence diversity and thus may not represent an authentic motif. The mRNA 3′ clusters contain sufficient sequence diversities, allowing us to obtain a more general rule. We found that the priming events in mRNA 3′ cluster also prefer a (A/U)G or WG (W representing A/U) motif since 56% of capped RNAs starts with G, and 56% and 25% of the virtual nt upstream of the start nt are A and U, respectively (Supplemental Fig. S5B). For example, the U1/U2 caps mapped to IAV mRNA 3′ clusters clearly utilize this WG motif (Fig. 5C; Supplemental Fig. S6A). Since the WG motif is based on multiple loci and includes the AG motif, it represents a more general rule for cap-mediated initial priming events.

Cap-snatching likely generates mRNA encoding truncated or new IAV proteins

We speculate that at least a fraction of the noncanonical capped RNAs mapped to IAV mRNA strands, mostly in the 3′ clusters, bear all the required structure elements of translation-capable mRNAs. First, the caps are stolen from Pol II transcripts, most of which are used for translation; second, these capped RNAs likely contain a poly(A) tail (Supplemental Fig. S4); third, the caps or internal sequences of annotated IAV mRNAs may provide a start AUG; fourth, since the consensus Kozak sequence (A/G)

NNAUGG is very short, 12.5% of any given AUG-containing sequence bears a Kozak motif for translation (Kozak 1986, 1987; Cavener 1987; Hamilton et al. 1987). Considering all these structure requirements, we developed a custom Perl script to predict potential coding frames on noncanonical capped RNAs mapped to IAV mRNA strands. We found several capped RNAs potentially encoding 28 proteins, all of which are composed of at least 50 amino acids and 12 of which have a size of at least 200 amino acids (Supplemental Table S1). The 5′ UTR size is usually less than 50 nt and each protein may be coded by several capped RNAs with 5′ UTRs of various sizes (Supplemental Table S1). Among these proteins, the host caps provide the start AUG for nine proteins ("hybrid" in column 2 of Supplemental Table S1), generating 1–3 amino acids, while IAV RNAs provide the start AUG for the rest.

Some of these proteins are encoded in capped RNAs upstream of mRNA 3′ clusters but well downstream from annotated IAV mRNAs (Fig. 2). For example, one internal AUG of PA can be used by several capped RNAs to encode a protein composed of 343 amino acids (Fig. 6). These RNAs are likely authentic capped RNAs generated by cap-snatching since: (1) the cap size is ~11 nt; (2) most of them use a G/C base-pairing-mediated priming; (3) most host caps start with A, a signature start nt of host caps (Gu et al. 2015); and (4) the same IAV locus can have multiple caps derived from different host RNAs (Fig. 6).

Capped RNAs mapped to vRNA 5′ regions likely belong to ncRNAs. Although its size could be as long as that of vRNAs, we can only identify five short coding frames, all of which have long 5′ and 3′ UTRs (Supplemental Table S1). Moreover, we do not know whether the capped RNAs contain a poly(A) tail, and if so, how the tail is added. Therefore, we propose that capped RNAs in vRNA 5′ regions may serve as ncRNAs instead of mRNAs.

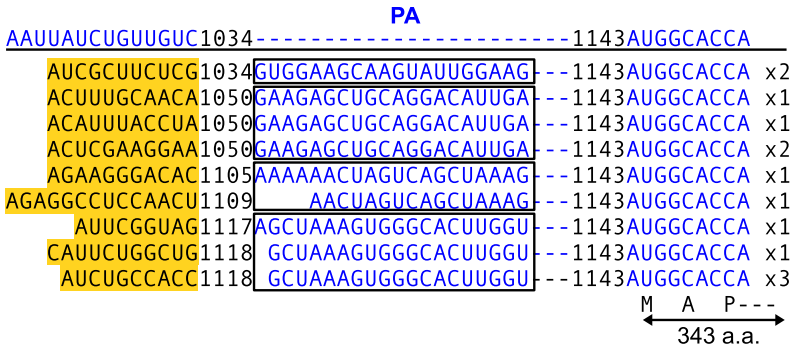

FIGURE 6. A coding frame is shared by multiple capped RNAs. Multiple capped RNA reads mapped to an internal position of IAV PA mRNA were aligned with yellow, boxed blue, and non-boxed blue indicating the host caps, 5′ UTRs, and shared coding frame, respectively. The start positions of the 5′ UTRs and coding frames are labeled in the sequences, the read number is labeled on the *right*, and the encoded amino acids are labeled at the *bottom*.

DISCUSSION

We provide multifaceted evidence to demonstrate that IAV utilizes noncanonical cap-snatching to generate capped RNAs. We found that both canonical and noncanonical cap-snatching share similar mechanisms. However, noncanonical cap-snatching bears unique features including infrequent realignment events and a more general priming motif, WG (Fig. 7). We also found that cap-snatching promotes the diversity of IAV mRNAs and ncRNAs.

Noncanonical cap-snatching is authentic. We demonstrated that noncanonical and canonical cap-snatchings share several signature features including: (1) the median size of snatched caps is ~11 nt; (2) U1 and U2 are the top cap donors totally contributing 7%–8% of snatched caps; (3) a G/C base pair and WG motif are preferred in the initial priming step; (4) the cap-mediated priming prefers the −2 C of template RNAs for synthesizing capped RNAs both at mRNA +1 and in vRNA 5′ regions (Fig. 7). In addition, we provided three negative internal controls, including vRNA +1, non-5′ regions of vRNAs, and host mRNA +1, to demonstrate that cloning artifacts barely generate any reads containing host caps. Were caps added via cloning artifacts, we would expect: (1) no G/C base-pairing preference and no WG motif (Li et al. 2020); (2) a random size distribution of caps; (3) cap addition to the +1 position of host mRNAs and IAV vRNAs since they represent more than 80% of the cloned reads; (4) no preference for template −2C. In conclusion, our data clearly support that noncanonical cap-snatching is as authentic as canonical cap-snatching.

Cap-snatching utilizes a more general WG motif to prime the initial RNA synthesis. IAV RdRP utilizes cap-independent and dependent manners to synthesize RNAs (Desselberger et al. 1980; Te Velthuis and Oymans 2018). Unlike DNA polymerases, most RNA polymerases do not require primers for initiating RNA synthesis. Although IAV RdRP appears to utilize a host cap to prime capped RNA synthesis, the single-nt base-pairing is technically equal to de novo synthesizing RNAs using a G nt modified by a capped RNA oligo in a primer-independent manner. In the cap-independent mode, IAV RdRP synthesizes ppp-vRNAs using the −1 U of template cRNAs and *vice versa*; in the cap-dependent mode, IAV RdRP primarily utilizes the −2 C of template vRNAs to synthesize mRNAs (Fig. 7). Here we demonstrate that IAV RdRP also utilizes the −2 C of template cRNAs to synthesize capped RNAs in vRNA 5′ regions, basically

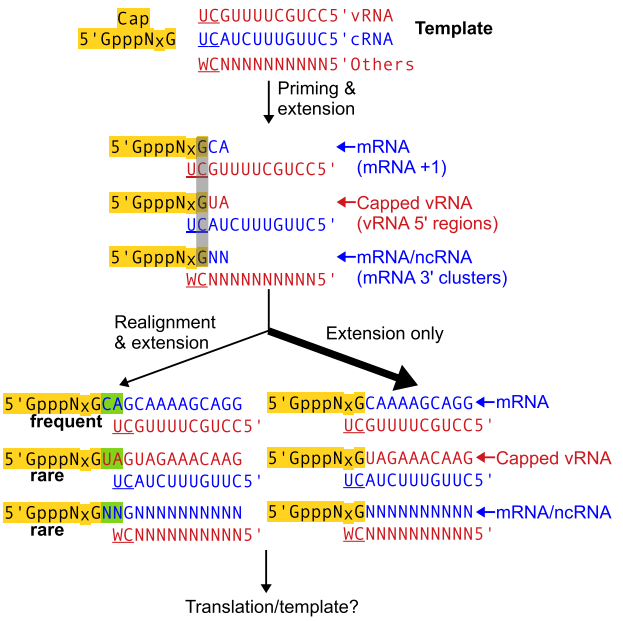

FIGURE 7. A unified model for canonical and noncanonical cap-snatching. Host caps are annealed to IAV RNA templates using the base-pairing between the cap −1 G and template −2 C in the initial priming step; the majority of host caps are extended to make full-size IAV RNAs; a small fraction of host caps are extended for a few nts usually ending with "A," realigned via the base-pairing between the −1 A of the extended sequences and the template −1 U, and extended again to generate full-size mRNAs and ncRNAs. W represents "A" or "U."

establishing a symmetry in synthesizing capped/ppp-RNAs using vRNA and cRNA templates (Fig. 7). However, this symmetry is imperfect since vRNA templates are predominantly for synthesizing capped mRNAs, while cRNA templates are predominantly for synthesizing ppp-vRNAs. In both cases, the level of the minor RNA species is only ~1%–2% of that of the major species. In summary, the template −2 C is preferentially used for synthesizing capped RNAs in both cases, and the template −1 U is preferentially used for synthesizing ppp-RNAs. Moreover, we demonstrate that IAV RdRP also preferentially utilizes the template −2 C, defined as a relative position, to synthesize capped RNAs in mRNA 3′ clusters (Fig. 7).

We demonstrate that the cleavage site preference on host caps likely plays a critical role in preferentially selecting template sites (C nt), since: (1) previous studies have demonstrated that cap-snatching preferentially cuts host caps 3′ of G; (2) our data show the cleavage site preference, 50%, for 3′ of G on host caps (Fig. 5A,B), is very close to ~60%, the start nt preference for G in IAV mRNA 3′ clusters and vRNA 5′ regions (Supplemental Fig. S5B,C).

The nts 3′ of template C's affect the selection of C sites for the initial priming events (Fig. 7). Based on canonical and noncanonical cap-snatchings, we obtained a general motif 3′WC5′ on template RNAs, corresponding to a 5′WG3′ motif on capped RNAs, in which "W" represents A or U encoded by templates but not expressed. This WG motif is not caused by realignment since realignment is infrequent (<5%) in mRNA 3′ clusters while ~50% of priming events utilize this motif. Theoretically this motif can be used to develop a novel therapeutic strategy.

Realignment occurs much less frequently in noncanonical cap-snatching. Realignment or rerealignment events constitute ~20% cap-snatching with recognizable sequence patterns at IAV mRNA +1. In contrast, realignment events with recognizable sequence patterns only constitute ~1.5% of cap-snatching in noncanonical regions. This suggests that IAV RdRP may use different modes to synthesize capped RNAs at different loci. Or this discrepancy is caused by template sequence differences, that is, 3′UCGUUUUCG5′ for mRNA +1, 3′UCAUCUUUG5′ for vRNA 5′ regions, and 3′WCNNNNNNN5′ for mRNA 3′ clusters (Fig. 7). A sequence swap assay may help address this hypothesis.

Trans-realignment utilizes IAV-derived caps. It has been hypothesized that IAV cap-snatching does not target IAV mRNAs as cap donors (Shih and Krug 1996). However, we clearly show that some caps utilized in vRNA 5′ regions are likely derived from IAV mRNAs based on the specific realignment patterns. Here we propose two models. In model 1, cap-snatching targets both host and IAV capped RNAs as cap donors. As a simple and straightforward model, it has serious caveats since (1) IAV mRNAs are exported to cytoplasm for translation, generating a physical barrier for cap-snatching; and (2) evidence showed that IAV RdRP does not target its own mRNAs (Shih and Krug 1996). In model 2, the realignment process may fail to prime with the same template RNAs, jumping onto a second template, a process called "*trans*-realignment." We prefer this model because it is well known that RNA/DNA polymerases often stall on promoters, generating PASRs (Seila et al. 2008; Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Nechaev et al. 2010; Gu et al. 2012). IAV RdRP may bear the same feature, generating partially extended caps and then falling off templates. These caps usually anneal back with the −1 U of the same template RNAs simply because (1) the initial priming/extension usually ends with A due to four template U's within the −2 to −7 positions of template vRNAs; (2) the selection of the same template RNA is due to physical proximity. These caps may be used to prime with another template at a much lower frequency, as we observed. We detected this phenomenon simply because the two templates, cRNAs and vRNAs, bear different 5′ sequences. If *trans*-realignment were to occur between two cRNA or two vRNA templates, the result would appear as *cis*-realignment on the same templates.

Noncanonical cap-snatching diversifies IAV mRNAs and ncRNAs. We demonstrate that in mRNA 3′ clusters, cap-

snatching generates capped RNAs bearing all the features of functional mRNAs including a cap, poly(A) tail, start AUG and Kozak motif. Actually, we can identify more translation-capable capped RNAs when we include other mRNA regions or relax the Kozak motif requirement. In many cases, host caps provide a start AUG and a coding frame for 1–3 amino acids. Although noncanonical cap-snatching only generates ~1% of capped RNAs mapped to IAV mRNA strands, the expression levels of some noncanonical mRNAs may reach to the median level of host mRNAs because IAV mRNA reads constitute 18% of total host/IAV reads and are derived from only eight mRNA strands. Interestingly, we found that on NA mRNAs, the noncanonical capped RNA level reaches ~9% of the canonical mRNA level. In addition, cap-snatching may promote the diversity of IAV mRNAs and ncRNAs via (1) introducing a new AUG and Kozak signal to the internal sequences of annotated mRNAs; and (2) obtaining new coding frames, especially on vRNAs, due to the high mutation rate of IAV RdRP.

In summary, we provide a comprehensive profile of IAV cap-snatching and a more general mode for the priming-realignment mechanism. We also propose that cap-snatching promotes the diversity of IAV mRNAs and ncRNAs. Insights from this study may help better understand the cap-snatching mechanism and design research and therapeutic tools.

MATERIALS AND METHODS

IAV infection

The cell culture and virus infection condition were described previously (Gu et al. 2015). Briefly, A549 cells were incubated with influenza A/Brisbane/59/2007 (H1N1) at a multiplicity of infection 1 at 37°C for 1 h, washed and cultured for 6, 12, 24, and 48 h.

RNA extraction

RNA was extracted from infected cells using TRI reagents (Sigma-Aldrich) according to the manufacturer's protocol. The resulting aqueous solution was phenol/chloroform extracted and coprecipitated with 20 µg glycogen.

Obtaining the 5′ end sequences of host and IAV RNAs

To simultaneously analyze the 5′ end of host and IAV RNAs, high-throughput sequencing libraries were constructed using CapSeq and then sequenced (Gu et al. 2015). In brief, 2 µg of total RNA was processed using Terminator exonuclease (Epicentre) to remove rRNAs and decapped using Tobacco Acid Pyrophosphatase (Epicentre). The linkers required for high-throughput sequencing were added to the 5′ end of target RNAs using ligation and to the 3′ end using random priming in reverse transcription. We obtained cDNA containing 50–200-nt RNA inserts and sequenced the first 50 or 100 nt using HiSeq 2000. The data were stored at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=apwnwimmvnupfkr&acc=GSE67493 (Gu et al. 2015).

Bioinformatic analyses

The RefSeq IAV sequences lack parts of the 5′ and 3′ UTRs, which are critical for our analyses (Pruitt et al. 2014; Sikora et al. 2014). We used a custom Perl script to assemble the 5′ UTRs of IAV mRNAs, cRNAs, and vRNAs based on our CapSeq data and then obtained the corresponding 3′ UTRs using reverse complement, as shown in Supplemental Figure S1A.

The bioinformatic analyses were performed using custom Perl scripts and Bowtie 0.12.7, as described previously (Langmead et al. 2009; Gu et al. 2015). Since the previous analyses focused on the 5′ end of mRNAs, we modified the scripts to fit the analyses of vRNA 5′ regions and mRNA 3′ clusters. In brief, we mapped the first 40-nt sequence of each read to human genome and annotations, the Ensembl GRCh37 release 71 (Zerbino et al. 2018). We obtained the unmatched reads and then split them into two parts, position 1–20 and 21–40. The latter was mapped to IAV RNAs, and the resulting full-size match was extended toward the 5′ end of the reference sequence using the position 20 −1 of the read with a score +1 and −3 for each match and mismatch, respectively. The longest extension was selected using the maximum score, and the sequence 5′ of the matched part was extracted as a non-IAV sequence. We also removed the most frequent extension stops caused by indels due to sequencing errors or mutations. Gbrowse was used to generate histograms of IAV reads (Stein et al. 2002). We used a custom Perl script to predict the coding frame encoded by noncanonical capped RNAs with the criteria: (1) it contains at least 50 amino acids; (2) there is a Kozak motif (A/G)NNAUGG in which AUG is the start codon encoded by IAV or host caps; (3) a poly (A) tail can be added using the "stuttering" mechanism (Luo et al. 1991; Pritlove et al. 1998; Poon et al. 1999; Zheng et al. 1999). The whole pipeline is available at https://github.com/guweifengucr/WG060519_capseq_flu_analysis.git.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work is supported by the University of California.

Received November 1, 2019; accepted May 12, 2020.

REFERENCES

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. 2009. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457:** 1028–1032. doi:10.1038/nature07759

Beaton AR, Krug RM. 1981. Selected host cell capped RNA fragments prime influenza viral RNA transcription in vivo. *Nucleic Acids Res* **9:** 4423–4436. doi:10.1093/nar/9.17.4423

Bouloy M, Plotch SJ, Krug RM. 1978. Globin mRNAs are primers for the transcription of influenza viral RNA *in vitro*. *Proc Natl Acad Sci* **75:** 4886–4890. doi:10.1073/pnas.75.10.4886

Bouvier NM, Palese P. 2008. The biology of influenza viruses. *Vaccine* **26:** D49–D53. doi:10.1016/j.vaccine.2008.07.039

Caton AJ, Robertson JS. 1980. Structure of the host-derived sequences present at the 5′ ends of influenza virus mRNA. *Nucleic Acids Res* **8:** 2591–2603. doi:10.1093/nar/8.12.2591

Cavener DR. 1987. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res* **15:** 1353–1361. doi:10.1093/nar/15.4.1353

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88:** 127–131. doi:10.1016/j.ygeno.2005.12.013

Datta K, Wolkerstorfer A, Szolar OHJ, Cusack S, Klumpp K. 2013. Characterization of PA-N terminal domain of Influenza A polymerase reveals sequence specific RNA cleavage. *Nucleic Acids Res* **41:** 8289–8299. doi:10.1093/nar/gkt603

Decroly E, Ferron F, Lescar J, Canard B. 2011. Conventional and unconventional mechanisms for capping viral mRNA. *Nat Rev Microbiol* **10:** 51–65. doi:10.1038/nrmicro2675

Desselberger U, Racaniello VR, Zazra JJ, Palese P. 1980. The 3′ and 5′-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene* **8:** 315–328. doi:10.1016/0378-1119(80)90007-4

Dhar R, Chanock RM, Lai CJ. 1980. Nonviral oligonucleotides at the 5′ terminus of cytoplasmic influenza viral mRNA deduced from cloned complete genomic sequences. *Cell* **21:** 495–500. doi:10.1016/0092-8674(80)90486-9

Dias A, Bouvier D, Crepin T, McCarthy AA, Hart DJ, Baudin F, Cusack S, Ruigrok RW. 2009. The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458:** 914–918. doi:10.1038/nature07745

Fodor E, Crow M, Mingay LJ, Deng T, Sharps J, Fechter P, Brownlee GG. 2002. A single amino acid mutation in the PA subunit of the influenza virus RNA polymerase inhibits endonucleolytic cleavage of capped RNAs. *J Virol* **76:** 8989–9001. doi:10.1128/JVI.76.18.8989-9001.2002

Geerts-Dimitriadou C, Goldbach R, Kormelink R. 2011a. Preferential use of RNA leader sequences during influenza A transcription initiation *in vivo*. *Virology* **409:** 27–32. doi:10.1016/j.virol.2010.09.006

Geerts-Dimitriadou C, Zwart MP, Goldbach R, Kormelink R. 2011b. Base-pairing promotes leader selection to prime *in vitro* influenza genome transcription. *Virology* **409:** 17–26. doi:10.1016/j.virol.2010.09.003

Gu W, Shirayama M, Conte D, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, et al. 2009. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol Cell* **36:** 231–244. doi:10.1016/j.molcel.2009.09.020

Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151:** 1488–1500. doi:10.1016/j.cell.2012.11.023

Gu W, Gallagher GR, Dai W, Liu P, Li R, Trombly MI, Gammon DB, Mello CC, Wang JP, Finberg RW. 2015. Influenza A virus preferentially snatches noncoding RNA caps. *RNA* **21:** 2067–2075. doi:10.1261/rna.054221.115

Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, Sehr P, Lewis J, Ruigrok RW, Ortin J, Hart DJ, et al. 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat Struct Mol Biol* **15:** 500–506. doi:10.1038/nsmb.1421

Hagen M, Tiley L, Chung TD, Krystal M. 1995. The role of template–primer interactions in cleavage and initiation by the influenza virus polymerase. *J Gen Virol* **76:** 603–611. doi:10.1099/0022-1317-76-3-603

Hainer SJ, Gu W, Carone BR, Landry BD, Rando OJ, Mello CC, Fazzio TG. 2015. Suppression of pervasive noncoding transcription in embryonic stem cells by esBAF. *Genes Dev* **29:** 362–378. doi:10.1101/gad.253534.114

Hamilton R, Watanabe CK, de Boer HA. 1987. Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Res* **15:** 3581–3593. doi:10.1093/nar/15.8.3581

Kobayashi M, Toyoda T, Ishihama A. 1996. Influenza virus PB1 protein is the minimal and essential subunit of RNA polymerase. *Arch Virol* **141:** 525–539. doi:10.1007/BF01718315

Koppstein D, Ashour J, Bartel DP. 2015. Sequencing the cap-snatching repertoire of H1N1 influenza provides insight into the mechanism of viral transcription initiation. *Nucleic Acids Res* **43:** 5052–5064. doi:10.1093/nar/gkv333

Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292. doi:10.1016/0092-8674(86)90762-2

Kozak M. 1987. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15:** 8125–8148. doi:10.1093/nar/15.20.8125

Krug RM, Broni BA, Bouloy M. 1979. Are the 5′ ends of influenza viral mRNAs synthesized in vivo donated by host mRNAs? *Cell* **18:** 329–334. doi:10.1016/0092-8674(79)90052-7

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi:10.1186/gb-2009-10-3-r25

Li L, Dai H, Nguyen AP, Gu W. 2020. A convenient strategy to clone small RNA and mRNA for high-throughput sequencing. *RNA* **26:** 218–227. doi:10.1261/rna.071605.119

Luo GX, Luytjes W, Enami M, Palese P. 1991. The polyadenylation signal of influenza virus RNA involves a stretch of uridines followed by the RNA duplex of the panhandle structure. *J Virol* **65:** 2861–2867. doi:10.1128/JVI.65.6.2861-2867.1991

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327:** 335–338. doi:10.1126/science.1181421

Pflug A, Lukarska M, Resa-Infante P, Reich S, Cusack S. 2017. Structural insights into RNA synthesis by the influenza virus transcription-replication machine. *Virus Res* **234:** 103–117. doi:10.1016/j.virusres.2017.01.013

Plotch SJ, Bouloy M, Krug RM. 1979. Transfer of 5′-terminal cap of globin mRNA to influenza viral complementary RNA during transcription *in vitro*. *Proc Natl Acad Sci* **76:** 1618–1622. doi:10.1073/pnas.76.4.1618

Plotch SJ, Bouloy M, Ulmanen I, Krug RM. 1981. A unique cap (m^7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell* **23:** 847–858. doi:10.1016/0092-8674(81)90449-9

Poon LL, Pritlove DC, Fodor E, Brownlee GG. 1999. Direct evidence that the poly(A) tail of influenza A virus mRNA is synthesized by reiterative copying of a U track in the virion RNA template. *J Virol* **73:** 3473–3476. doi:10.1128/JVI.73.4.3473-3476.1999

Pritlove DC, Poon LLM, Fodor E, Sharps J, Brownlee GG. 1998. Polyadenylation of influenza virus mRNA transcribed in vitro from model virion RNA templates: requirement for 5′ conserved sequences. *J Virol* **72:** 1280–1286. doi:10.1128/JVI.72.2.1280-1286.1998

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM,

et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42:** D756–D763. doi:10.1093/nar/gkt1114

Rao P, Yuan W, Krug RM. 2003. Crucial role of CA cleavage sites in the cap-snatching mechanism for initiating viral mRNA synthesis. *EMBO J* **22:** 1188–1198. doi:10.1093/emboj/cdg109

Reich S, Guilligay D, Pflug A, Malet H, Berger I, Crépin T, Hart D, Lunardi T, Nanao M, Ruigrok RWH, et al. 2014. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516:** 361–366. doi:10.1038/nature14009

Saira K, Lin X, DePasse JV, Halpin R, Twaddle A, Stockwell T, Angus B, Cozzi-Lepri A, Delfino M, Dugan V, et al. 2013. Sequence analysis of *in vivo* defective interfering-like RNA of influenza A H1N1 pandemic virus. *J Virol* **87:** 8064–8074. doi:10.1128/JVI.00240-13

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322:** 1849–1851. doi:10.1126/science.1162253

Shaw MW, Lamb RA. 1984. A specific sub-set of host-cell mRNAs prime influenza virus mRNA synthesis. *Virus Res* **1:** 455–467. doi:10.1016/0168-1702(84)90003-0

Shi L, Summers DF, Peng Q, Galarz JM. 1995. Influenza A virus RNA polymerase subunit PB_2 is the endonuclease which cleaves host cell mRNA and functions only as the trimeric enzyme. *Virology* **208:** 38–47. doi:10.1006/viro.1995.1127

Shih SR, Krug RM. 1996. Surprising function of the three influenza viral polymerase proteins: selective protection of viral mRNAs against the cap-snatching reaction catalyzed by the same polymerase proteins. *Virology* **226:** 430–435. doi:10.1006/viro.1996.0673

Sikora D, Rocheleau L, Brown EG, Pelchat M. 2014. Deep sequencing reveals the eight facets of the influenza A/HongKong/1/1968 (H3N2) virus cap-snatching process. *Sci Rep* **4:** 6181. doi:10.1038/srep06181

Sikora D, Rocheleau L, Brown EG, Pelchat M. 2017. Influenza A virus cap-snatches host RNAs based on their abundance early after infection. *Virology* **509:** 167–177. doi:10.1016/j.virol.2017.06.020

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The Generic Genome Browser: a building block for a model organism system database. *Genome Res* **12:** 1599–1610. doi:10.1101/gr.403602

Sugiyama K, Obayashi E, Kawaguchi A, Suzuki Y, Tame JR, Nagata K, Park SY. 2009. Structural insight into the essential PB1–PB2 subunit contact of the influenza virus RNA polymerase. *EMBO J* **28:** 1803–1811. doi:10.1038/emboj.2009.138

Te Velthuis AJW, Oymans J. 2018. Initiation, elongation, and realignment during influenza virus mRNA synthesis. *J Virol* **92:** e01775-17. doi:10.1128/JVI.01775-17

Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, He X, Lv Z, Ge R, Li X, Deng T, et al. 2009. Crystal structure of an avian influenza polymerase PA_N reveals an endonuclease active site. *Nature* **458:** 909–913. doi:10.1038/nature07720

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46:** D754–D761. doi:10.1093/nar/gkx1098

Zheng H, Lee HA, Palese P, García-Sastre A. 1999. Influenza A virus RNA polymerase has the ability to stutter at the polyadenylation site of a viral RNA template during RNA replication. *J Virol* **73:** 5240–5243. doi:10.1128/JVI.73.6.5240-5243.1999