The Prime-and-Realign Process of the Influenza A Virus Occurs to Rescue Cap-Snatched Primers on the Basis of Length and RNA Duplex Stability

Corey De Vlugt

Dr. Martin Pelchat (Supervisor)

A thesis submitted in partial fulfillment of the requirements for the Master's degree in Biochemistry

Department of Biochemistry, Microbiology and Immunology

Faculty of Medicine

University of Ottawa

Abstract

Cap-snatching by the influenza A virus (IAV) RNA-dependant RNA polymerase (RdRp) is driven by the abundance of transcripts being actively transcribed by host RNA polymerase II (Pol II)[1]–[3]. Deviations from a direct correlation with abundance do arise, due to selective cleavage of transcripts with a compatible length (10 to 13 nucleotides) and nucleotide sequence (ending in 3'AG)[4]–[7]. Some cap-snatched primers are not directly used to transcribe mRNA, but instead undergo a prime-and-realign mechanism (PAR). As of yet it is unknown why this process occurs. **My hypothesis is that the prime-and-realign process is related to the physical characteristics of the primers and their interactions with RdRp and the vRNA template.** Here, I used four published deep sequencing datasets of the 5' ends of IAV mRNA obtained from IAV infected A549 cells to examine PAR[1], [7]–[9]. Primers are biased towards PAR on the basis of length (<12 nucleotides) and RNA duplex stability (mediated by the base directed at 3'U1 and the pyrimidine-purine base pair at position four). PAR typically adds a GCA addition resulting in a primer three nucleotides longer ending in a compatible nucleotide sequence with 3'U1. Prime-and-realign converts poor primers on the basis of length and sequence compatibility with the 3' end of the vRNA into one that can efficiently undergo transcription of the critical conserved sequence without errors, or failure. Prime-and-realign, therefore, affords tremendous flexibility to RdRp in cap snatched primer length and sequence compatibility.

Acknowledgements

I would like to thank my thesis supervisor Dr. Martin Pelchat for giving me the opportunity to work on this project, teaching me how to use R, helping to guide this project through to completion, and assisting in the preparation of my master's thesis. I couldn't have asked for a better supervisor.

I would also like to thank my Thesis Advisory Committee members Dr. Tommy Alain and Dr. Mathieu Lavallée-Adam for their helpful guidance. In particular Dr. Mathieu Lavallée-Adam introduced me to the concept of Monte Carlo simulations as a statistical test for big data. I would also like to thank Dorota Sikora and Lynda Rocheleau for the work they did in generating the mRNA data sets which I used for this project. I would also like to thank Gabrielle Goodrum and Lynda Rocheleau for their helpful comments on my practice presentations.

I would like to thank the people who took the time to proof read my thesis, in particular Montana McLaughlin-Tom and Victoria Ajila who proof read and helped with editing of this document.

On a personal note, I would also like to thank my parents Robert and Evelyn De Vlugt, my sister Carmela De Vlugt, and my friends, for their support as completed my Master's degree requirements, and I would like to thank God for helping me through my academic endeavors.

Table of Contents

1. Introduction 1
1.1 Influenza A Virus Genome and Proteins 1
1.2 Influenza A Virus Life Cycle 8
1.2.1 Virion Uptake and Entry 8
1.2.2 Transcription and Replication Phase 8
1.2.3 Export of Viral Ribonucleoproteins and Release of Progeny Virions 9
1.3 RNA-dependent RNA Polymerase 12
1.3.1 Nuclear Localization and Assembly of RdRp 13
1.3.2 Structure of Trimeric RdRp 14
1.3.3 Polymerase Structure 22
1.3.4 RdRp as a Replicase 24
1.3.5 RdRp as a Polyadenylase 25
1.3.6 RdRp as a Transcriptase: Cap-Snatching and Prime and Realign 28
1.3.6.1 Association with Host DNA-Dependent RNA Polymerase II 32
1.3.6.2 Polymerase Basic Protein 2 m^7G Cap Binding 34
1.3.6.3 Polymerase Acidic Mediated Endonuclease Cleavage 34
1.3.6.3.1 Length and Sequence Specific Cleavage 36
1.3.6.4 Cap-snatching is Based on Abundance 40
1.3.6.5 Priming Transcription and Prime-and-Realign 43
1.4 Rational and Hypothesis 47
2. Methods 49
2.1 Summary of Previously Obtained Data Set Preparation 49
2.2 Data Set Preparation by Sikora *et al*. 50
2.3 Pre-Processing of Obtained Data 54
2.4 Primer Classification 58
2.5 Prime-and-Realign Frequency Determinations 58
2.6 Prime-and-Realign Rounds 58
2.7 Primer 3’ Sequence Ends and the 3’C2 Directed Guanine 59
2.8 Primer Length Determinations 61

2.9 Nucleotide Sequence Additions 63
2.10 Free Energy Determinations 65
2.11 Relative Guanine and Cytidine content 71
2.12 Effect Modeling 72
3. Results 74
3.1 Prime and Realign Frequencies 74
3.2 Prime and Realign Rounds 78
3.3 Complementarity between the Primer and Template Sequences 84
3.4 Length of Cap Snatched Primers 105
3.5 Length and Nucleotide Sequence Additions 118
3.6 Duplex Stability from Bond Energy and Stacking Interactions 128
3.7 3’ vRNA Conserved Sequences 135
3.8 Conserved Sequences and Nucleotide Sequence Additions 143
3.9 Primer Sequence End 151
3.10 Relative Guanine and Cytidine Content 161
3.11 Effect Modeling 169
4. Discussion 174
4.1 PAR Frequency 174
4.2 PAR was not Related to the G+C Content of Cap Snatched mRNA 175
4.3 Length of Cap Snatched Primers 176
4.4 Free Energy of the RNA Duplex 179
4.5 Free Energy and Sequence Specific Cleavage 180
4.6 Free Energy and the vRNA Conserved Sequences 188
4.7 The Prime and Realign Process 194
5. Conclusion 201
6. References 202

I. Abbreviations

RNA – ribonucleic acid

IAV – Influenza A Virus

PAR – prime-and-realign

Puerto Rico – A/PuertoRico/8/1934 (H1N1)

Hong Kong – A/HongKong/1/1968 (H3N2)

WSN – A/WSN/1933 (H1N1)

Brisbane – A/Brisbane/59/2007 (H1N1)

H#N# – hemagglutinin and neuraminidase

PB2 – polymerase basic 2

PB1 – polymerase basic 1

PA – polymerase acidic

RdRp – RNA-dependent RNA polymerase

HA – hemagglutinin

NP – nucleoprotein

NA – neuraminidase

M1 – matrix protein 1

M2 – matrix protein 2

NS1 – nonstructural protein 1

NEP – nuclear export protein

vRNA – viral RNA

vRNP – viral ribonucleoprotein

cRNA – complimentary RNA

cRNP – complimentary ribonucleoprotein

mRNA – messenger RNA

Pol II – DNA-dependent RNA polymerase II

NLS – nuclear localization sequence

CAT - chloroamphenicol acetyltransferase

CTD – C-terminal domain

ChiP – chromatin immunoprecipitation

RGSV – rice grassy stunt virus

RSV – rice stripe virus

UTR – untranslated region

SNV – Sin Nombre Virus

A – adenine

C – cytidine

G – guanine

U – uridine

T – thymidine

Y – pyrimidine

R – purine

N – nucleotide

nt – nucleotide

II. Table of Figures

Figure 1.1 Simplified Influenza A Virus Life Cycle 11

Figure 1.2. Cartoon Representation of RdRp bound the vRNA Promoter and Priming Transcription 19

Figure 1.3. IAV vRNA Promoter Structure in the Corkscrew Conformation 21

Figure 2.1 Matched Trimming Flowchart 53

Figure 3.2.1 Loge Distribution of Realigned Primers at a Given Number of Rounds of PAR 83

Figure 3.3.1. Nucleotide Enrichment 3' and 5' of the End of the Matched Sequence and Frequency Distributions of PAR relative to presence of a Complimentary Guanine Residue 92

Figure 3.4.1. Length of Cap Snatched Primers 109

Figure 3.4.2. PAR Frequency at a given Length 113

Figure 3.4.3. Log Linear relation of PAR Frequency and Primer Length 115

Figure 3.4.4. Log Linear relation of PAR Frequency and Primer Length between Nine and 13 Nucleotides 117

Figure 3.5.1. Sequence Additions at all Start and End Lengths 123

Figure 3.5.2. Loge Transformed Frequency of Nucleotide Sequence Additions of a Given Length 125

Figure 3.5.3. Sequence Additions at all Start and End Lengths at a Given Length 127

Figure 3.6.1. Free Energy ($\boldsymbol{\Delta G37o}$) of Duplex Structure formed after a Three Nucleotide Sequence Addition Length 132

Figure 3.7.1. Frequency Distributions of PAR relative to IAV Subunit Transcribed and vRNA Conserved Sequences 140

Figure 3.8.1. Nucleotide Sequence Additions on a Given Viral Conserved Sequence 150

Figure 3.9.1. Nucleotide Enrichment and Free Energy When Priming Transcription 156

Figure 3.9.2. Free Energy of the 3'U1 Directed Nucleotide and PAR Frequencies 158

Figure 3.9.3. Free Energy of the 3'U1 Directed Nucleotide and PAR Frequencies Factoring in the 3'C2 directed Guanine 160

Figure 3.10.1. Relative Guanine and Cytidine Residue Enrichment of the Cap-Snatched Genome, No Realignment, Pre-Realignment, and Post-Realignment Subpopulations 164

Figure 3.10.2. Frequency Distributions of PAR Relative to Guanine and Cytidine Content 168

Figure 3.11.1. PAR Frequencies Based on Length and Nucleotide Sequence used to Prime Transcription 173

Figure 4.1. Graphical Model of the Determinants of Prime-and-Realign 200

III. Table of Tables

Table 1.1 Influenza A Virus Genome, Proteins, and Function 4

Table 1.2. Nucleotide Sequences which may be Added by the Prime and Realign Process 30

Table 2.1. Example of Organization of the Data for the Puerto Rico Data Set with Example Primers 57

Table 3.1.1. Relative Distribution of Cap-Snatched Primers between the Realigned and No Realignment Subpopulations 76

Table 3.2.1. Relative Distribution of the Realigned Subpopulation by Rounds 80

Table 3.3.1. Guanine at +1 Enrichment and Randomly Generated Enrichment 93

Table 3.3.2. Cytidine at -2 Enrichment and Randomly Generated Enrichment 95

Table 3.3.3. Adenine at -1 Enrichment and Randomly Generated Enrichment 97

Table 3.3.4. Uridine at -1 Enrichment and Randomly Generated Enrichment 99

Table 3.3.5. Enrichment of an AG Dinucleotide at the 5’ End of Primers and Randomly Generated Enrichment 101

Table 3.3.6. Enrichment of a CAG Trinucleotide at the 5’ End of Primers and Randomly Generated Enrichment 103

Table 3.4.1. Length Distributions and p-values 110

Table 3.6.1. Free Energy ($\boldsymbol{\Delta G37o}$) Distributions and p-values 133

Table 3.7.1. 3’ 5’ Base Pair Region, Free Energy, and PAR Frequency for the Hong Kong Strain 141

Table 3.8.1 Free Energy from Hydrogen Bonds and Stacking Interactions for PAR additions of Three Nucleotides or Longer on both vRNA Conserved Sequences 147

Table 3.10.1. Relative G+C Content Distributions and p-values 165

1. Introduction

The influenza A virus (IAV) is a highly pathogenic virus, which is responsible for yearly epidemics, with the occasional pandemic. In humans, IAV targets respiratory epithelial cells; in this regard, IAV may induce morbidity or mortality through a variety of mechanisms including cell death, 'cytokine storm' (host immune response), and/or secondary infection[10]. It is estimated that seasonal influenza is responsible for approximately 650,000 deaths annually through respiratory complications[11], with pandemic strains potentially exceeding this mortality rate[10]. The single most fatal event in recorded human history stands as the 1918 Spanish influenza pandemic, which is estimated to have killed upwards of 50 million people[10]. It is estimated that if a similar pandemic were to occur now, in the US alone, it would result in a financial loss of at least 5% of the gross domestic product with an additional patient care cost of approximately $675 billion[12]. For these reasons, IAV remains an important area of research.

1.1 Influenza A Virus Genome and Proteins

The influenza A virus is a negative sense single stranded RNA virus (Group V) of the *Orthomyxoviridae* family. Subtypes of influenza A are categorized based on the type of the two surface antigens present on the virion: hemagglutinin and neuraminidase[13]. The IAV genome consists of eight segmented genes of varying lengths, each of which encodes for at least one protein[13]. A segmented genome is a feature shared with other negative sense single stranded RNA viruses in both the Bunyavirales order and the *Arenaviridae* family.

IAV uses post transcriptional methods to increase the number of proteins encoded for by these eight segments to at least 17 known proteins. Strategies to increase the number of proteins encoded for by a single gene include alternative splicing sites, open read frame shifting, and

ribosomal leaky scanning. The genes and proteins are discussed in Table 1.1. Of these 17 known proteins, only ten are regarded as critical for the IAV life cycle[13], with the other six being referred to as accessory proteins. These ten critical proteins are indicated by an asterisk (*) next to the protein in Table 1.1, and are the full length mRNA of each of the eight genes, as well as the spliced variants of the matrix gene (M2) and nonstructural gene (NEP).

Each individual viral RNA (vRNA) segment is associated with a protein complex. The 3' and 5' ends are associated with RNA-dependent RNA polymerase (RdRp) with the remainder of the gene associated with nucleoprotein (NP). These viral ribonucleoprotein (vRNP) complexes perform both transcription and replication of the viral genome. Replication is a two stage process, with a positive sense intermediate (complimentary RNA; cRNA). cRNA only undergoes replication, and not transcription of mRNA. IAV RNA is negative sense, and requires a positive sense mRNA copy to produce proteins. mRNA is transcribed directly from the vRNA in a cap-dependent manner[14], and omits the nucleotide region beyond a 5' polyuridine track[15].

Genetic variability of influenza A strains arises from two different sources: mutation and reassortment. A single strain may increase the genetic variability through mutations. Mutations arise at between 10^{-4} to 10^{-6} nucleotides per strand copied[16]–[18]. Despite mutation being an important aspect of IAV evolution, many genes have highly conserved regions that are essential for proper protein function. If a mutation increases the fitness of the gene, it will be selected for over the course of multiple viral life cycles[19], whereas mutations that decrease the relative fitness are lost. Mutations are important in overcoming the effects of antiviral drugs, and bypassing host immune defenses, as well as adaptation of a strain to a new host species[19]–[21]. The second method of increasing genetic variability is a mechanism called "reassortment". Reassortment occurs when two or more IAV strains infect the same cell. In such cases genes

from each IAV strain are cross-packaged into the progeny virions[19], [21]; allowing novel strains of IAV to potentially be formed, or genes from one strain can be shared with another[19], [21], [22]. This process is of particular importance to the generation of pandemic strains, as genes from different host species can be incorporated to form a highly pandemic strain, such as the 1957 H2N2 and the 1968 Hong Kong H3N2 flu pandemics[10]. These reassortment events do not always lead to a pandemic. The circulating H1N1 and H3N2 seasonal strains commonly reassort into an H1N2 subtype that is unlikely to cause a pandemic[22], as it is a reassortment product of two endemic strains of low pathogenicity. This process is of particular concern in swine which may be infected by both human and avian strains, allowing for highly pathogenic avian IAV genes to be incorporated into a human adapted strain.

Table 1.1 Influenza A Virus Genome, Proteins, and Function

Gene Segment	Length (nt)	Protein	Expression Method	Function
Polymerase Basic 2	2341	PB 2*	Full length mRNA	RdRp component[23] Disrupts mitochondrial antiviral signaling pathway lowering IFN-β[24]–[26] Induces mitochondrial fragmenting[26]
		PB2-S1	Spliced (5'-G1512 to 5'-C1895)	Mitochondrial targeting, may disrupt mitochondrial antiviral signaling pathway[27]
Polymerase Basic 1	2341	PB 1*	Full length mRNA	RdRp component[23]
		PB1-F2	Ribosomal leaky scanning (4th AUG)[28]	Reduces mitochondria membrane potential and disrupts mitochondrial ultrastructure[28] Initiates mitochondria mediated cell death via cytochrome c release[28] Disrupts mitochondrial antiviral signaling pathway lowering IFN-β[29]
		PB1-N40	Ribosomal leaky scanning (5th AUG)[30]	Unknown; loss of expression lowers IAV growth[30]
Polymerase Acidic	2233	PA*	Full length mRNA	RdRp component[23] Proteolysis[31] (may be a function of PA-X rather than PA) RNA Polymerase II degradation[32]
		PA-X	+1 open read frameshift (5'-UCC UUU CGU C at 192nd codon)[33]	Down regulation of genes related to immune response/inflammation, apoptosis, tissue remodeling, and cell differentiation[33]; important for low pathogenic or seasonal IAV[34] Proteolysis[33] Degradation of pre-mRNA with a polyadenylation signal[35]
		PA-N155	Ribosomal leaky scanning (11th AUG)[36]	Unknown[36]
		PA-N182	Ribosomal leaky scanning (13th AUG)[36]	Unknown[36]

Gene Segment	Length (nt)	Protein	Expression Method	Function
Hemagglutinin	1778	HA*	Full length mRNA; cleaved post-translationally into HA_1 and HA_2; forms a trimer of 3 HA_1 and 3 HA_2[37]–[39]	Principal (~80%) viral membrane protein (spike glycoprotein), present in lipid rafts[40] IAV species are classified based on the Hemagglutinin protein (ex., the "H" in H1N1) Key immune system target against IAV virions[41] HA_1 binds to receptor associated sialic acid residues allowing for the uptake of virions via endocytosis into the endosome[42] Residue at 226 within the receptor binding region confers species specificity; glutamine α2,3-linked sialic acid (avian, swine), leucine α2,6-linked sialic acid (humans, swine)[43] Endosomal membrane fusion, via the HA_2 membrane fusion region, following acidification (pH < 6.0) of the virion core in the endosome[37], [38], [42] Nuclear export of vRNPs via MAPK signaling[44]
Nucleoprotein	1565	NP*	Full length mRNA	Key viral ribonucleoprotein component Binds to viral RNA in a non-sequence specific manner via associations with the phosphate backbone and stacking interactions with the bases[45]–[49] One NP binds to 20 nucleotides[50] Forms oligomers with the long vRNA strands which take on a helical conformation[49], [51] Prevents RNA secondary structure, allowing for efficient transcription via RdRp[46] Nuclear import of vRNPs after release into the cytoplasm from the virion[52], [53] Antitermination factor for replicative transcription[54] Nuclear accumulation required for efficient genome replication[54], [55]
Neuraminidase	1413	NA*	Full length mRNA; forms homotetramers[56], [57]	Second most abundant virion membrane protein (~17%) associated with lipid rafts[40] Cleaves sialic acid, with cleavage affinity being related to hemagglutinin binding affinity[58] Not critical for viral life cycle, but absence lowers viral titer and infectivity[59]

Gene Segment	Length (nt)	Protein	Expression Method	Function
Matrix	1027	M1*	Full length mRNA	Major virion structural protein Associates with virion membrane[60] Binds to RNA in a non-sequence specific manner[61] One M1 binds to 200 nucleotides[61], [62] Selectively binds to vRNA over cRNA to allow for selective packaging and export[63]–[67] Prevents vRNP entry into nucleus[64] Binding is related to pH; a pH below 6 leads to the disassociation of M1 with vRNA, allows for efficient release of vRNA after virion entry[64]
		M2*	Spliced (5'G51 to 5'G740); forms homotetramer[68], [69]	Third and least abundant virion membrane protein[40] Allows for one way flow of protons in a manner that is sensitive to the N-terminus side pH[70]–[72] Allows for acidification of the virion core[70]–[72] in the endosome and thereby fusion of the virion of membrane and the endosome membrane[37], [38], [42] and release of vRNP from M1[64] Prevents premature deployment of the HA fusion peptide during trafficking to the membrane[72]
Non Structural	890	NS1*	Full length mRNA	Non critical for viral life cycle and function[73] Selective translation of viral mRNA[74] though interactions with eIF4GI, PABP1, and hStaufen[75]–[77] Retains polyadenylated mRNA in the nucleus[78] Blocks polyadenylation via CPSF[79] Blocks transcription of antiviral genes[80] Blocks mRNA export[81] Suppresses the host immune response via various mechanisms[82]
		NEP*	Spliced (5'G56 to 5'G526)[83]	Present in virions in small amounts[84] Nuclear export of vRNPs via Rab/hRIP1, yRip2, yNup100, and yNup116 or CRM1[85]–[87] Binds to M1 associated vRNPs[67] Switch from transcription to replication[88], [89]
		NS3	Spliced (5'A399 to 5'G526)[20]	Poorly conserved, but expressed in IAV strains associated with a recent change in species[20] Similar functions to NS1[20]

The eight gene segments of the Influenza A virus, the length of each gene in nucleotides, the name of the protein transcribed, the method of expression, and some highlighted functions. An asterisk (*) next to the protein indicates it has been deemed essential to an IAV infection.

1.2 Influenza A Virus Life Cycle

The life cycle of IAV can be divided into three distinct phases: virion uptake and content release, transcription and replication, and export and release of progeny virions. A graphical representation of this process can be found in Figure 1.1.

1.2.1 Virion Uptake and Entry

An influenza A virus infection may occur after the host organism takes up a small number of virions, with infection occurring at a low multiplicity of infection. The neuraminidase proteins (NA) on the virion cleave through the mucus lining the airway and allow for the virion to reach the respiratory cells[90]. Once the respiratory cells are reached, the hemagglutinin (HA) protein binds to sialic acid associated receptors. Uptake of the virion requires multiple binding events between HA and sialic acid associated receptors. Once sufficient binding is reached, the virion is taken up via receptor mediated endocytosis[42], [91]. The endocytosed virion is trafficked to the endosome for degradation[42]. Degradation is mediated through acidification of the endosome core. Virions posses a pH sensitive proton channel that allows for the influx of protons into the virion core once the pH drops below approximately 6.0[70]–[72]. Once this pH is reached in the endosome, the HA fusion peptide is exposed, allowing for the fusion of the virion and endosome membranes[37], [38], [42]. Concomitantly, the acidification of the virion core releases the vRNPs from the bound M1[64]. The vRNPs are released into the cytoplasm and then trafficked to the nucleus wherein transcription and, later, replication occur.

1.2.2 Transcription and Replication Phase

Transcription of mRNA and production of viral proteins is the next key phase of the viral life cycle. vRNPs bound to different viral gene segments are trafficked to different portions of the nucleus[92] wherein transcription begins in a cap dependant manner[14] via association with

host DNA-dependent RNA polymerase II (Pol II)[2]. Genes are initially transcribed in equimolar amounts[93]; however, transcription becomes differentially regulated shortly thereafter[93]–[97]. In the early phase, transcription of PB2, PB1, PA, NS and NP genes is predominant. This shifts towards HA, NA, and M genes later in infection. Once sufficient nucleoprotein and RdRp are synthesized replication begins (~4 hours post infection)[54], [55], [97] and occurs in parallel with transcription.

1.2.3 Export of Viral Ribonucleoproteins and Release of Progeny Virions

After the replication and transcription phase ends, and the vRNPs are exported from the nucleus. M1 binds to the vRNPs, stopping transcription[98], and then NEP associates with M1 to export the vRNPs from the nucleus[66], [67], [85]–[87]. The vRNPs are likely selectively packaged[63]–[67], [99]–[101], though the exact determinants are unknown. The vRNPs, M1, HA, NA, and M2 proteins associate to form virions which bud from the cytoplasm, though the cleavage mechanism remains to be elucidated[40]. NA cleaves nearby sialic acid residues to allow for efficient release of the virions[58], [59]. These virions may then infect other cells.

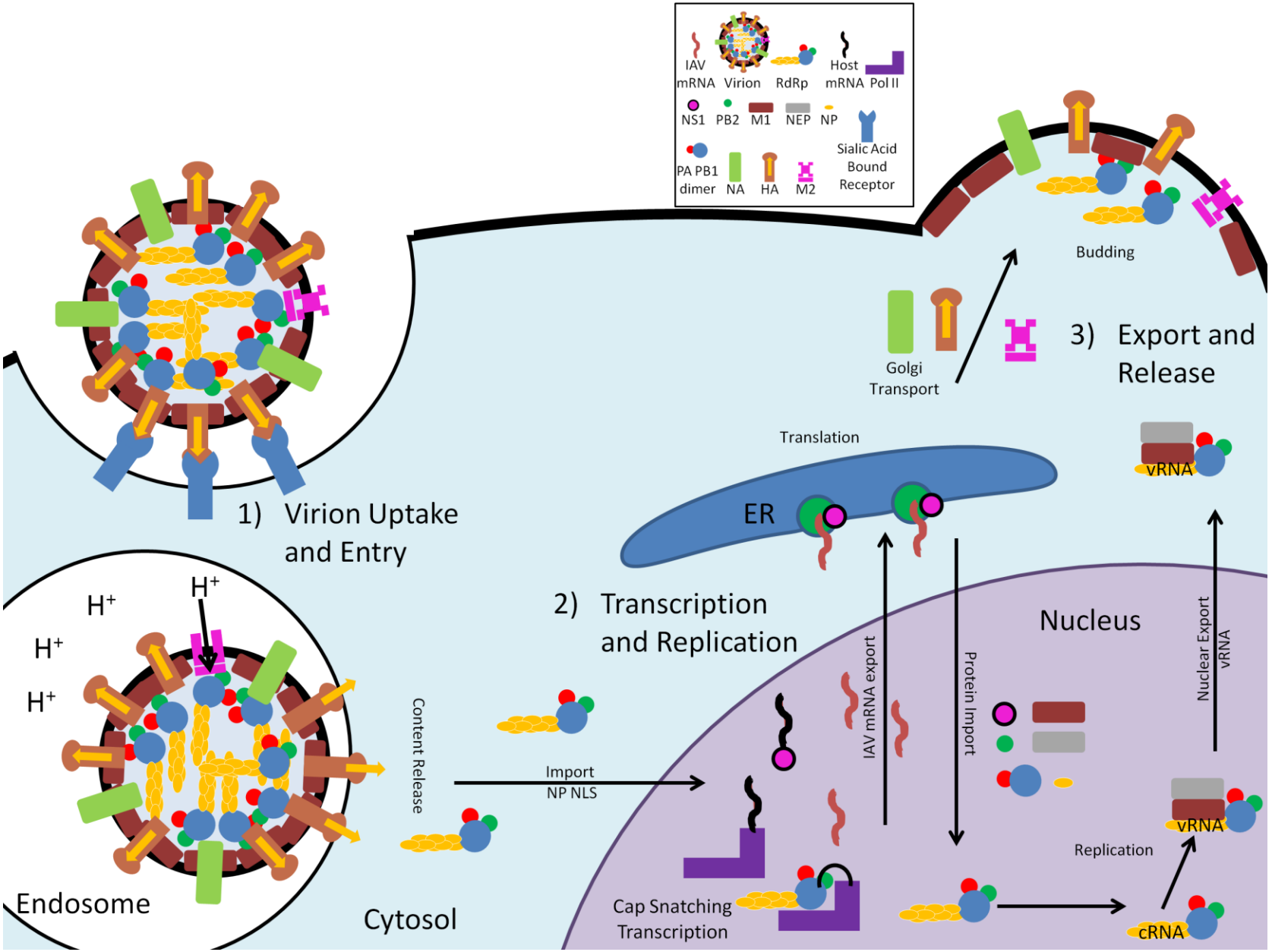
IAV mRNA
Virion
RdRp
Host mRNA Pol II
NS1
PB2
M1
NEP
NP
PA PB1 dimer
NA
HA
M2
Sialic Acid Bound Receptor
Budding
Golgi Transport
3) Export and Release
Translation
ER
vRNA
1) Virion Uptake and Entry
2) Transcription and Replication
Nucleus
H+
H+
H+
H+
Content Release
Import
NP NLS
IAV mRNA export
Protein Import
Nuclear Export vRNA
vRNA
Replication
Endosome
Cytosol
Cap Snatching Transcription
cRNA

Figure 1.1 Simplified Influenza A Virus Life Cycle. Virions are taken up via multiple binding events between HA and sialic acid presenting cell receptors. Virions are trafficked to the endosome. Acidification of the endosome leads to activation of the M2 ion channel, which in turn acidifies the virion core causing M1 and vRNPs (RdRp) to dissociate. Acidification also causes a conformational change in HA, causing the fusion peptide to associate with the endosome membrane, allowing for release of vRNPs into the cytosol. These vRNPs are then trafficked to the nucleus via nuclear localization signals within the bound proteins. Once in the nucleus the transcription and Replication phase begins. Transcription of mRNA occurs in a cap-dependent manner, via cap-snatching. Cap-snatching cleaves a 5' capped oligonucleotide sequence actively being transcribed by Pol II, and uses this oligonucleotide to prime transcription of IAV mRNA. IAV mRNA are selectively exported and translated due to the actions of NS1. Once sufficient RdRp and NP have been produced, replication begins. Replication is a two step process; first a cRNA intermediate is transcribed, then a vRNA copy is transcribed from the cRNA. NS1, NEP, and M1 are also imported to the nucleus. HA, M2, and NA are trafficked to the plasma membrane via the golgi. The export and release phase occurs next. vRNP are selectively trafficked to the plasma membrane via M1 and NEP. Together with M2, NA, and HA, these proteins are incorporated into budding virions. NA cleaves the sialic acid residues from receptors preventing reuptake of virions. Virions are then released from the plasma membrane via an unknown mechanism.

1.3 RNA-dependent RNA Polymerase

The RdRp is a heterotrimer comprised of the full-length proteins encoded by segments one to three of the IAV genome (PB2, PB1, PA)[23]. RdRp primarily exists as a viral ribonucleoprotein (vRNP), bound to the 3' and 5' conserved sequences of a single strand of the segmented IAV genome[102], with the remained of the vRNA being bound to NP[49], [51]. RdRp is trafficked to the nucleus[52], wherein it transcribes mRNA, replicates the IAV genome, and plays a role in host shut-off[103], [104].

While each subunit of the heterotrimer perform multiple functions each has a primary function in relation to transcription. PB1 is the central subunit of the heterotrimer to which the other subunits connect[23]. PB1 also contains the functional motifs associated with transcription of RNA[105], [106]. Both PB2 and PA subunits are important in transcription of mRNA via the cap-snatching process, wherein RdRp cleaves a 5' capped host-derived pre-mRNA, which it will use as a primer for transcription of mRNA. PB2 contains the cap binding region, and binds to 5' capped pre-mRNA[107]. PA contains the endonuclease region responsible for cleavage of PB2 bound host pre-mRNA[108]. Due to extensive interactions, the primary function of each individual subunit is related to the proper binding and association of the other subunits as well as the vRNA/cRNA [23], [109]–[111].

The overall structure and function of the trimeric RdRp of IAV is similar to the RdRp of other negative sense single-stranded RNA viruses in the order Bunyavirales and the families *Arenaviridae* and *Orthomyxoviridae*. IAV and other members of *Orthomyxoviridae* differ from *Arenaviridae* and Bunyavirales virus in that *Orthomyxoviridae* members engage in transcription and replication in the nucleus rather than the cytoplasm. There are also some key structural differences; in particular *Orthomyxoviridae* RdRp is a heterotrimer rather than a monomer as

seen in *Arenaviridae* and Bunyavirales RdRp. These viruses all transcribe mRNA in a cap dependant manner via cap-snatching, and have similar methods of replication.

1.3.1 Nuclear Localization and Assembly of RdRp

RdRp is imported to the nucleus in two distinct manners. Following release of the virion contents into the infected cell, mature vRNP are trafficked to the nucleus primarily by the nuclear localization signals (NLS) contained within the associated NP[53], though the effects of the RdRp NLS have yet to be determined[52]. Host synthesized PB2 and a PA:PB1 complex are imported into the nucleus via NLS contained within each protein[52]. PB2 is efficiently imported into the nucleus via a bipartite NLS that can associate with importin α1, α3, α5, and α7 isoforms[112]–[114]. The association of PB2 with importin requires a conformation change[114], which may be due to association with the chaperone Hsp90, which is trafficked to the nucleus with PB2[115]. PA shows partial localization in both the cytoplasm and the nucleus when expressed alone[116]. Nuclear accumulation of N-terminally tagged PA requires the region containing the UCC UUU CGU C motif responsible for the +1 open read frame shift that allows for translation of PA-X[33], [117]; it is possible that the observed nuclear accumulation of PA is due to PA-X rather than PA. Co-expression of PA and PB1 is the minimum viral protein requirement for efficient nuclear import of both PA and PB1[116]. PA and PB1 form a dimer in the cytoplasm prior to import to the nucleus[52], [116], [118], despite both proteins containing a NLS[52], [116], [117], [119]. The PA:PB1 dimer complexes with the importin-β RanBP5 for nuclear import[120]. This binding of PA:PB1 with RanBP5 also likely performs an essential chaperone function which along with Hsp90 PB2 interactions allow for proper assembly of RdRp with nuclear compartmentalization[115], [121], [122]. Association of polymerase subunits proceeds at a slower rate than is estimated by the Stokes-Einstein hydrodynamic theory,

indicating that other proteins may be involved in forming the RdRp complex[118]. It is likely that host chaperone proteins are important in both the compartmentalization and assembly of RdRp. This compartmentalization and assembly is essential for nuclear retention and accumulation of PB2 and the PA:PB1 dimer[112], [118].

1.3.2 Structure of Trimeric RdRp

While the structure of RdRp a human infecting strain of IAV remains to be determined, most of the structure of RdRp from a bat strain of IAV has been solved. Inter-subunit interactions are extensive, with the three subunits being tightly interwoven; the buried area between PB1 and PA is estimated to be 17,330 angstroms^2, between PB2 and PB1 14,100 angstroms^2, and between PB2 and PA 2,880 angstroms^2 [23]. RdRp forms a "U" shape with PB1 forming the central portion, and PA and PB2 forming the upper portions; the cap-binding region of PB2 faces the endonuclease containing region of PA[23] (see Figure 1.2 for a cartoon representation). This conformation is not static, however, and RdRp may change conformation.

vRNA becomes associated with trimeric RdRp in the nucleus where it binds to either cRNA or vRNA that has recently been transcribed from vRNA or cRNA respectively. Association of RdRp with vRNA or cRNA is the requirement for replication[54]; if this association happens during or after replication remains to be elucidated. Binding of RdRp to both the 3' and 5' ends of vRNA allows RdRp to act as transcriptase, and to effectively cap-snatch primers for mRNA transcription or to initiate replication; the conformation when bound to cRNA seems to only be conducive to replication[111]. 5' cRNA bound polymerase has been shown to have a different structure than vRNA bound polymerase[111], though the structure of RdRp bound to both the 3' and 5' cRNA promoter remains to be elucidated. It is likely that when acting as a replicase, rather than a transcriptase, RdRp changes conformation; and given that

mRNA is not synthesized from cRNA, a change in conformation of RdRp is a key component of the virus life cycle. It has been proposed that the difference in M1 binding to vRNP versus cRNP is due to a conformational change which allows for the selective export of vRNP over cRNP[63], and incorporation into the virion. cRNA bound RdRp has not been observed to engage in mRNA transcription *in vivo*. It is likely that other important interactions with RdRp are mediated by the conformational differences between cRNA and vRNA bound structures.

RdRp binds to vRNA at the 3' and 5' conserved sequences, with the binding site relocating the conserved sequences if they are not located at the 3' and 5' end[102]. The 5' vRNA conserved sequence is 5'-AGUAGAAACAAGG-3'[123] which is followed by a short variable region of three nucleotides (which may end in an uridine residue) and a polyuridine track five or six uridines in length[15]. The 3' vRNA conserved sequence is 3'-UCG[U/C]UUUCGUCC-5'[123] with the cytidine residue at position four being present on the PB2, PB1, and (usually) the PA segments. Nucleotides beyond the first 13 of the 5' and the first 12 of the 3' are not conserved among genes, but do contribute to the vRNA promoter region. The structure of the bat IAV polymerase bound to the vRNA promoter region shows interaction between the 3' and 5' ends as well as interactions between RdRp and the vRNA[23]. Five base pairs are formed between nucleotides 11 to 15 of the 5' end pair with nucleotides ten to 14 of the 3' end. This region incorporates the last three residues of each conserved sequence as well as the first two residues of the 3' and 5' segment specific untranslated regions[23]. Using the Turner Energy Rules[124], which estimate the free energy of an RNA duplex from both hydrogen bonds and stacking interactions, the free energy of the five base pairs ranges from -7.65 kcal/mol to -9.61 kcal/mol (excluding the energy penalty of forming an RNA duplex). The 3' end and 5' end are then separated by PA amino acids 503-Arg-Leu-His and 466–475, which form a wedge[23]

directing both ends to separate binding pockets. Overall this causes the conserved sequences to adopt a corkscrew conformation[23], [125] (Figure 1.3). The remainder of the vRNA is directed out of RdRp and is bound to NP[45]–[50].

The 5' end forms a tetraloop with 5'G2 pairing with 5'C9 and 5'U3 pairing with 5'A8[23], [125]. There are extensive contacts between the PA and PB1 subunits and the 5' end via both polar and stacking interactions up to the 11^{th} base pair[23]; this leads to a K_d of approximately 2.2nM for the 5' end[125]. This tight hold on the 5' end is extremely important in non-enzymatic polyadenylation of vRNA, as it prevents tracking of the 5' end into the polymerase active site, allowing for iterative copying of the polyadenine track[15], [126]–[130]. Less extensive contacts are seen for the 3' end of the vRNA, with only nucleotides six to nine of 3' end having interactions with RdRp[23]. Residues from all three polymerase subunits interact with the phosphate backbone of the 3' end between 6-UUCG-9, and position the remainder of the 3' end towards the polymerase template entry tunnel[23]. These interactions also force a sharp turn of the vRNA between 3'C8 and 3'G9[23]. The less extensive enzyme-vRNA interactions likely contribute to the much higher K_d for the 3' end (>1 μM)[125]. While base pairs between 3'C2 and 3'G9 as well as 3'G3 and 3'C8 have been suggested via FRET[125], the structure of the 3' end of the vRNA 3' of 3'U6 could not be resolved in the crystal structure[23], indicating that the RNA secondary structure is not static. These results have been confirmed by FRET analysis, where the 3' vRNA end has been shown to exist in two conformations in the absence of polymerase activity[131].

The 3' and 5' ends has also been shown to be related to selective packaging of genome segments[99], [100]. The hemagglutinin gene from the A/WSN/33 (H1N1) strains with the 3' and 5' non-coding regions from a different IAV type were shown to produce lower viral titers

with minimal changes to mRNA, vRNA, and cRNA production[99]. The changes in replication were shown to be due to a reduction of packaging of the HA gene, with no decreases to other genes[99]. The loss of viral titer was shown to be related to the length of the non-coding regions relative to the length of the wild type WSN coding region, with the 3' non-coding region being more important for successful packaging than the 5' end[99]. Similar results were obtained with switching the non-coding regions of neuraminidase with the non-coding regions of other genes; lower viral titers were observed, with only the NA gene's packaging being affected[100]. Packaging could be rescued in mutant NA genes through lengthening mutations and point mutations for most of the other 5' and 3' non-coding regions[100]. The exact mechanisms of how these regions impact packaging remains to be elucidated, however other regions of the vRNA are implicated in this process[101]. Selective packaging may be mediated by interactions of the 3' and 5' noncoding region with other RNA elements in the gene, as well as the conformation of RdRp.

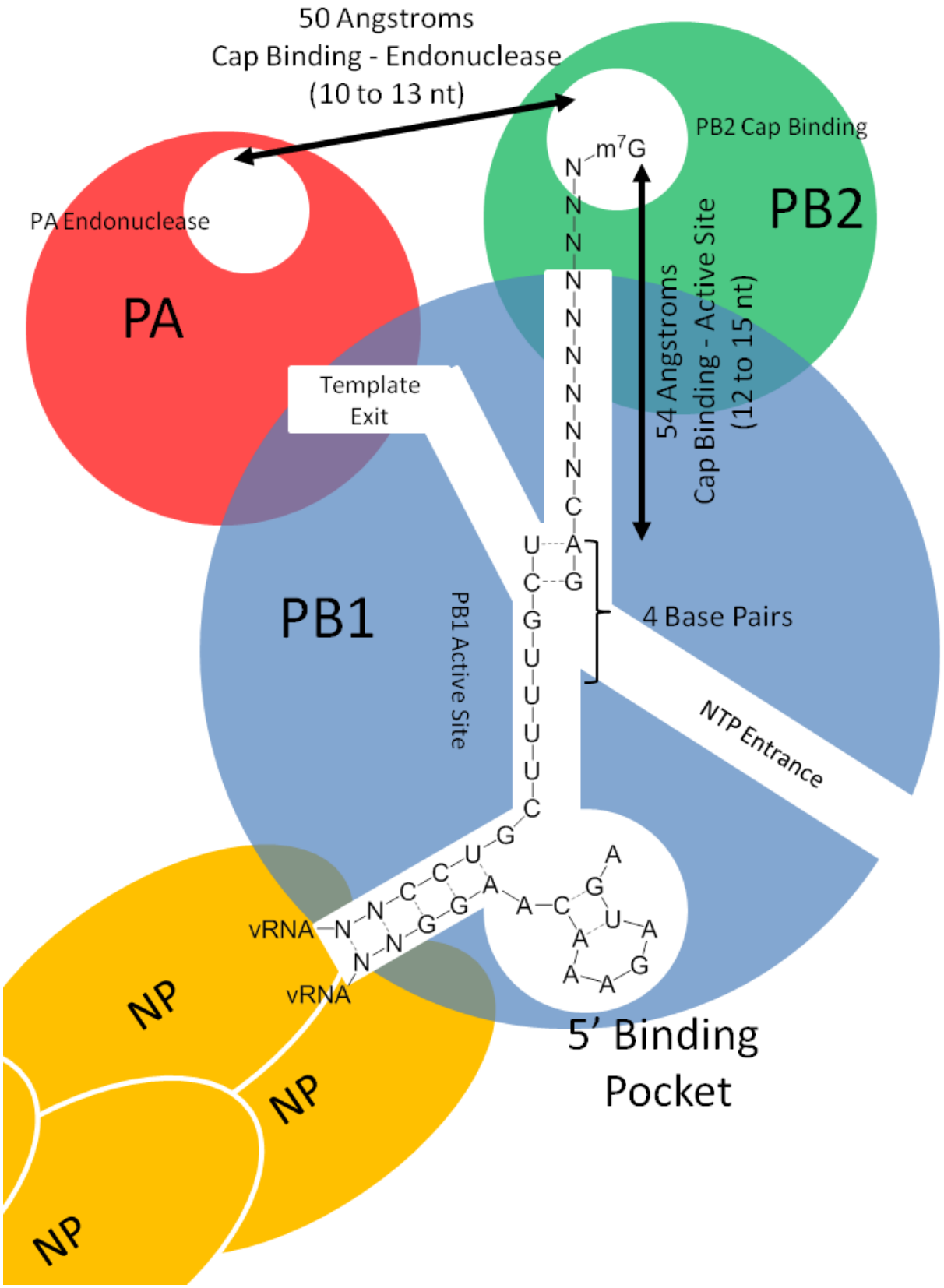
50 Angstroms
Cap Binding - Endonuclease
(10 to 13 nt)
PA Endonuclease
PA
PB2 Cap Binding
PB2
m7G
54 Angstroms
Cap Binding - Active Site
(12 to 15 nt)
Template
Exit
PB1
PB1 Active Site
4 Base Pairs
NTP Entrance
vRNA
vRNA
NP
NP
NP
5' Binding
Pocket

Figure 1.2. Cartoon Representation of RdRp bound the vRNA Promoter and Priming Transcription. A cartoon representation of RdRp, the PB1 subunit (blue) has two entrance tunnels, one occupied by the vRNA coming from the nucleoprotein (orange), the other allowing NTPs to enter the active site. The two entrance tunnels join at the active site, and split into the template exit tunnel (left; where the vRNA exits near PA opposite to the endonuclease), and the product exit tunnel. Cap snatched primers are threaded through the product exit tunnel to prime transcription. The distance from the PB2 cap binding region to the PB1 active site has been estimated to be 54 Angstroms; the distance from the PB2 cap binding region to the endonuclease active site has been estimated to be 50 angstroms[130]. The 5' end of the vRNA is tightly held in a binding pocket, and there are five base pairs between the 3' end and the 5' end of the vRNA. The nucleotides 3'-UCGUUUU are not bound to the polymerase. The 3' nucleotide sequence is that of the 3'U4 template (typically five out of eight IAV gene segments). An A indicates and adenine residue, C indicates a cytidine residue, G indicates a guanine residue, U indicates a uridine residue, and N indicates a nucleotide reside. Nucleotides labeled as N are gene segment specific. Black lines indicate phospholinkages, and hashed lines indicate base pairs.

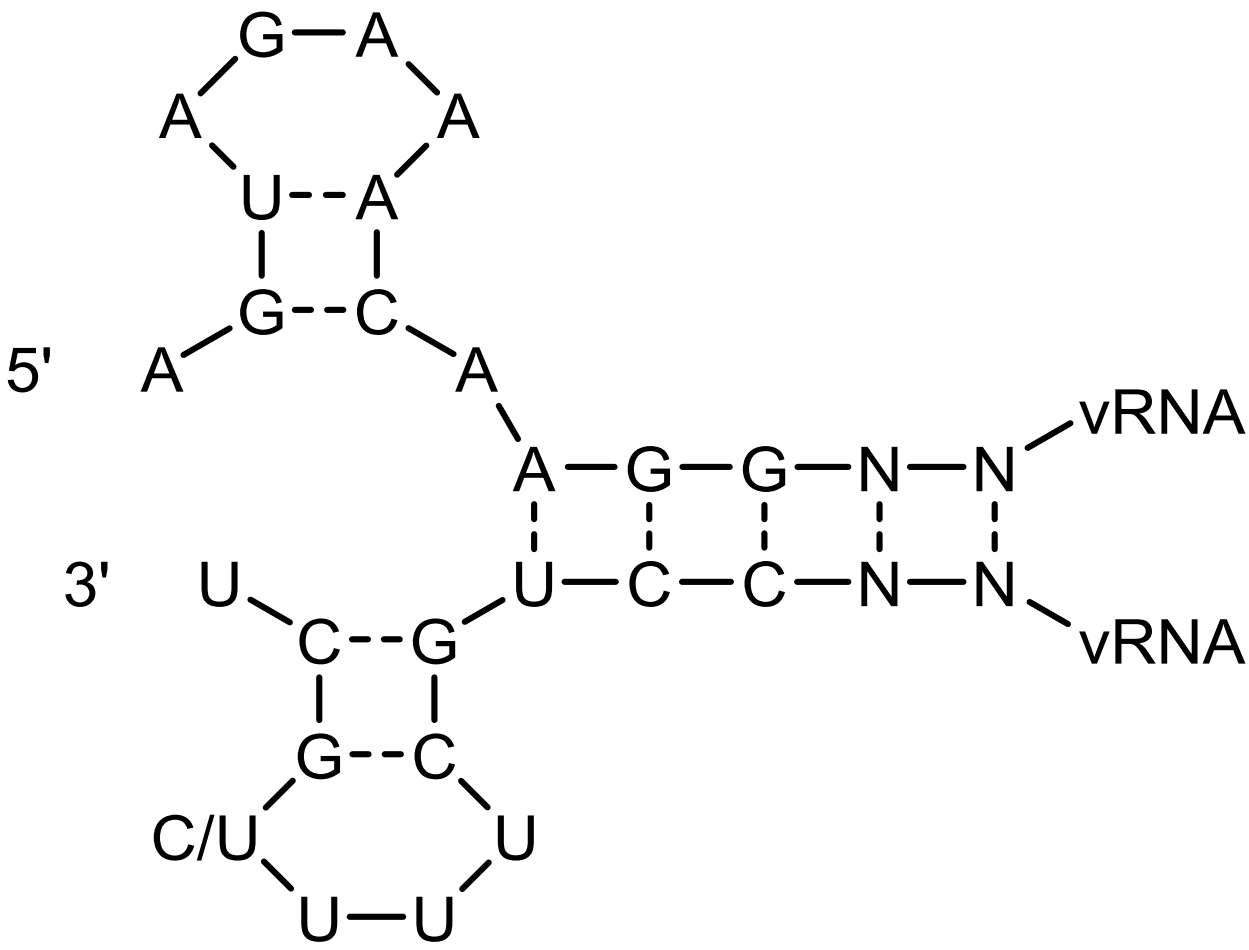
G
A
A
A
U
A
G
C
5'
A
A
vRNA
A
G
G
N
N
3'
U
U
C
C
N
N
vRNA
C
G
G
C
C/U
U
U
U

Figure 1.3. IAV vRNA Promoter Structure in the Corkscrew Conformation. The 3’ and 5’ conserved sequences of IAV and the typical secondary structure they form within RdRp. An A indicates and adenine residue, C indicates a cytidine residue, G indicates a guanine residue, U indicates a uridine residue, and N indicates a nucleotide reside. Nucleotides labeled as N are gene segment specific. Black lines indicate phospholinkages, and hashed lines indicate base pairs.

1.3.3 Polymerase Structure

The central core of RdRp forms the active site, and shows a structural similarity to the hepatitis C virus polymerase, and bacteriophage Φ6[23], and a very high degree of similarity to the Bunyavirales polymerase[132]–[134]. The active site is formed by the finger, thumb, palm, and priming loop regions of the PB1 subunit of RdRp. Interactions with PA, PB2, and the bound vRNA stabilize this region in a functionally active conformation[23], and may play a role in polymerase function[135].

The four functional motifs of the PB1 palm domain are 303-TGDN-306, 403-LSPGMMMGMF-412, 438-WDGLQSSDDFALI-450, and 474-GINSKKKSYI-484[106], [136]. Asp^{305} of motif I and Asp^{445} and Asp^{446} of motif III form the divalent cation binding pocket necessary for polymerase function[106]. These cation binding sites allow for the polymerase reaction and the addition of the incoming NTP to the growing chain by promoting the nucleophilic attack of the 3' OH of the RNA chain on the α-phosphate of the NTP. Motif IV is involved in binding to incoming NTPs, while Motif II acts to stabilize the base pairs being formed by the polymerase[106]. Each of these four regions shows a high degree of conservation among IAV strains, particularly in strains which affect humans, as mutations generally result in a loss of transcriptional activity[136]. Mutations to the highly conserved Asp^{305}, Gly^{406}, Asp^{445}, Asp^{446}, and Lys^{480} residues generally abolish polymerase activity[106], [136]. The structure of RdRp brings these motifs close together, and enables polymerase function[23].

The PB1 priming loop (thumb domain) is directly across a small solvent channel from the palm domain[23]. This region plays a critical role in *de novo* (primer independent) initiation of replication, as well as a stabilizing role in transcription. Pro^{451} acts as a stacking platform for the initial NTP used in replication[137]; deletion of the region containing this residue greatly reduces replication, though it does not affect transcription[138], [139]. Deletion of the entire

priming loop, or either end of the priming loop, greatly decreases transcription of mRNA[138], indicating that the priming loop is still important for mRNA transcription. Furthermore, deletion mutants of residues 636-642 or 642-656 result in a large increase in polymerase stuttering events when priming transcription of mRNA[138]. The priming loop may be important for the association of the polymerase and the template during transcription and replication.

There are four channels leading to the core of RdRp that are important for transcription activity: two entrance channels which join before the polymerase active site, and two exit tunnels which branch off from the active site[140]. The template entrance tunnel is formed by the 3' vRNA promoter region, and allows for entry of the 3' end of the template to easily access the active site[140]. The 3' end of the vRNA is capable of reaching the polymerase active site through this channel without breaking of the 3' and 5' base pair region[110]. The NTP entrance channel is lined with conserved positive residues and allows for NTPs to enter the active site in parallel with the template[23], [140]. The NTPs exit proximal to the PB1 priming loop[23], where they can associate with the growing RNA chain or be used to prime replication. Opposite to these two channels are the template and product exit channels. The product exit channel goes from the PB1 polymerase site to the PB2 cap binding region[140]. It is through this exit channel that cap snatched primers are threaded to prime transcription of mRNA[109], [110], [130]. The template exit tunnel splits off from the polymerase active site towards the same side as the template enters[140]. It is hypothesized that the PB2 helical lid region splits the template product duplex[140]. Together these channels direct the template through the active site opposite the bound metal ions, and direct the NTPs proximal to the bound metal ions; these entering NTPs can be added to the growing chain, and exit the polymerase active site towards the PB2 cap binding region.

1.3.4 RdRp as a Replicase

RdRp is the enzyme that handles replication of the IAV genome. Both replication and transcription occur early in infection; however, replication products (cRNA) do not accumulate in the absence of host synthesized RdRp and NP[54], [55]. RdRp binds to the 3' and 5' ends of cRNA and vRNA preventing their degradation[54]. NP appears to play an anti-termination function, binding to cRNA/vRNA as it is replicated[55]. In the context of a cellular infection (*in vivo*), this leads to replication occurring later after synthesis of RdRp and NP in the host cell. The polymerase subunits and NP are the minimum protein requirements for replication.

Nuclear export protein (NEP) has been shown to promote replication[88], [89], [141]. NEP has been shown to bind to PB2 and PB1[89]; it is possible that this binding promotes the "replicase" conformation[111] of RdRp. NEP also helps to promote the production of short viral RNAs by RdRp which are similar to the last 22 to 27 nucleotides of the 5' end of the vRNA, and may act as a substitute 5' promoter region allowing the 5' end of the vRNA to enter the polymerase active site[141]. Other host and viral factors are likely involved in this process, and the mechanism of replication is not yet known.

Replication of full length vRNA or cRNA is either primed with an AG dinucleotide or occurs *de novo*[142]. Using the IBV polymerase and a fluorescence polarization assay, it has been estimated that *de novo* replication is ~300 fold slower than initiation with a dinucleotide, with cRNA being replicated at a slower rate than vRNA[110]. The Pro^{651} residue of the PB1 β-hairpin priming loop, which is hypothesized to act as an NTP stacking platform, is a requirement for *de novo* initiation on the vRNA template, but not the cRNA template[143]. *De novo* initiation for cRNA occurs by synthesis of a dinucleotide at 3'4-UC-5, which is then realigned to the 3' end to initiate replication[142]. This realignment process in replication is also seen in members of *Arenaviridae*[144], and likely occurs in members of Bunyavirales. Following priming with a

dinucleotide, the next rate-limiting step is the melting of the 3' 5' base-pairing region. Decreasing the bond strength (higher free energy) of this region increased the rate of replication[110].

Replication occurs for the entire length of the vRNA/cRNA, including the 5' end, which is not transcribed in viral mRNA. To allow for this, RdRp must release the 5' end from the binding pocket. This does not occur during mRNA synthesis, wherein the polymerase stutters while transcribing the polyuridine track with the 5' end remaining held in the binding pocket, and thereby generates a polyadenylated mRNA[126]. Though the complete mechanism of how this occurs is unknown, short viral RNAs replacing the 5' end[141] or other proteins may be involved.

A potential mechanism of replication has been described where two RdRp act in trans to efficiently replicate vRNA/cRNA. Isolated vRNP may initiate replication in the absence of other proteins, however, isolated cRNPs do not engage in replication[145]. Treatment of isolated cRNPs with recombinant RdRp rescues replication, and allows for the accumulation of vRNA[145]. Furthermore, isolated replication deficient RdRp could be rescued for replication through treatment with replication competent RdRp[146]. Replication likely occurs via a trans-acting mechanism. It is unknown if the RNA bound RdRp or the free RdRp engages in the replication, and if the template RNA remains associated with the same RdRp throughout this process.

1.3.5 RdRp as a Polyadenylase

IAV mRNA transcribed by RdRp are 3' polyadenylated. This polyadenylation process is not brought about by a specific IAV encoded polyadenylase or by a co-opted host factor[15]. 3' polyadenylation is accomplished through iterative copying of a polyuridine track near the 5' end

of the vRNA template by polymerase stuttering[126]. RdRp increases the probability of polymerase stuttering to ~100% near the 5' end through a combination of a tight hold on the 5' end by RdRp, and a lower free energy duplex between the polyuridine track and the transcribed adenine residues. All eight IAV genes have a region of five to seven consecutive uridine residues 15 to 17 nucleotides from the 5' end[15], wherein polymerase stuttering is forced to bring about polyadenylation.

Evidence for the uridine track being used for polyadenylation comes from three sources: first, mRNA is a perfect copy of the vRNA up until the polymerase reaches the beginning of the uridine track[15]. Secondly, no nucleotides 5' of the uridine track are transcribed into mRNA. Mutations of individual uridine residues within the uridine track prevented the polyadenylation of IAV mRNA[126], save for mutation of the final uridine residue which resulted in reduced, but present, polyadenylation from polymerase stuttering[126]. The third line of evidence is that mutation of this polyuridine region to a polyadenine region resulted in a polyuridinylated mRNA with the polyuridine track beginning at the same location of the mRNA[126].

The presence of a polyuridine track alone is not enough to explain the 100% polymerase stuttering rate observed. The free energy relative to the hold on the 5' end and the length and positioning of the polyuridine track are important factors in increasing the probability of polymerase stuttering. The ten 5' most nucleotides of the vRNA are held within a binding pocket in RdRp[23], [130]. The affinity of the binding pocket for the 5' has been shown to have a K_d of 2.2 nM[125] which would correspond to a free energy of approximately -12 kcal/mol. Decreasing either the free energy of the transcription duplex or decreasing the affinity of the 5' end for its binding pocket decrease the probability of polymerase stuttering. Mutations to 5' residues two, three, seven, eight, and nine which affect RdRp binding to the vRNA[23], [147]

severely reduced polyadenylation (<20% wild type)[127]. Additionally, the release of the 5' end from the binding pocket during replication results in full length RNA copy, typically without polymerase stuttering events[128]. The hold on the 5' end must be sufficient to offset the strength of the vRNA mRNA duplex and the stabilizing effects of RdRp during transcription. Decreasing the free energy of the transcription duplex by mutation of the polyuridine track (free energy of -1.72 kcal/mol at four nucleotides) to a polycytidine or polyguanine track (-9.57 kcal/mol), but not a polyadenine track (-1.72 kcal/mol), results in a loss of polymerase stuttering in this region (polyadenylated and polyuridinylated RNA, but no polycytidinylated or polyguanidinylated RNA)[126]. The free energy of the 5' end binding pocket (approximately -12 kcal/mol) must remain higher than the sum of the stabilizing effects of RdRp and the transcription duplex, or the 5' end will be released and the polymerase will read through the remainder of the template.

Based on the structure of RdRp, this polyuridine track must start 17 nucleotides 3' of the 5' end. This allows for the 5' conserved sequence to remain within the binding pocket while the polyuridine track enters the polymerase[130]. Changes in the position or length of the polyuridine track reduces the levels of polyadenylated mRNA. The insertion or deletion of two or more nucleotides between the 5' end and the polyuridine track resulted in decreased levels of a chloroamphenicol acetyltransferase (CAT) reporter (with IAV 3' and 5' UTRs of the NA gene) to <2% of expression of CAT with the wild type UTR, with no change in cRNA production[129]. Moving the uridine track closer to the 5' end would position the optimal region for stuttering in an area with stronger bonds than the polyuridine track, allowing for the 5' end to be dislocated from the binding pocket. Similarly, moving the polyuridine track away from the 5' end would allow it to be efficiently read through prior to the steric hindrance promoting

polymerase stuttering. It is likely these shifts would decrease the probability of polymerase stuttering from 100%.

Decreasing the length of the polyuridine track to four uridine residues reduced CAT expression to 4% of the wild type polyuridine track[129]. Current estimates of the IBV RdRp active site suggest that it can accommodate at least three base pairs[110], though accommodation of four base pairs may be possible for the IAV RdRp. Polymerase stuttering during replication near the 5' end has been observed to iteratively add one to four nucleotides[128], and prime-and-realign (polymerase stuttering following initiation of transcription) has been shown to be a similar one to six nucleotides[1], [4], [5], [7]. It is possible that the four A-U base pairs from this four uridine long polyuridine track may be stabilized within the RdRp active site, and transcription through the remainder of the vRNA may occur. Simply, the stabilizing effects of RdRp may offset the energy penalty from the hold of RdRp on the 5' end decreasing the probability of polymerase stuttering through a four nucleotide long polyuridine track. In summary, the length and positioning of this polyuridine track and the hold on the 5' end adjust the probability of polymerase stuttering to 100% and allow for efficient polyadenylation.

1.3.6 RdRp as a Transcriptase: Cap-Snatching and Prime and Realign

IAV mRNA transcribed by RdRp are 5' capped without an IAV encoded protein for this process or co-opting of the host 5' capping machinery[14]. 5' capped mRNA are generated through a process called "cap-snatching"; where a 5' capped oligonucleotide is cleaved from a host derived 5' capped transcript and used to prime transcription of viral mRNA[14]. Other viruses in the *Orthomyxoviridae* family[1], [4], [5], [7]–[9], [14], [110], [138], [148], [149] (though Thogotovirus has a somewhat distinct process[150]–[152]), Bunyavirales order[153]–[174], and *Arenaviridae* family[175]–[180] use a similar cap-snatching process. This results in chimeric

mRNA with a host derived leader sequence followed by the viral mRNA. For IAV, cap-snatching is carried out in four steps: association with host Pol II, PB2 m^7G cap binding, PA mediated endonuclease cleavage, and priming transcription and, potentially, prime-and-realign.

Nucleotide sequence repeats complimentary to the 3' end of the vRNA conserved sequence not present in the cap snatched primer have been observed between the viral conserved sequences and the 3' most nucleotide of the host derived sequence[1], [4], [5], [7] (Table 1.2). These insertions arise through an iteratively aborted transcription attempt of the second to seventh nucleotides of the viral conserved sequences[4], [5], [138]. It is hypothesized that the polymerase transcribes one to six nucleotides, falls off the template, realigns to the start of the vRNA, and initiates another round of transcription with the elongated primer. This process is called prime-and-realign (PAR). PAR is not limited to a singular round, and some host-derived sequences have multiple nucleotide sequence insertions from multiple rounds of PAR. mRNA transcribed from the genomes of *Orthomyxoviridae*, Bunyavirales, and *Arenaviridae* show evidence of PAR, and it appears that PAR is related to cap-snatching. While the exact determinants of this process remain to be elucidated, it appears as though there is no single cause of PAR.

Table 1.2. Nucleotide Sequences which may be Added by the Prime and Realign Process

	1 nt	2 nt	3 nt	4 nt	5 nt	6 nt
Possible Additions	G	GC	GCA GCG	GCAA GCGA	GCAAA GCGAA	GCAAAA GCGAAA

Indicates all possible nucleotide sequence additions which may be added via prime-and-realign for every nucleotide of length from 3’C2 to 3’U7 of the IAV conserved sequences.

1.3.6.1 Association with Host DNA-Dependent RNA Polymerase II

While RdRp can initiate transcription of mRNA from a free capped primer[14], *in vivo* cap-snatching involves association with Pol II. The first evidence of Pol II being involved in this process was the observation that α-amanitin treated cells (a specific inhibitor of Pol II at low concentrations) also inhibited RdRp mediated transcription[181]. Expression of IAV proteins was below the level of detection in cells treated with α-amanitin six hours post infection, whereas untreated cells, or cells with an α-amanitin resistant Pol II mutant showed accumulation of IAV proteins[181]. RdRp was shown to co-immunoprecipitate with a slow migrating Pol II form via the C-terminal domain (CTD)[182]. The PA subunit was also shown to partially colocalize with Pol II in the nuclei of infected cells[182]. This interaction appears to be specific to Pol II; using a chromatin immunoprecipitation (ChIP) assay with lysate from IAV infected cells and a PA specific antibody, only Pol II promoter DNA was shown to be present (no DNA from Pol I or Pol III transcribed genes)[183]. Furthermore, this association with Pol II appears to be specific to the serine-5-phosphorylated state of the Pol II CTD[3].

The serine-5-phosphorylated state is indicative of Pol II pausing for 5' capping of nascent host pre-mRNAs. Treatment with DRB, which is known to inhibit transition of the CTD of Pol II from the serine-5 phosphorylated state to the serine-2-phosphorylated state, did not inhibit cap-snatching and mRNA transcription of RdRp[183]. This data shows that RdRp mediated cap-snatching only requires transcription initiation and 5' capping of nascent host pre-mRNA. Following this association RdRp pauses Pol II at the initiation site; analysis of Pol II distribution on the β –actin gene using ChIP with a Pol II-specific N20 antibody showed a two-fold decrease on Pol II in the coding region of the gene with no change in the promoter region in IAV infected cells[183]. This binding has been shown to be a direct and specific association with the serine-5-phosphorylated CTD[3]. Using a CTD mimic protein, RdRp only copurified with the serine-5-

phosphorylated, but not the serine-2-phosphorylated or scrambled control, protein[3]. RdRp associated with this serine-5-phosphorylated protein were transcriptionally active, indicating that this association is compatible with transcription of mRNA[3]. Taken together, these data indicate that RdRp binds with the serine-5-phosphorylated CTD of Pol II during pre-mRNA capping, and pauses Pol II transcription to allow for cap-snatching and IAV mRNA transcription.

Recently the crystal structure of bat RdRp associated with a truncated four-heptad repeat (YSPTSPS sequence) CTD of serine-5-phosphorylated Pol II has been solved[2]. The binding to the Pol II CTD is mediated by two binding sites with the C-terminal domain of PA, at a site distal to other key functional domains. The first binding site is in a grove formed by three PA α-helix (16, 20, 21) in which six amino acid residues of the first CTD repeat are inserted and adopt a β-like conformation[2]. The phosphate of the serine within this repeat interacts with a positively charged pocket formed by Lys^{630} and Arg^{633} of the bat polymerase[2]. The second, IAV specific, binding site is formed by the PA 550 loop and β18-β19 ribbon in which amino acid residues from the second to fourth CTD repeat are inserted and adopt a β-turn conformation[2]. The serine from the third repeat is bound to Lys^{289} and Arg^{449} of the bat polymerase via hydrogen bonds[2]. Binding of the bat polymerase to these phosphorylated serine residues had a K_d of 0.9μM, versus 6.1μM for only one repeat binding, and >10μM for unphosphorylated binding[2]. Mutation of both the Lys and Arg residues of either of the two serine binding sites, which allow for RdRp to associate with the Pol II CTD, resulted in a non-viable virus. Mutations to one of the two binding residues lead to a >10,000 fold reduction in viral titers[2], likely due to a loss of capped primer dependant transcription. These data indicate that efficient Pol II binding in the serine-5-phosphorylated state is of critical for the viral life cycle, and that the binding affinity must be high to allow for association in the transient serine-5-phosphorylated state.

1.3.6.2 Polymerase Basic Protein 2 m^7G Cap Binding

Following the association with Pol II, PB2 binds to the 5' capped end of the pre-mRNA, which has been transcribed to a length of 20-30[184] nucleotides by the bound Pol II. Cap-bound structures for RdRp of both IBV[110] and the bat IAV[109] have been partially solved. There are multiple contacts between the PB2 cap binding domain (residues 318-483)[107] and the m^7Gppp cap structure. Phe^{323}, Phe^{404}, and His^{357} form the 5' cap structure binding pocket, and engage in stacking interactions with the base[107], [109]. Glu^{361} and Lys^{376} form hydrogen bonds with the N1 and N2, and O6 of the guanine (respectively)[107]. Gln^{306} binds with the ribose sugar hydroxyls[109]. Asn^{429}, Arg^{264}, and Ser^{519} interact with the phosphates[109]. There are further interactions with the second and third nucleotide residues 3' of the 5' capped end[109]. Together, the binding of the 5' capped end has a free energy of approximately -40 kcal/mol[185], [186], with a K_d of 170 μM[107]. The affinity for PB2 to the 5' cap structure is much lower than that of eIF4E (K_d~130 nM)[187], possibly due to no direct interaction with the positive charge of the methylated cap structure[107]. This lower affinity may be of importance in release of the 5' cap during transcription, which is thought to happen due to lengthening of the primer during transcription rather than via an enzymatic process[105], [130].

1.3.6.3 Polymerase Acidic Mediated Endonuclease Cleavage

PB2 bound nascent host mRNA must be cleaved before transcription of mRNA can be primed. Cleavage of the bound mRNA is mediated by the endonuclease region within the N-terminal half of PA[108]. While endonuclease activity of the N-terminus is viable without association of PA (or PA-X which contains the same functional region)[33], [35], [108], in the context of RdRp this region requires key interactions with PB1[23], [188], PB2[23], [111], and the binding of the 5' and 3' vRNA (but not cRNA) promoter[111], [189]–[191].

Mutational analysis showed that mutation of any base pair in the 3' vRNA promoter except for 3' 4-7 abolished endonuclease activity; mutations to 3' 2-3 or 8-9 could be rescued by a complimentary mutation to 3' 9-8 or 3-2 respectively to restore the hairpin conformation[190] (Figure 1.3). Interestingly, the natural variation at 3'U4 to C4 present in the PB2, PB1, and PA genes results in a slight loss of endonuclease activity[190], [191]. This loss is possibly due to a decrease in the time spent in the hairpin conformation[191], as the residues 3' of 3'U6 seem to be quite mobile, and could not be resolved in the crystal structure[23]. Similar to the 3' end, only mutations within the hairpin region of the 5' vRNA promoter are tolerated, except for 5'G5[189]. Complimentary mutations restoring secondary structure of the hairpin (5' 2-3 and 8-9) or the base pairing region (5' 11-15 and 3' 10-14) with 3' vRNA promoter can partially rescue mutations at these points[189]. 5'G5 and 5'A10 are both involved in interactions with the PA-arch (residues 366 to 397), with the Pro^{392} in this region stacking on 5'G5[23]. The whole of the 5' end is more extensively buried in the PA subunit than in the PB1 subunit[23] further shown the importance of the interaction between the 5' end and PA. When both the 3' and 5' vRNA promoter are present, the PA subunit of RdRp is able to adopt its active formation[23], [111].

The PA endonuclease site is anchored to both PB1 and PB2 through the PB1-Cter–PB2-Nter interface[130]. Through these interactions, all three subunits participate in positioning the endonuclease so that it is facing the cap binding region of PB2[23]. The distance between the cap-binding region and endonuclease region is estimated to be a linear distance of approximately 50 Angstroms, which is consistent with cleavage between ten and 15 nucleotides[130]. There is some variability in this distance, as the PB2 subunit has a high degree of mobility with different crystal structures showing as much as 70^{o} of rotation in the PB2 cap binding region between

Ile^{319} and Arg^{495} residues[130]. It is likely that modulation of this distance is important in both the length selective and sequence specific cleavage observed for the PA endonuclease.

The endonuclease region of PA is a PD-(D/E)XK endonuclease, formed by Pro^{107}, Asp^{108}, Glu^{119}, Lys^{134} residues[108], [192], and is similar to the Bunyavirales and *Arenaviridae* L protein endonuclease[178], [193]. Cleavage is mediated by two active sites bound to either manganese or magnesium[194], the first is formed by Glu^{80}, Asp^{108}, and Glu^{119} residues, and the second is formed by His^{41}, Asp^{108}, Glu^{119}, and the carbonyl of Ile^{120} residues[108]. Mutation to any of these key binding residues, with the exception of Ile^{120} (cation interaction is mediated by the carbonyl), greatly reduces transcription of mRNA when endonuclease activity is required[192], [194]. Mutation of Lys^{134} also abolishes activity, demonstrating that it is indeed the catalytic lysine of the PA endonuclease[194]. The IAV endonuclease also contains a catalytic histidine residue. This histidine residue increases the endonuclease activity by about 180 fold the activity of a similar endonuclease lacking a catalytic histidine[193]. While both Mn^{2+} and Mg^{2+} have been observed to bind in this region, manganese ions bind with a higher affinity. The K_d for the two binding sites in the N-terminally truncated PA are 0.3 μM and 6.5 μM for manganese and 148 μM and 4000 μM for magnesium[194]. These differences in affinity lead to a cleavage rate that is approximately 20 fold slower for Mg^{2+} than Mn^{2+} (using 5mM $MgCl_2$ versus 2mM $MnCl_2$)[193]. Despite the affinity for manganese being much higher, magnesium is at a higher concentration *in vivo* and is therefore still biologically relevant in this process. PA cleaves bound mRNA via its endonuclease region with a catalytically active lysine and two bound Mn^{2+} or Mg^{2+}.

1.3.6.3.1 Length and Sequence Specific Cleavage

Similar to cap-snatching from other viruses, RdRp cleaves 5’ capped host transcripts with a given preference for both length and nucleotide sequence compatibility with the 3’ end of the

vRNA. The cleavage of mRNA was shown to be between generally ten to 13 nucleotides 3' of the 5' cap, typically 3' of a guanine residue. The length distribution of cap snatched primers increases above background at approximately ten nucleotides and peaks at 11 and 12 nucleotides depending on the gene transcribed[1], [6]–[8], [14]. This length range is consistent with the estimated distance between the PB2 cap-binding and PA endonuclease regions (~ 50 Angstroms, estimated to be at least ten nucleotides[130]), but is shorter than the estimated distance between the PB2 cap-binding region and the PB1 active site (54 Angstroms, estimated to be at least 12 nucleotides[130]). This indicates that RdRp may cleave primers at a length shorter than optimal for priming transcription. Furthermore, despite this optimal length range, primers shorter than ten nucleotides have been observed to prime transcription *in vivo*[1], [7]–[9]. Primers may be selected for at a length slightly shorter than optimal in an attempt to obtain a complimentary nucleotide sequence for priming transcription internally rather than terminally on the vRNA template. The cleavage preference based on length alone is 12 nucleotides 3' of the 5' cap, though this preference is ignored when a guanine residue (complimentary to the cytidine residue at -2 of the vRNA) is present between the 10^{th} and 13^{th} nucleotide[6]. When a shorter host derived sequence is cleaved, PAR may be favored, as RdRp would likely have to compress the template exit tunnel, adopting a strained conformation, to prime transcription with a shorter host derived sequence. Shorter length has been shown to lead to PAR for some members of Bunyavirales[161]–[163], however, length was not the sole determinant of PAR for these viruses.

While length appears to be more important in determining the cleave site[5]–[7], cleavage also appears to be sequence selective. When possible, RdRp appears to preferentially cleave mRNA 3' of a purine residue, to obtain a primer with a 3' end that is complimentary to

the 3’ vRNA template[4]–[7], [14], [195], [196]. RdRp cleavage assays, capped primer competition assays, and deep sequencing all show preferential use of capped substrates that can be cleaved 3’ of a guanine residue, generating a capped oligonucleotide ending in G[4]–[7], [14]. Cleavage at a nucleotide sequence that allows for internal initiation appears to be the preference of the endonuclease, rather than a sequence which allows for terminal initiation, and the terminal nucleotides of the vRNA may not undergo base pairing[4], [5], [7], [138]. Sequence compatibility for internal initiation is an important feature of both *Orthomyxoviridae* and Bunyavirales mRNA cleavage[162], [163], [167], [172], [174].

For IAV RdRp, cleavage 3’ of a guanine residue is most efficient at a 5’GC3’ motif[6] which matches the 3’U**CG** region of the vRNA; this data is consistent with preferential usage of pre-mRNA with a 5’GC3’ motif within the length-determined cleavage range[4], [5]. It had previously been hypothesized that cleavage followed the cytidine residue and base paired with 3’G3 of the vRNA template, but given the cleavage preference this is unlikely. A primer ending in a guanine residue is generally used to prime transcription internally at 3’C2 of the vRNA template rather than terminally at 3’U1[14], though the guanine residue being directed at 3’U1 to prime transcription has also been observed[7]. Priming at 3’C2 is consistent with a recent crystal structure of IBV RdRp superposed on the Norwalk virus polymerase primer-template complex showed that 3’C2 of the vRNA template was in position -1 allowing for priming to occur at this position[110]. The internal initiation strategy is likely selected for as this has a lower free energy than terminal initiation due to the overhanging nucleotide sequence[137]. This preference is seen in replication for cRNA[110], [137], wherein a dinucleotide primer is transcribed *de novo* and then realigned to the 3’ end of the vRNA, and in transcription of mRNA of Bunyavirales. While

priming internally may be important for an energetically favored initiation of transcription, it is currently unknown if internal initiation decreases the probability of PAR *in vivo*.

While cleaving transcripts 3' of a guanine residue appears to be the primary determinant of sequence specific cleavage, there appears to be further preference in the nucleotides residues at both the penultimate and antepenultimate positions within the cap-snatched primer. Adenine is favored as the penultimate position[4]–[7], [14], [195], [196], with cytidine being strongly favored as the antepenultimate position[7], [195], [196]. This leads to priming of transcription with a $\begin{matrix} CAG \\ UCG \end{matrix}$ duplex when mRNA can be cleaved based on the preference of RdRp. Both *in vivo* and *in vitro* primer competition assays have shown that mRNA containing a 5'-AGC-3' motif within the compatible length range were preferentially cap-snatched over primers with a less compatible motif (5'-AGU-3')[4], [5]. Primers with this 5'-AGC-3' motif remained enriched in the IAV mRNA analyzed even when a less compatible primer was supplied at a tenfold higher amount[4], [5]. Selection of both the ultimate and penultimate residues is consistent with formation of base pairs between host primers and vRNAs; however the preference for a cytidine residue at the antepenultimate residue remains to be elucidated. It is tempting to speculate that this cytidine residue may be related to interactions with RdRp in priming transcription, as either adenine or guanine residues would be favored at the position from an energetic perspective[124]. PA mediated cleavage of pre-mRNA does seem to be selective with a preference for nucleotides complimentary to the 3' end of the vRNA template. Similar to the preference for internal initiation, selective cleavage allowing for multiple base pairs upon priming, as well as sequence complementarity 3' of the cleavage site, has also been observed in members of Bunyavirales[162], [163], [167], [172], [174]. It is possible that the variations in structure of these endonucleases confer differences in sequence specificity, and it is possible that the

specificity evolved concomitantly with the 3’ end of the vRNA sequence allowing for cleavage at a sequence that matches the vRNA 3’ end.

Despite the preference for cleaving at a complimentary sequence, not all cap snatched primers have the complimentary guanine residue[1], [7]–[9]. Nucleotide complementary to the first vRNA conserved nucleotide (a uridine) is often absent[4], [5], [7], [138], with transcription being primed with a mismatch nucleotide[6], [7], [138]. This flexibility is related to the uridine residue at position 3’U1, as mutation to an adenine residue greatly reduced this flexibility[138]. This flexibility allows for priming transcription both with a primer ending in a G residue directed at 3’C2 without complementarity to 3’U1, or use of a primer with no sequence compatibility. It has been shown *in vitro* that PAR is related to the nucleotide residue directed at 3’U1 rather than initiating internally[138], in such cases the free energy of the RNA duplex may lead to PAR, though this has not been examined *in vivo*.

1.3.6.4 Cap-snatching is Based on Abundance

When mRNA primers are provided in near equimolar amounts, or in excess, cap-snatching occurs solely on the basis of sequence compatibility with the 3’ end of the vRNA, within the compatible length range[4]–[6]. These analysis were performed either *in vitro*, or by examining only a small and specific population of mRNA that could serve as primers, and does not likely represent the nature of this process *in vivo*. *In vivo*, there is a much larger pool of host transcripts to serve as cap donors, and use of a transcript is subtractive. When an mRNA is used to prime transcription, it is removed from the pool of potential cap donors. Despite limited studies examining cap-snatching on the basis of cap donor availability, this process does appear to be related to the abundance of the mRNA[1], [7].

An initial analysis of cap snatched mRNAs for A/Hong Kong/1/1968 (H3N2) indicated that this process was related to abundance, as there was no significant difference of enriched

gene ontology terms between cap snatched mRNA and host cell mRNA levels[7]; however, non-coding RNAs were not analyzed in this study. Two subsequent studies using different IAV strains (A/WSN/33 (H1N1)[9] and A/Brisbane/59/2007 (H1N1)[8]) showed that mRNA corresponding to snRNAs and snoRNAs were enriched within the cap snatched sequences[8], [9]; however, these enrichments were not analyzed on the basis of abundance. These three data sets, as well as an additional data set (A/Puerto Rico/8/1934 (H1N1)), were reanalyzed by Sikora *et al*., who showed that mRNA are cap snatched on the basis of abundance[1]. When controlling for either nuclear or cellular RNA levels, the enrichment of most transcripts used as cap donors can be explained by the abundance of these RNAs, including the frequent use of snRNAs and snoRNAs[1].

There are, however, some RNAs that are enriched within the cap snatched population for IAV beyond levels which can be directly explained by abundance - particularly protein coding genes related to cellular metabolism, protein localization, regulation of cell cycle, and apoptosis[1]. One possibility is the change in the cellular transcriptome in response to viral infection and host shut-off. Given that Pol II association is key in allowing for cap-snatching, as loss of one of the two Pol II binding sites leads >10,000 fold reduction in viral titers due to a large reduction in cap-snatching activity[2], it is likely the actively transcribed genes are the reservoir for cap-snatching for IAV. IAV upregulates genes involved in metabolism and in positively regulating apoptosis in infected cells[197]–[199]. It is likely that the increased transcription of these mRNA is directly related to their increased use as cap donors.

The other explanation of the selective enrichment is that the enriched RNAs have a compatible nucleotide sequence within the preferred length range for IAV cleavage. Using transfected cells with plasmids containing a known sequence and varying degrees of

compatibility, it has previously shown that relatively less abundant mRNAs with a higher degree of sequence compatibility (5'-GUAUUAAUAUC**AGC**A) can be cap snatched more frequently than a less compatible (5'-GUAUUAAUACCAUUU) relatively more abundant (tenfold higher) mRNA[4]. An enrichment of cap donors that could be cleaved within a 5'-CAGC-3' motif was observed in the A/Hong Kong/1/1968 (H3N2) data set[7], showing that this preference likely maintained in the context of host derived transcripts. Furthermore, the cleavage preference of the PA endonuclease appears to be 3' of a guanine residue within a 5'-GC-3' motif[6]. Taken together these results indicate that there may be a bias towards compatible primers within the abundant primer subset, that lead to either an enrichment or a reduction levels of an RNA species used as a cap donor relative to host expression levels.

The enrichment of a cap snatched sequence appears to be on the basis of abundance. Furthermore, given the relationship of this process to abundance, length and sequence compatibility, it is unlikely that a specific group of transcripts are preferentially cap snatched. The enrichment of a given transcript as a cap donor appears to change on the basis of the RNA expression patterns of the host[1], [197]–[199] or the abundance of an RNA within the cellular region in which cap-snatching is compartmentalized[1], [172], [174]. Given the choice between multiple abundant cap donors, transcripts are cap snatched with a preference towards those with a compatible nucleotide sequence within the preferred cleavage length[4]–[7]. This selection of transcripts with a compatible nucleotide sequence likely leads to deviations from a direct correlation with abundance; however, it is unlikely that sequence and length compatibility cannot result in a given RNA of low abundance being a major cap donor. However, the transcript enrichment of the population of cap snatched mRNA which have undergone PAR has not been analyzed. It is possible that the subpopulation which underwent PAR is of an abundant group of

transcripts that are poor cap donors based on sequence compatibility, yet are still cap-snatched due to their abundance.

1.3.6.5 Priming Transcription and Prime-and-Realign

Once cleaved by PA, cap-snatched primers are threaded into the PB1 active site through the 54 Angstrom exit tunnel[109], [110], [130], [140]. This follows a conformation change from the 'cap-binding' or 'cleavage' configuration into a 'priming' configuration[109]. The cap binding domain of the PB2 subunit rotates approximately 70 degrees, and associates with the PB2 midlink domain pointing the primer towards the active site[109], [130]; this conformation change is stabilized by a salt bridge between PB2 Arg^{423} and PB1 Glu^{277}, and PB2 Asn^{425} and PB1 Glu^{276} residues[109]. In addition to interactions between the 5' cap and the PB2 binding region, there are interactions between the second and third bases within the primer and the PB2 subunit. Ile^{260} stacks on the first base, the aromatic residue at position 432 (His^{432} in human IAV strains) stacks on the second base[109]. The structure of the primer beyond the third nucleotide could not be resolved in any of the crystal structures due to partial occupancy[109], [110], [130]. Further interactions with the 424-loop and cap-627 linker regions of PB2 guiding the primer into the active site are likely[130]. In total, the length the primer must traverse is 54 Angstroms (estimated to be at least 12 nucleotides)[130]. Given that priming of transcription occurs with primers less than 12 nucleotides of length, it would appear as though this distance may be compressed further.

The 3' end of the vRNA template must also be moved into the active site for transcription to be primed. Only the phosphate of 3'U1 has been clearly resolved in the IBV RdRp in the 'priming' conformation, due to partial occupancy of the first five bases of the 3' end of the vRNA[110]. Superposing this structure on the Norwalk virus polymerase primer-template complex showed that 3'C2 of the vRNA template was in position -1 allowing for priming to

occur at this position[110]. To achieve the 'priming' conformation, only small side chain movements are necessary for trafficking of the vRNA template into the active site. Furthermore, the base pair region, wherein 5'11-15 end pair with 3'10-14[23], remains unbroken[110]. Given that this conformation may be an equilibrium state (as noted by the partial occupancy) and that the breaking of the 3' 5' base pair region is not required, it is unlikely that this conformation change is an active process. The base pair region could then be broken at a later point during transcription, wherein new bonds between the 3' end and the primer could offset this energy penalty[110]. The estimated free energy[124] of these bonds range from -7.65 to -9.61 kcal/mol; the free energy of this structure may be responsible, at least in part, for PAR (polymerase stuttering) when transcribing through the first few nucleotides of the vRNA conserved sequences in a manner similar to polyadenylation.

Once in the active site, the primer associates with either 3'U1 or 3'C2, and transcription beings with the addition of GTP or CTP respectively. Base pairing is not a requirement for priming at 3'U1[7], [9], [14], [138]; however, base pairing (ending in a guanine residue) is a requirement for priming at 3'C2[4], [5], [14], [138]. Despite being able to pair with 3'C2, some primers ending in a guanine residue prime transcription at 3'U1[7], [138], though the reason for priming at 3'U1 over 3'C2 in such cases is unknown. Mutational analysis of the 3' vRNA promoter showed that this base pairing for priming at 3'N2 is indeed a requirement. When 3'C2 was mutated to 3'A2, primers that ended in a guanine residue could only able to prime transcription terminally at 3'U1, however primers ending in an uridine residue could prime transcription internally at 3'A2[138]. While priming at 3'C2 requires base pairing, priming at 3'U1 is more flexible; any nucleotide may be used to prime transcription at 3'U1 with or without a guanine residue paired to 3'C2[7], [14], [138]. This flexibility is conferred by the uridine

residue. Mutation to an adenine residue at this position resulted in minimal transcription activity when a primer ending in A (forming an A-A pair) was used as a substrate; whereas a primer ending in U on the wild type template (forming a U-U pair) showed normal transcription activity[138]. While base hybridization between the 3' end of the vRNA and the 3' end of the capped primer is important, it is likely that interactions with RdRp are involved in stabilizing the initiation and early elongation products.

The initiation conformation of RdRp has yet to be solved; however, the PB1 priming loop is close to the active site and may interact with the template[109], [130], [138]. Deletion mutants of the priming loop (residues 631 to 662) do suggest that this region is important in stabilization and transcription of mRNA[138], [139]. Residues between 636 and 642, as well as between 642 and 656 appear to be important for stabilizing transcription though deletion of the priming loop tip (648-AHGP-651) has no effect on the *in vitro* lengthening assay[138]. Deletion of this region abolishes transcription[138], [139]. It is likely that this region is involved in stabilizing the primer template duplex and allowing for elongation of the primer, though the exact interactions remain to be elucidated.

Following priming, transcription of mRNA begins with the addition of a G directed at 3'C2, or, if the primer ends in a guanine residue, the addition of a C directed at 3'G3. The 5' untranslated region of the mRNA therefore consists of a 5' cap, a host derived sequence, a copy of the conserved sequence beginning at the second nucleotide 3'2-GCAAAAGCAGG or 3'2-GCGAAAGCAGG (the uridine residue at position 3'1 is not transcribed), followed by the gene specific sequence[1], [4], [5], [7]–[9], [14], [138]. In addition to this, small sequence repeats of one to six nucleotides consistent with nucleotides two to seven of the 3' vRNA template have

been observed between the host derived sequence and the vRNA conserved sequences which arise due to PAR[1], [4], [5], [7]–[9], [14], [138] (Table 1.2).

The exact determinants of PAR have yet to be determined; however, it is likely RdRp and the 3' 5' base pair region play a role in this process. Disruption of the PB1 priming loop has been shown to increase the frequency of PAR in an *in vitro* assay, disruption of the priming loop was shown to greatly increase the frequency of PAR during mRNA transcription[138], though the exact mechanism is unknown. It is also unknown how the base pairs in the 3' and 5' base pair region of the vRNA are melted to allow for transcription. Using the Turner Energy Rules[124] the free energy of these five base pairs ranges from -7.65 kcal/mol to -9.61 kcal/mol. It is hypothesized that energy from base pairing between the 3' vRNA conserved sequences and the transcription product (in both replication and transcription) is used to facilitate the melting of the base pairs within the 3' and 5' base pair region. Indeed PAR is observed less frequently in genes encoding for a cytidine at 3'4 of the vRNA template[1]. Given that a polymerase stuttering has only been observed at 5' end of mRNA (not cRNA) it is likely that differences between priming in these two processes play a role in PAR. Specifically, in transcription of mRNA, a base pair is not necessary at 3'U1, whereas a Watson-Crick base pair is required at 3'U1 for replication. Given that these processes are mechanistically different, it is likely that differential interactions with RdRp prevent polymerase stuttering in replication. *In vitro* studies, however, have shown that base pairing at 3'U1 does decrease the probability of PAR[138]. It is likely that other characteristics of the cap snatched primer, such as length, overall sequence compatibility, and internal initiation, also affect the probability of PAR.

1.4 Rational and Hypothesis

Based on previous results, cap-snatching by the IAV RdRp is driven by the abundance of transcripts being actively transcribed by Pol II[1]–[3]. Deviations from a direct correlation with abundance do arise, likely from selective cleavage of transcripts with a compatible length and nucleotide sequence[4]–[7]. For IAV, primers are preferentially cleaved between ten and 13 nucleotides in length, though primers outside this range have been observed[1], [4]–[8]. This length preference is related to the 50 angstrom distance from the PB2 cap binding region to the PA endonuclease, though this may result in selection of primers shorter than the 54 angstrom distance from the PB2 cap binding region to the PB1 active site where transcription is primed[130]. Within this compatible length range, RdRp will attempt to generate a primer with multiple bases of complementarity with the 3' vRNA to prime transcription with a $\begin{matrix} CAG \\ UCG \end{matrix}$ duplex[4], [5], [7]. The minimum preference of RdRp is cleavage 5' of a guanine residue[6] so that transcription can be primed internally at 3'C2[110], [138]. Despite this preference, cleavage without a complimentary nucleotide sequence and thus priming transcription in the absence of base pairing have been observed[6]–[9], [138]. Overall, RdRp preferentially cap snatches relatively abundant transcripts with a compatible motif to the 3' vRNA within a range of ten to 13 nucleotides.

Some of these primers are not directly used to transcribe mRNA, but instead undergo a prime-and-realign mechanism. As of yet it is unknown why this process occurs, however, it is likely that these primers are suboptimal for transcription. PAR has been shown to be related to the length of the cap snatched primer for other negative sense single stranded RNA viruses [158], [161]–[163]. This process has also been shown to be related to the nucleotide directed at 3'U1 to

prime transcription [138]. It is possible that a breakdown in length and sequence selective cleavage leads to PAR.

My hypothesis is that the prime-and-realign process is related to the physical characteristics of the primers and their interactions with RdRp and the vRNA template. These physical characteristics include the length, nucleotide sequence, and nucleotide composition of the cap snatched primers. Here, I used four published deep sequencing datasets of the 5' ends of IAV mRNA obtained from IAV infected A549 cells[1], [7]–[9]. I determined the frequency of PAR, and the number of rounds of PAR for each primer. I then examined the physical aspects of the primer and vRNA template and their relation to PAR, specifically: G complementarity and internal initiation, length, free energy from hydrogen bonds and stacking interactions, and relative G+C content. I find that 14.3±0.6% undergo at least one round of PAR primer to transcription. Primers are biased towards PAR on the basis of length, with primers less than 12 nucleotides in length being more likely to undergo PAR. PAR frequencies decrease as the free energy from hydrogen bonds and stacking interactions decreases; these changes are mediated by the base directed at 3'U1 and the pyrimidine purine base pair at position four of the vRNA conserved sequences. PAR typically adds a nucleotide sequence that is at least three nucleotides long ending in an adenine residue; this rescues primers on the basis of length and free energy. An increase of length by three nucleotides and a 3'U1 directed adenine residue will yield a primer that is less likely to require a subsequent round of PAR. For this reason, of the cap snatched primers which undergo PAR, 95.6±2.5% required only one PAR event. PAR therefore rescues suboptimal primers on the basis of length and free energy, and converts this primer into one that is more optimal for transcription.

2. Methods

The PAR process was examined using previously obtained deep sequencing data of cap snatched primers (5' ends of IAV mRNA). A549 cells were infected with one of four different strains of IAV: A/PuertoRico/8/1934 (H1N1; SAMN06011516)[1], A/HongKong/1/1968 (H3N2; SRR1301043)[7], A/WSN/1933 (H1N1; GSE65101)[9], A/Brisbane/59/2007 (H1N1; GSE67493)[8]. mRNA was extracted from these cells, and 5' ends of IAV mRNA were sequenced.

2.1 Summary of Previously Obtained Data Set Preparation

Data from the A/PuertoRico/8/1934 and the human influenza isolate A/HongKong/1/1968 was obtained from A549 cells which were infected with the given strain at a multiplicity of infection of two plaque forming units. mRNA was extracted 4h post infection. Purified mRNA was subjected to 5' rapid amplification of cDNA ends. IAV mRNA was extracted, and tagged with sequence identifiers for multiplexing and deep sequencing. Data from the A/WSN/1933 data set was obtained from A549 cells infected at a multiplicity of infection of 1-3 plaque forming units. mRNA was harvested at an unspecified time thereafter. IAV mRNA was extracted, and subjected to a template switching oligo procedure which added a variable length 5' polyguanine head group to allow for these unique identifiers to be attached. Data from the A/Brisbane/59/2007 (virus originally obtained from the National Institutes of Health Biodefense and Emerging Infections Research Resources Repository, National Institute of Allergy and Infectious Diseases) data set was obtained from A549 cells infected at one in 100 μL of influenza virus growth medium. mRNA was extracted 24h post infection. All mRNA (human and IAV) was cloned by CapSeq[8]. All sequencing was performed with Illumina HiSeq. 2000 Systems.

These data sets were retrieved from the Sequence Read Archive of NCBI by Sikora *et al*.,[1] and reanalyzed. Briefly, the raw data sets for the A/PuertoRico/8/1934 (H1N1) and A/HongKong/1/1968 (H3N2) strains were de-multiplexed using cutadapt v1.8.1 using the gene-specific and RACE primer sequences. Since the A/Brisbane/59/2007 data set contained all cellular mRNA[8], the viral mRNA was extracted using cutadapt v1.8.1 and the sequences of the 5' UTRs of the eight IAV genes. Due to the template-switching procedure used for library production for the A/WSN/1933 strain, mRNA reads with at least three guanine residues on the 5' end were extracted; all guanine residues on the 5' end of these mRNA were trimmed to yield the mRNA sequence. For all data sets except the WSN, the nucleotide sequence 5' of the vRNA conserved sequences (5'-GC[A/G]AAAGCAGG-3') was extracted; these sequences were the host-derived primers. The conserved sequences were not present in the WSN data set, only the host-derived sequences.

2.2 Data Set Preparation by Sikora *et al*.

To determine if these primers had underwent prime-and-realign, a matched trimming strategy was performed (Figure 2.1) by Sikora *et al*[1]. Host derived primers nine nucleotides and longer were matched to a ±100 nucleotide window around transcription start sites annotated in the human reference genome GRCh38 using Bowtie v.1.0.0[200]. If no match was found, the longest single nucleotide sequence which could be added by prime-and-realign (Table 1.2) was trimmed from the 3' end of the host-derived sequence. This iterative matching and trimming procedure was performed until a match was found. Once a match was found, no further trimming would take place, even if trimming was possible. Sequences which could not be matched were discarded from the data set (~10 to 20% of the data)[1]. For example, the sequence 5'-GTTTTCGGCAGCA-3' would first be matched to all sequences ±100 nucleotides (arbitrarily

chosen threshold; mRNA are capped at the first nucleotide transcribed, so cap-snatched primers would originate at correct transcription start cites) from annotated transcription start sites. No match is found for this sequence, so 5'-GCA-3' is trimmed leaving 5'-GTTTTCGGCA-3'. This sequence can be matched to one initiation site of HDAC4 gene, therefore no further rounds of trimming would occur, despite this sequencing ending in "GCA" which may have been added by PAR. In the event of multiple matches, the match closest to the transcription start site was selected; if two or more matches were found and these matches were equidistant from the transcription start site, the individual reads were randomly assigned amongst all matches. For example, the primer 5'-GACGGGGGC-3' which has five matches (ARF5, MCFD2, EXOC7, ZNF623, FAM179B) at the same relative position (±1 of a known transcription start site), will have the 84 reads of this sequence randomly distributed amongst those five matches (14, 15, 18, 17, and 20 to each gene respectively). These matches were then assumed to be correct, despite limitations due to the unknown nature of the true transcription start sites for genes, and a potential preference for a given gene despite multiple matches. Simply, the transcription start sites for the five matches of the primer 5'-GACGGGGGC-3' may not be accurate. It is also possible that the primer 5'-GACGGGGGC-3' was only ever snatched from a small subset of these genes, or another gene entirely for which the transcription start site is incorrect. This method of random matching is the best available strategy at the time of publishing. A graphical representation of this process can be found in Figure 2.1. These data sets were made available by Sikora *et al*[1].

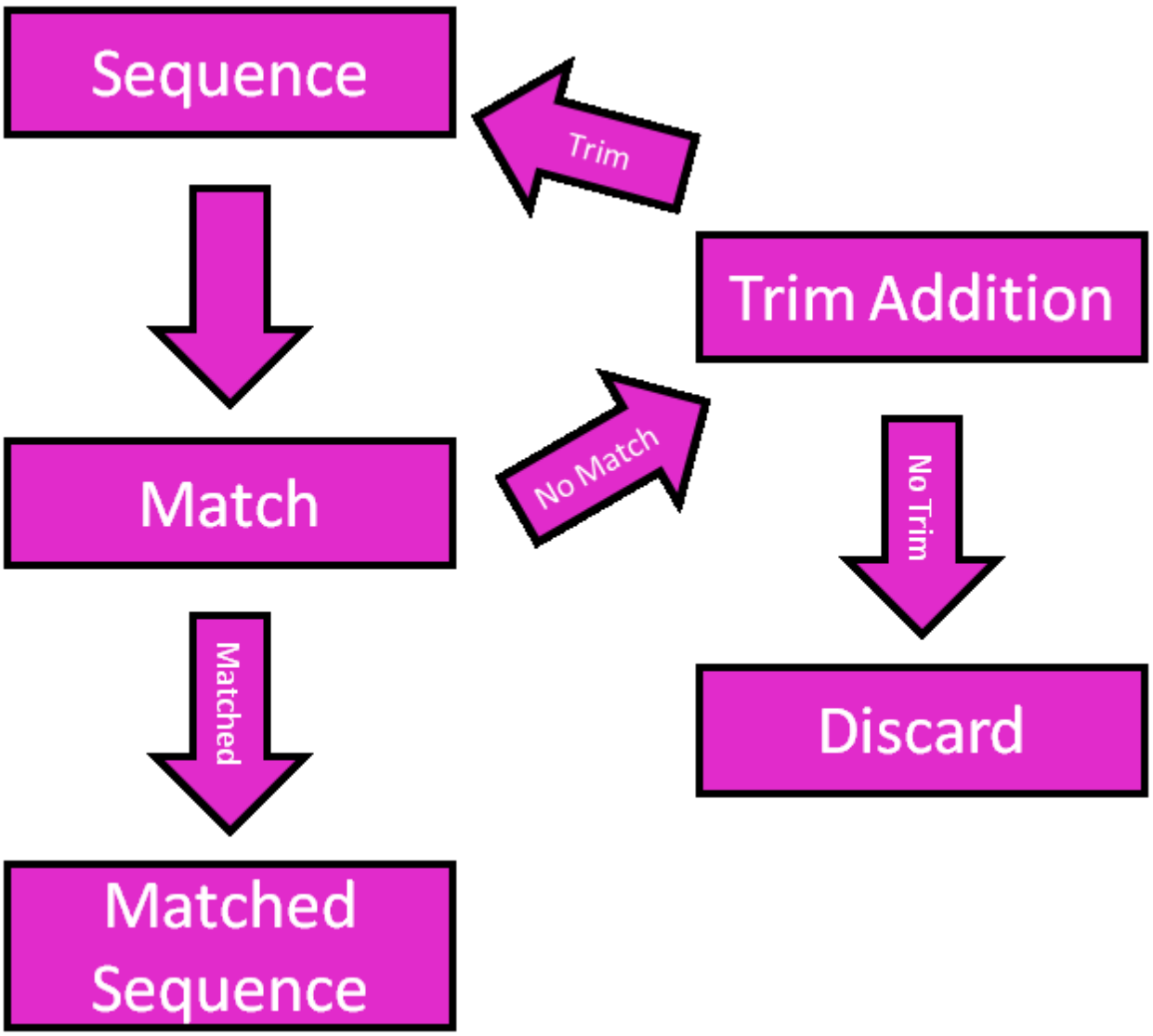
Sequence
Trim
Trim Addition
Match
No Match
No Trim
Matched
Discard
Matched Sequence

Figure 2.1 Matched Trimming Flowchart. Matches for a given mRNA read are searched for within a 100 nucleotide window on either side of the transcription start site in the human reference genome GRCh38. If the sequence is matched, no further trimming occurs; if the sequence is not matched, the longest possible addition from PAR is trimmed. If the sequence cannot be trimmed it is discarded; if the sequence can be trimmed, matches for the trimmed mRNA are searched for within a 100 nucleotide window on either side of the transcription start site.

2.3 Pre-Processing of Obtained Data

Further data analysis was performed using custom R scripts with R v.3.5.1[201] using R Studio v.1.1.383[202]. The R packages doParallel[203] and compiler[201] were used for data processing in most scripts. References for used R packages were generated using the knitr package[204]. Data analysis was performed on a custom desktop computer with an Intel Core i9-7960X @ 2.80GHz (32 CPUs), 128GB RAM, and a 200GB page file. A copy of the scripts used for this analysis are available at https://github.com/cdevlugt/iav_par

To reduce noise in these data sets, only cap snatched primers with at least ten reads (arbitrarily chosen threshold) matched to a given gene were included. Data was further limited to primers matched between nine and 16 nucleotides in length. The WSN data set had a duplicate of the NS reads, this duplicate was removed to avoid biasing the data set towards the duplicated gene.

The nucleotide sequences added by PAR (removed by the iterative matching procedure) were not directly available in the published data sets. A recursive algorithm was used to iteratively extract the individual nucleotide sequences that were trimmed. For example, the published data sets contained the read sequence (5'-ACAGAGCGCGCAGCAAA), the matched sequence (5'- ACAGAGCGC), and the round of matching in which this primer was matched (3). Using the number of rounds, it was determined that a GCAAA addition had been removed, followed by a GCA addition. These additions were then reversed to order them as they would have been added by PAR rather than removed by the matching procedure. Simply, 5'-ACAGAGCGC underwent one round of PAR adding a GCA followed by a subsequent round of PAR adding a GCAAA, which lead to the sequence 5'-ACAGAGCGCGCAGCAAA which did not undergo PAR. The number of rounds of matching (3) was reduced by one to indicate the rounds of PAR (2) that a given primer underwent prior to transcription. With this notation,

primers directly matching TSSs would be listed as undergoing zero rounds of PAR. This algorithm also extracted the added nucleotide sequence (5'-GCAGCAAA), the length of the read primer (the 17 nucleotide length of 5'-ACAGAGCGCGCAGCAAA), the length of the trimmed primer (matched primer; the nine nucleotide length of 5'- ACAGAGCGC), the length of the added sequence (the eight nucleotide length of 5'-GCAGCAAA), the average length added per round (four nucleotide average of three and five nucleotides), and the number of nucleotides added each round. This data set was used for all subsequent data analysis, and can be found in the script 0 – Prime and Realign; an example of this table can be found in Table 2.1.

HA	M	NA.	NP	NS	PA	PB1	PB2	ENSEMBL	index	Sequence	Chr	seq_length	pos_relative	Seq_Bowtie	NT_coupure	rounds
0	0	0	3766	0	0	0	0	ENSG00000000460	5446	AAGCAGGCTGCA	chr1	12	-27	AAGCAGGCTGCA	TGAGC	0
0	165	0	0	0	0	0	0	ENSG00000163040	10918	ACCAGGCGTCGCAGCAAAAGCA	chr2	10	21	ACCAGGCGTC	TGTTT	3
50	39	0	4027	0	0	0	0	ENSG00000001167	1095	AGTGGCGGCGGCA	chr6	13	4	AGTGGCGGCGGCA	GCGGC	0

Nb_Same_Pos	SeqCompl	NTCompl	transcript_id	gene_name	transcript_biotype	strand	start	end	mini_genome
1	GCA		ENST00000498289	C1orf112	processed_transcript	+	169662007	169853085	AGCACCCTGTCAAAACAGACCACTTGACTCTACCAATCAGCAGGATGTGGGTGGGGCCAGATAAGAGAA TAAAAGCAGGCTGCATGAGCCAGCAGTGGCAACCCGCTCGGGTCCCCTTCCACACTGTGGAAGCTTTGTT CTTTCGCTCTTTGCAGTAAATCTTGCTGCTGCTCACTCTTTGGGTCCACACTGCCTTTATG
1			ENST00000434330	CCDC74A	protein_coding	+	131529181	131532941	GATGTGGCTCCGCTCCTCGATCTTGCTTCAGGAGGGCTCCAGCAAGGGCACCAGTGCCCGCGGCCTCCTC TCCCGCCCCGTTCCCTGGTGCTGCCTTACCTGAGTGCCGAGTATCCCATCACCAGGCGTCTGTTTTGTCAG GACCCCGCTTCCCCATGTCTTCTTGAGCCTCAGGACGCCCAGTTCCTAATGTCTTCTTG
1	GCA	GCG	ENST00000353205	NFYA	protein_coding	+	41072983	41097950	AACTGGAGTTAGTGGGCGTGGCCTCAGAGGCGCCTGCGCGGAGGCGGTTGGAGGGAGGCCCGATTCCC CTTTGTTCGGGTTCGCCATTTTGCTAGGCAGCGGCAGTGGCGGCGGCAGCGGCGGCTGGAGCCTCTGAT TGGGTTTCGGAGTCCGGTACTGGAGCCAATCAGCGCGGGCAGCGAACCGGGGGAGCGAGGCAC

transcript_name	length_transcript	transcript_support_level	Trim_Sequence	Added_Sequence	Seq_Trim_R1	Seq_Trim_R2	Seq_Trim_R3
C1orf112-004	191078	2	AAGCAGGCTGCA				
CCDC74A-009	3760	5	ACCAGGCGTC	GCAGCAAAAGCA	GCA	GCAAAA	GCA
NFYA-002	24967	5	AGTGGCGGCGGCA				

Length	Trim_Len	Len_Adjust	Velocity	Num_Trim_R1	Num_Trim_R2	Num_Trim_R3
12	12	0	0	0	0	0
22	10	12	4	3	6	3
13	13	0	0	0	0	0

Table 2.1. Example of Organization of the Data for the Puerto Rico Data Set with Example Primers. HA, M, NA., NP, PA, PB1, PB2 columns indicate the number of reads of the given primer matched to that IAV segment. ENSEMBL, index, Chr, transcript_id, gene_name, Transcript_biotype, strand, start, end, transcript_name, length_transcript, transcript_support_level are data about the gene, and the match; these columns were inherited from the data set prepared by Sikora *et al*.,[1], and were not used in this analysis. Sequence is the mRNA sequence that was read. Seq_Bowtie and Trim_Sequence both indicate the matched sequence. NT_coupure indicates the 5 nucleotides 3' of the end of the matched sequence. mini_genome is the ±100 nucleotide window around the transcription start site of the gene to which the primer was matched. Added_Sequence indicates the entire nucleotide sequence that was presumed to be added by PAR. Seq_Trim_R# columns indicate the sequences added in each round of PAR. seq_length and Length are the length of the read primer. Trim_Len is the length of the matched primer. Len_Adjust indicates the total length adjustment due to PAR. Velocity indicates the average number of nucleotides added each round. Num_Trim_R# indicates the number of nucleotides added in each round of PAR. rounds indicates the number of rounds of PAR (number of rounds of trimming prior to the sequence being matched).

2.4 Primer Classification

Primers matched within the ±100 nucleotide window without any rounds of trimming were categorized as "No Realignment". Primers which were matched after at least one round of trimming were classified as "Realigned". For the purposes of some analysis, the Realigned primers are further classified as either "Pre-Realignment" (matched sequence) or "Post-Realignment" (read sequence). For example, the sequence 5'-GTTTTCGGCAGCA-3' which was matched as 5'-GTTTTCGGCA-3' would be classified as "Realigned", with the matched version (5'-GTTTTCGGCA-3') being further classified as "Pre-Realignment", and the read version (5'-GTTTTCGGCAGCA-3') further classified as "Post-Realignment".

2.5 Prime-and-Realign Frequency Determinations

Matched sequences were classified into the Realigned and No Realignment subpopulations as described. The total number of primers in each of these subpopulations as well as the total number of primers was extracted from each strain. The PAR frequency was then calculated as the number of primers which underwent at least one round of trimming prior to matching (the Realigned subpopulation) divided by the total number of matched primers. The frequency of primers which did not undergo PAR was calculated in a similar method. These data were tabulated for each strain and displayed in Table 3.1.1. This data analysis was performed in the script 1 – PAR Frequencies and Rounds.

2.6 Prime-and-Realign Rounds

Matched sequences were classified into the Realigned and No Realignment subpopulations as described. The number of primers remaining after each round of PAR was extracted, and displayed as a percentage of the total primers. For example, the primer 5'-GGTAGCGCGGCAGCAGCAA-3' was matched after three rounds of trimming (5'-GCAA, 5'-

GCA, 5'-GCA) as 5'-GGTAGCGCG-3' to the TSS of RBL1 gene. This primer would be counted in the groups of primers which underwent PAR in round one, two, and three; since the fourth attempt resulted in transcription of mRNA it would not be counted with the primers that underwent four rounds of PAR. This data was then tabulated and was displayed in Table 3.2.1.

Since the percentage of primers that underwent a given number of rounds of PAR appeared to follow an exponential decrease with rounds (Table 3.2.1), the natural logarithm of the number of primers that remain in each round was taken. The $\log_e$ of the frequency was correlated to the round of PAR, between one and the maximum number, by fitting the data to a linear model using the lm() function in R. These data were also graphed using the R packages ggplot2[205] and cowplot[206], and were displayed in Figure 3.2.1. This data analysis was performed in the script 1 – PAR Frequencies and Rounds.

2.7 Primer 3' Sequence Ends and the 3'C2 Directed Guanine

Cleavage 3' of a guanine residue has been previously reported for the IAV RdRp[6]. This guanine would not be present within the matched sequence. For example, the primers 5'-GACGGAAGAACA that would prime transcription with a $\begin{matrix} CA \\ UCG \end{matrix}$ duplex and 5'-GACGGAAGAACA**G** that could prime transcription with a $\begin{matrix} CA\boldsymbol{G} \\ UCG \end{matrix}$ duplex would be indistinguishable after a guanine residue directed at 3'C2 has been transcribed on the first primer to give a $\begin{matrix} CAG \\ UCG \end{matrix}$ duplex. The presence of this guanine residue was extracted from the reference genome as the nucleotide immediately 3' of the end of the matched sequence. In the data sets obtained from Sikora *et al*[1], this is the first nucleotide in the column "NT_coupure" which contains the five nucleotides immediately 3' of the end of the matched sequence in the reference genome. The last three nucleotides of the matched sequence and the three nucleotides

immediately 3’ of the cap snatched sequence were extracted. This was performed for all primers, and those in the Realigned and No Realignment subpopulations. These data were tabulated and also graphed using the R packages ggplot2[205], cowplot[206], and ggeqlog[207], and displayed in figure 3.3.1. This data analysis was performed in the script 3 – Nucleotide Enrichment.

PAR frequency of primers with or without a guanine residue one nucleotide 3’ of the end of the matched sequence was extracted. Primers were classified as either Realigned or No Realignment, and then these groups were further classified based on the presence of a guanine residue one nucleotide 3’ of the end of the matched sequence. The PAR frequency was then calculated as the total number of primers in the Realigned subpopulation in each subcategory (G, other) divided by the total number of primers in that subcategory. These data were tabulated and graphed using the R packages ggplot2[205] and cowplot[206], and were displayed in figure 3.3.1. The strain specific PAR frequency was overlaid on this graph. Fold change values were calculated using the strain specific values and then averaged. This data analysis was performed in the script 3 – Nucleotide Enrichment.

To see if the abundance of nucleotide residues at the penultimate (-2), ultimate (-1), and immediately 3’ of the end of the matched sequences (+1) were due to the abundance of these residues within the genome and random selection, or potentially due to a selective process. I performed a 100,000 iteration Monte Carlo procedure using the R doParallel package[203]. 100,000 iterations was chosen due to the large size of the data set. While a larger number of iterations would have resulted in better resolution of the p-value, the computational time for a 10 fold increase in iterations would have required additional weeks for a minimal increase in resolution. First I calculated the actual enrichment of each nucleotide at these three positions (-2, -1, and +1). Next I simulated random selection of nucleotide residues within the reference

genome ("mini_genome" column of the data set obtained from Sikora *et al*[1]). The frequency of each nucleotide being selected was saved for each of these 100,000 iterations. From these simulated values, the mean and standard deviation of the simulated enrichment was obtained. Using this sample distribution, p-values were calculated as the number of iterations wherein the actual nucleotide enrichment was greater than (if the actual value were higher than the simulated mean) or less than (if the actual value were lower than the simulated mean) the actual value. For example, a cytidine residue is found in the Realigned subpopulation at position +1 (immediately 3' of the end of the matched sequence) in 27.8% of primers. The mean of the 100,000 simulated values was 27.7%. Since the actual value (27.8%) is higher than the simulated mean (27.7%), all iterations where a value higher than the actual value will be used to determine the p-value. A value higher than 27.8% occurred in 46,679 of the 100,000 iterations, therefore the p-value is 0.467 (rounded). The sample distribution was compared to the nucleotide abundance of the overall population, and the Realigned and No Realigned subpopulation using the strain specific randomly generated values. The Monte Carlo procedure was performed by script 3 – Nucleotide Enrichment Monte Carlo.

To examine selective cleavage of a dinucleotide and trinucleotide sequence, I performed a Monte Carlo procedure as above, but examining the selection of up to four nucleotides. This Monte Carlo procedure was performed by script 3 – All Cleavage Sites.

The results of these Monte Carlo procedures were tabulated using the script 3 –Monte Carlo Tables. This script also obtained the PAR frequency for each of these nucleotide residues at the given position. These data were presented in Table 3.3.1, Table 3.3.2, Table 3.3.3, Table 3.3.4, Table 3.3.5, and Table 3.3.6.

2.8 Primer Length Determinations

The length of primers was determined for both the matched sequence and the read sequence using the script 0 – Prime and Realign. Since the enrichment of guanine one nucleotide 3' of the end matched sequences was non random (Table 3.3.1), this guanine was factored into the length of primers as applicable. For example, the 12 nucleotide long primer 5'-GACGGAAGAACA had a guanine residue immediately 3' of the end of this sequence on the mini-genome used, therefore this sequence assumed to be 3'-GACGGAAGAACA**G** and had the length increased by one nucleotide to 13 nucleotides. The length of the cap snatched primers was examined for both the Realigned and No Realignment subpopulations. Since PAR adds a nucleotide sequence to the cap snatched primer, increasing the length, I classified the Realigned subpopulation into two groups: Pre-Realignment – the matched cap snatched primer; Post-Realignment – the cap snatched primer prior to the iterative trimming process. For example, the sequence 5'-GTTTTCGGCAGCA (13 nucleotides), which was matched as 5'-GTTTTCGGCA (ten nucleotides) would be classified as Realigned, with the trimmed 5'-GTTTTCGGCA (ten nucleotides) being further classified as Pre-Realignment, and the read sequence 5'-GTTTTCGGCAGCA (13 nucleotides) further classified as Post-Realignment. The abundance of cap snatched mRNA in each of these groups was extracted relative to the total number of cap snatched primers. These data were also graphed using the R packages reshape2[208], ggplot2[205], and cowplot[206], and were displayed in Figure 3.4.1, using the script 4 – Length + G Populations. The mean, standard deviation, and confidence intervals of these length ranges for each of these three subpopulations were calculated and tabulated using the script 4 – Length + G Stats. This script also performed t-tests of the three different subpopulations, using the R function t.test(). These data were displayed in Table 3.4.1.

PAR frequency was examined as a function of length. Matched primers were classified as No Realignment or Realigned (this would be the Pre-Realignment group, "Trim_Len" column in the data set), and then further classified on the basis of length factoring in the presence of the complimentary guanine, as applicable. The PAR frequency for each length was then calculated as the number of matched primers which underwent PAR at that given length relative to the total number of matched primers at that given length. These data were graphed using the R packages ggplot2[205] and cowplot[206], and were displayed in Figure 3.4.2. A horizontal black line was added to each graph representing the strain specific PAR frequency. Fold change values were calculated using the strain specific values and then averaged. A copy of these data was also tabulated and saved. This data analysis was performed using the script 4 - Length + PAR Frequency.

These PAR frequencies seemed to be following an exponential decrease. The saved data from the script 4 - Length + PAR Frequency, was analyzed on a logarithmic scale. The natural logarithm of the PAR frequency was calculated and graphed in relation to the length of the primer factoring in the complimentary G. These data were correlated by fitting the data to a linear model using the lm() function in R, and graphed using the R packages ggplot2 and cowplot. These data are in figure 3.4.3. This analysis was repeated between nine and 13 nucleotides, omitting longer primer lengths; these data are in figure 3.4.4. These analyses were performed by the script 4 - Length + G Correlations.

2.9 Nucleotide Sequence Additions

The nucleotide sequence added with each PAR event was examined. Since length was a key determinant of PAR, this process was examined as a function of the length of the primer prior to PAR. Since the No Realignment subpopulation does not undergo PAR, it was omitted from these

analyses. For the Realigned subpopulation, the nucleotide sequence added (G to GC[A/G]AAA) at each PAR event was extracted. Some sequences underwent multiple rounds of PAR, and each of these PAR events was extracted and treated as an independent event. For example, a primer originally nine nucleotides in length that underwent two rounds of realignment having three nucleotides and four nucleotides added for a final length of 16 nucleotides will be examined as a primer of nine nucleotides lengthened to 12 nucleotides, and again as a primer of 12 nucleotides lengthened to 16 nucleotides. These data were then analyzed relative to all PAR events, and graphed. Additionally, the frequency of each length prior to PAR was extracted and graphed, and the frequency of each nucleotide sequence addition was extracted and graphed. These data were graphed using the R packages ggplot2[205], cowplot[206], and ggpubr[209], and displayed in Figure 3.5.1. This analysis was repeated graphing the data relative to the number of PAR events at a given length rather than the total number of PAR events; these data were displayed in figure 3.5.3. These analyses were performed by the script 5 - Sequence Addition and Length Heatmaps. The length of these sequence additions was then correlated by calculating the Pearson's correlation coefficient using the cor.test() function in R. This analysis was performed by the script 5 - Sequence Addition Correlations.

For further data analysis, it was necessary to tabulate the nucleotide sequences added, the total number of PAR events for that given addition, and the frequency. This analysis was performed for all PAR events as above. Primers were further categorized as all primers, the gene that the primer was used to transcribe, and the conserved sequence of that gene. For example, the primer 5'- ACACTGAGCG, which underwent a PAR addition of GCA, had 6014 reads mapped to the NS gene and 11 reads mapped to the PB1 gene. The GCA addition of this primer would contribute 6025 GCA additions into the "All" group, 6014 into the NS group, 11 GCA additions

into the PB1 group, 6014 GCA additions into the 3'U4 group (the NS gene has the 3'U4 conserved sequences), and 11 GCA additions into the 3'C4 group (PB1 has the 3'C4 template. This primer would contribute zero reads to the other six groups for the other six genes. These data were tabulated and saved by the script 5 – Sequence Addition Table.

Using the data obtained from the script 5 – Sequence Addition Table, the frequency of sequence additions of a given length and the length of these sequence additions was extracted for all PAR events. It was observed that PAR events between three and six nucleotides appear to decrease in frequency exponentially. The natural logarithm of the frequency of all additions was taken, and graphed in relation to the length of these additions using the R packages ggplot2[205] and cowplot[206]. The logarithm of the frequency and the length of the addition were correlated using the lm() function for additions between three and six nucleotides in length and overlaid on this graph. These data are in Figure 3.5.2. This analysis was performed using the script 5 - Sequence Addition Correlations.

The effect of the vRNA conserved sequences on the nucleotide sequence addition was examined. The data from the script 5 – Sequence Addition Table, the nucleotide sequence addition frequencies for All primers and those with the 3'U4 and 3'C4 templates were extracted. These data were graphed using the R packages ggplot2[205] and cowplot[206], and displayed in Figure 3.8.1.

2.10 Free Energy Determinations

Free energy was calculated using the Nearest Neighbor method[124], [210], [211]. This method takes into account the free energy from both hydrogen bonds and stacking interactions of base pairs. The highest free energy values (least stable) were used for these calculations, and an initiation penalty (normally 4.31 kcal/mol) was omitted. For the purposes of these calculations,

A-C pairs are considered to be a mismatch. When a mismatch is incorporated into an RNA duplex, the proximity to both the 3' and/or 5' ends determines the free energy[210], [211]; this is assumed to be linear. Note: since only three nucleotides of transcription (ex. A $\begin{matrix} NNGCA \\ UCGCU \end{matrix}$ duplex) were analyzed, incorporation of the AC mismatch was not factored into any of the displayed free energy values. An example calculation can be found with equation 1.

Equation 1:

$$\Delta G^o_{37} = \Delta G^o_{37}(Watson - Crick) + \Delta G^o_{37}(dangling\ ends) + \Delta G^o_{37}(terminal\ mismatch) + \Delta G^o_{37}(end\ penalty) + (\Delta G^o_{37}(mismatch) + \Delta G^o_{37}(5'shift) + \Delta G^o_{37}(3'shift))$$

For the GCA addition to the GTTTTCGGCA primer to the UCGUU portion of the 3' conserved sequence $\begin{matrix} CAGCA \\ UCGUU \end{matrix}$:

$$\Delta G^o_{37} = \Delta G^o_{37}\left(\begin{matrix} AG \\ UC \end{matrix} + \begin{matrix} GC \\ CG \end{matrix} + \begin{matrix} CA \\ GU \end{matrix}\right) + \Delta G^o_{37}\left(\begin{matrix} CA \\ U \end{matrix} + \begin{matrix} A \\ UU \end{matrix}\right) + 0\,{}^{kcal}/_{mol} + \Delta G^o_{37}\left(2 * \begin{matrix} A \\ U \end{matrix} end\right) + 0\ {}^{kcal}/_{mol}$$

$$\Delta G^o_{37} = (-2.02 + -3.36 + -2.04)\,{}^{kcal}/_{mol} + (-0.3 + -0.2)\,{}^{kcal}/_{mol} + 0.98\,{}^{kcal}/_{mol}$$

$$\Delta G^o_{37} = -6.92\ {}^{kcal}/_{mol}$$

I calculated the free energy for all possible RNA duplexes following a three nucleotide sequence addition (all $\begin{matrix} NNGCR \\ UCGYU \end{matrix}$ duplexes where N is a nucleotide, R is a purine, and Y is pyrimidine). I then extracted all transcription events. For example, a primer which underwent two rounds of realignment having a GCAA nucleotide sequence addition, followed by a GCA nucleotide sequence addition, then the transcription of the conserved sequences without PAR, will be treated as three separate transcription attempts (PAR1, PAR2, and transcription). The No

Realignment and transcription subpopulation of the Realigned subpopulation were then combined into the Transcribed population. Note: analyzing the first transcription attempt only resulted in similar numbers; all transcription events where used as this reflects the all of the transcription rather than just the first attempt. Since I was interested in the free energy of all primers following a three nucleotide sequence addition, all transcription events of less than three nucleotides were omitted from these data. These transcription attempts were classified into three categories: Will Realign (those that undergo PAR after four or more nucleotides have been transcribed); Realigns (those that realign after three nucleotides have been transcribed), and Transcribed (those that did not undergo PAR). For example, a primer in the Realigned population which underwent PAR after the addition of a GCA would be classified as Realigns; a primer in the Realigned population which underwent PAR after the addition of a GCGAA would be classified as Will Realign. The last two nucleotides primer to the transcription attempt were extracted as well as the nucleotide sequence transcribed. Nucleotide sequences transcribed were either the addition from PAR, or assumed to be the complimentary sequence to the first three nucleotides of the vRNA template (GCA for 3'U4 templated genes, and GCG for 3'C4 templated genes). The template sequence, forming the other half of the RNA duplex, was assumed to be the correct sequence for each gene (3'UCG[C/U]U). The free energy of each of these duplexes was then calculated. For example, the sequence 5'-ACACTGAGCG which underwent PAR following a GCA addition had reads mapped to both the NS gene and the PB1 gene. For the NS reads the RNA duplex at three nucleotides would be $\frac{CGGCA}{UCGUU}$ when transcribing the NS gene, and $\frac{CGGCA}{UCGCU}$ when transcribing the PB1 gene. The free energy of all primers was graphed using the R packages ggplot2[205] and cowplot[206]. mRNA within a subpopulation were aggregated based on free energy, and the transparency of each point reflects the number of mRNA that had

the given free energy. The mean free energy and the 95% confidence interval (mean ±2 standard deviations) were overlaid on each series. These data were graphed and tabulated using the script 6 – 3nt Addition Energy, and displayed in Figure 3.6.1 and Table 3.6.1. P-values were obtained using the t.test() function in R.

The free energy differences associated with the vRNA conserved sequence variation at position four affected PAR the frequency was examined. The free energy when transcribing from 3'C2 to 3'Y4 was calculated for all four possible combinations ($\frac{GCA}{CGUU'}$ $\frac{GCA}{CGCU'}$ $\frac{GCG}{CGUU'}$ $\frac{GCG}{CGCU}$ duplexes). The nucleotide directed at 3'U1 was omitted due to the variability in free energy among the three nucleotide residues directed at it; simply, we examined only the effects of the vRNA template. These free energies were calculated and tabulated using the script 6 – Conserved Sequence Energies.

The PAR frequency for each subunit and vRNA conserved sequence transcribed was examined. Primers were classified into both the Realigned and No Realignment subpopulations, and then further classified on the basis of gene segment transcribed, and conserved sequence transcribed, in a similar manner as was performed by the script 5 – Sequence Addition Table. The PAR frequency was then determined by dividing the primers in the Realigned Subpopulation by the total number of primers for the given gene segment or conserved sequence. These data were graphed using the R packages ggplot2[205] and cowplot[206], and are displayed in Figure 3.7.1. The strain specific PAR frequency was overlaid on these graphs. The PAR frequencies were also saved for each gene and conserved sequence transcribed. Fold change values were calculated using the strain specific PAR frequencies, and then averaged. This data analysis was performed using the script 6 - Segment and Conserved Sequence PAR Frequencies.

The differences in PAR rates could be due to the different base pairs in the 3' 5' base pair region was examined. These sequences were retrieved from the NCBI database for the Hong Kong strain (KY321924, KY321925, KY321926, KY321927, KY321928, KY321929, KY321930, KY321931)[212]. The free energy of these regions was calculated using the Nearest Neighbor method[124], as in equation 1. The PAR frequencies for the gene segments of the Hong Kong strain from the script 6 - Segment and Conserved Sequence PAR Frequencies were correlated, using the lm() function, with the free energy from the 3' 5' base pair region of the respective gene segment. This analysis was repeated factoring in the energy difference following three nucleotide of transcription ($\binom{GCR}{CGYU}$ duplex), either in a subtractive manner, or as a ratio. These data were tabulated and saved along with the R^2 values; the tabulated data is in Table 3.7.1. These analyses were performed by the script 6 – 3' 5' Base Pair Region and PAR.

The free energy of each 3'U1 directed nucleotide reside was examined. Since the Nearest Neighbor method is calculated based on dinucleotide duplexes rather than a single base pair, these free energies were calculated assuming a guanine paired with 3'C2 ($\binom{NG}{UC}$ duplex). Adenine and guanine residues both obtain a free energy decrease from a dangling end. The maximum value was used in these calculations (-0.2 kcal/mol), and added to the values for adenine and guanine.

The PAR frequency of each 3'U1 directed nucleotide both with and without a distinction for the complimentary guanine residue was examined. Primers were classified as either Realigned or No Realignment, and then further classified based on the last nucleotide of the matched sequence (3'U1 directed nucleotide). The PAR frequency was then extracted. This analysis was repeated with further distinction for the presence of the complimentary guanine (a guanine one nucleotide 3' of the end of the matched sequence). These PAR frequencies were

tabulated using the script 6 – 3'U1 Directed Nucleotide PAR Frequencies. The PAR frequencies saved by the script 6 – 3'U1 Directed Nucleotide PAR Frequencies were graphed as a function of the 3'U1 directed nucleotide using the R packages ggplot2[205] and cowplot[206]. These PAR frequencies were averaged and these data were added to the graph. These PAR frequencies were then correlated to the previously calculated free energy of each 3'U1 directed nucleotide using the lm() function in R. Fold change values were calculated based on the strain specific values, and then averaged, if applicable. These data are displayed in Figure 3.9.2. This analysis was repeated making the distinction for the complimentary guanine, and displayed in figure 3.9.3. The distance between the lm() models with and without a guanine directed at 3'C2 for the Average group (average of all four strains) was analyzed to determine the free energy difference the complimentary guanine would confer. We used the PAR frequencies generated by the linear model of primers with a G to predict the free energy using the model of primers without a G; simply, the average free energy difference between the two linear models was calculated. This was performed using the predict() function in R. These analyses were performed by the script 6 - 3'U1 Directed Nucleotide Correlations.

The selection of nucleotides at the end of the matched sequence was examined. The abundance of each of these nucleotides at the end of the matched sequence had been saved by the script 3 - Nucleotide Enrichment Monte Carlo, and this data was retrieved for this analysis. The abundance of each 3'U1 directed nucleotide was graphed in relation to the free energy using the R packages ggplot2[205] and cowplot[206]. These values were averaged, and added to the graph. The abundance of each nucleotide was correlated to the free energy using the lm() function, and these data were added to the graph. These data are in figure 3.9.1, and were obtained using the script 6 - 3'U1 Directed Nucleotide Correlations.

2.11 Relative Guanine and Cytidine content

The relative G+C content of the primers was examined. Primers were grouped into the No Realignment, Pre-Realignment, and Post-Realignment groups as previously described. The number of guanine and cytidine (G or C) residues within each primer was extracted; these values were divided by the length of the primer to obtain the relative G+C content. I also extracted the relative G+C content of the reference mini-genome ("mini_genome" column) in the same manner as the primers. The G+C content was graphed for the genome and the three subpopulations using the R packages ggplot2[205] and cowplot[206]. Multiple primers had the same relative G+C content within a subpopulation, and these values were aggregated; the transparency of each point in the population reflects this aggregation, with more transparent values have less aggregated primers. The mean G+C content and the 95% confidence interval (±2 standard deviations) were overlaid for each subpopulation and the genome. These data are graphed in figure 3.10.1, and tabulated in Table 3.10.1. The data were graphed using the script 7 – GC Content Stats, and tabulated using the script 7 – GC Content Stats. P-values were obtained using the R function t.test().

The PAR frequency in relation to G+C content was examined. Due to the large variability, G+C content was clustered in 10% increments. Primers were then classified into the Realigned and No Realignment subpopulation, and then further grouped by the G+C content of the matched sequence ("Trim_Sequence" column) in these 10% increments. PAR frequencies were then calculated as the number of primers within the Realigned subpopulation relative to all primers within that G+C content increment. These data were then graphed using the R packages ggplot2[205] and cowplot[206]. The strain specific PAR frequency was overlaid as a horizontal black line. These data are in Figure 3.10.2, and this analysis was performed by the script 7 - GC Content PAR Frequencies.

2.12 Effect Modeling

The effects of the four key variables on PAR together were examined. These variables were the length, 3'U1 directed nucleotide, 3'C2 directed guanine, and the vRNA conserved sequences. Since length values for the WSN strain were not indicative of the actual length of some of the primers due to the template switching oligo procedure performed, this data set was omitted from this analysis. Matched primers were divided into the Realigned and No Realignment subpopulations. Primers were then further grouped based on their length and nucleotide sequence used to prime transcription. The complimentary guanine was extracted as previously described and factored in to both the length and the nucleotide sequence used to prime transcription. For example, the sequence read as 5'-ACAGAGCGCGCA was matched as 5'-ACAGAGCGC following the trimming of a GCA; this primer was matched to the CYP51A1 gene, and has a guanine residue immediately 3' of the matched sequence. This primer would be classified as Realigned having undergone one or more round of PAR. This primer would be further grouped on the basis of its length factoring in the complimentary guanine (5'-ACAGAGCGC**G**; ten nucleotides) and the sequence used to prime transcription factoring in the complimentary guanine (5'-CG); therefore this primer would be grouped with those ten nucleotides in length ending in CG. The PAR frequency was then calculated as the number of primers in the Realigned subpopulation within this group divided by the total number of primers within this group. This analysis was repeated for mRNAs transcribed on the 3'U4 template and 3'C4 template. This tabulation was performed by the script 8 - Effect Modeling Probability Tables.

The data generated by the script 8 - Effect Modeling Probability Tables was then further analyzed. To reduce the noise of these data, thresholds were applied. For all data a threshold of $1/6400^{th}$ of the total reads for each strain was used (1% times the 64 possible groupings). Since more PAR occurred on the 3'U4 template this threshold was increased to $1/1200^{th}$ of reads. This

threshold was decreased similarly for the 3'C4 template to 1/12800th of reads. Following application of these thresholds, the PAR frequencies for the Puerto Rico, Hong Kong, and Brisbane strains were averaged. If a value was absent from a strain, that value was omitted from the average, and the PAR frequency displayed reflects the average of the remaining strains. These data were graphed as a heatmap using the R packages ggplot2[205] and cowplot[206]. The coloring scale of the heatmap was generated such that the color changed consistently with the fold change from the average PAR frequency of the three strains. This analysis was performed using all primers, and those transcribed on either the 3'U4 or 3'C4 templates. These three graphs are in figure 3.11.1. These analyses were performed by the script 8 – Heatmaps of PAR Frequencies.

3. Results

3.1 Prime and Realign Frequencies

Prime-and-realign (PAR) frequencies between 10% and 70% have been reported for Group V viruses that prime transcription of mRNA using a cap snatched primer[138], [163], [166], [167]. Rice grassy stunt virus and rice stripe virus have very different PAR frequencies of 12.8% and 35.2% (respectively) despite both being viral species within the Tenuivirus genus[163]. While these Tenuivirus shared some key determinants of PAR, such as the length of the cap snatched primer, it is hypothesized that the difference between these two species arise from other species specific determinants[162], [163]. I examined the PAR frequencies in each of the four IAV strains with published deep sequencing databases of mRNA 5' ends[1], [7]–[9] to determine if these strains were similar in PAR frequency. If these PAR frequencies were to be similar, it would be likely that the key determinants are shared among these IAV species without any key strain specific determinants. To determine the PAR frequency of each strain, I extracted the number of primers within the Realigned and No Realignment subpopulations. Primers were assigned to the Realigned subpopulation if the read primer sequence was matched to the reference genome after one or more rounds of iterative trimming. Primers were assigned to the No Realignment subpopulation if the read sequence was matched to the reference genome without iterative trimming in the data sets published by Sikora *et al*[1]. For example, the two primers 5'- AATGGGCGGGCA-3' and 5'-GTTTTCGGCAGCA-3', which are cap snatched by the Hong Kong strain, would be matched to the reference genome. 5'- AATGGGCGGGCA-3' is matched to the SLC25A13 gene without iterative trimming, and is therefore assigned to the No Realignment subpopulation. No match is found for 5'-GTTTTCGGCAGCA-3', and therefore a 5'-GCA-3' is trimmed leaving 5'-GTTTTCGGCA-3'. 5'-GTTTTCGGCA-3'is matched to the

HDAC4 gene; since this occurred following one round of iterative trimming, this primer is assigned to the Realigned subpopulation. To determine the PAR frequency the number of primers assigned to the Realigned subpopulations was divided by the total number of matched primers. These data are displayed in table 3.1.1.

The majority of cap-snatched primers (85.7±0.6%) do not undergo PAR, indicating that PAR is generally not necessary prior to transcription. The relative frequency of PAR is very similar between all four strains at approximately 14.3±0.6% of all cap snatched mRNA. This indicates that there are likely no key strain specific determinants, and that the shared PAR determinants are likely similar among all four strains. This PAR frequency was higher than the previous reported value of ~10%[138], and may be due to the variability in the length and sequence of cap snatched primers, as well as the IAV gene segment transcribed in these deep sequencing data sets versus the previous *in vitro* study.

Table 3.1.1. Relative Distribution of Cap-Snatched Primers between the Realigned and No Realignment Subpopulations

	Puerto Rico	Hong Kong	WSN	Brisbane
No Realignment	85.8	85.0	85.8	86.4
Realigned	14.2	15.0	14.2	13.6

The percentage of primers in both the No Realignment (No Realign) and Realigned (Realign) subpopulations for each of the four strains. Overall average frequency of PAR is 14.3±0.6%.

3.2 Prime and Realign Rounds

When PAR occurs, it is typically limited to a single nucleotide sequence addition (one round)[1], [4], [5], [7], [163], [167]. Nucleotide sequence additions consistent with multiple rounds of PAR have been observed in deep sequencing data sets[1], [7]–[9], [163], [167]. PAR occurs 14.3±0.6%; however, it is unknown if this rate remains constant between rounds. If this rate is consistent with 14.3±0.6% of primers continuously undergoing PAR, this would indicate that PAR may occur randomly or that the determinants of PAR are not changed by PAR; however, if less than 14.3±0.6% of primers require a subsequent round of PAR, this could indicate that PAR rescues primers. To examine this process I extracted all primers within the Realigned subpopulation. The number of primers remaining after each round of PAR was extracted and displayed as a percentage of the total primers. For example the primer 5'-GGTAGCGCGGCAGCAGCAA-3' was matched after three rounds of trimming (5'-GCAA, 5'-GCA, 5'-GCA) as 5'-GGTAGCGCG-3' to one TSS of the RBL1 gene. This primer would be counted in the groups of primers which underwent PAR in round one, two, and three; since the fourth attempt resulted in transcription of mRNA it would not be counted with the primers that underwent four rounds of PAR. These data are in table 3.2.1.

Of the 14.3±0.6% of primers which underwent at least one round of PAR, 95.6±2.5% (13.6±0.8% of all original primers) did not undergo a subsequent round of PAR. Similarly, of the primers that underwent at least two rounds of PAR (0.629±0.338% of all primers, or 4.44±2.47% of the Realigned subpopulation) 92.6±5.0 did not undergo a subsequent round of PAR. Very few primers underwent three or more rounds of PAR.

These data appear to be decreasing exponentially with each round at a similar rate. These data were $\log_e$ transformed and graphed in Figure 3.2.1. The percentage of primers that underwent a given number of rounds of PAR decreased in a linear manner on the $\log_e$ scale (R^2 =

0.986±0.015). This indicates that PAR was capable of consistently rescuing the same approximate percentage of primers each round. Using the linear models to determine the PAR rate after the first round for these strains, PAR is less frequent in subsequent rounds for the Puerto Rico (1.30%), Hong Kong (7.80%), and Brisbane (8.36%), and more frequent for the WSN strain (22.0%). It is unknown why WSN exhibits a higher PAR rate in subsequent rounds, though this may be related to the sequencing method and data set preparation, as these were different from the data for the other three strains. Omitting WSN PAR appears to be a highly efficient rescue process, reducing the frequency of a subsequent round of PAR. It would appear as though the addition of a G to GC[A/G]AAA nucleotide sequence was sufficient to address the factors that led to a higher PAR rate (14.3±0.6%) at round one. These additions modified the length by one to six nucleotides, the 3' end of the primer that was directed at 3'U1 of the vRNA template, and modified the G+C content of the primer.

Table 3.2.1. Relative Distribution of the Realigned Subpopulation by Rounds

	Puerto Rico	Hong Kong	WSN	Brisbane
Round 1	14.2	15.0	14.2	13.6
Round 2	0.170	0.578	0.863	0.904
Round 3	0.0024	0.0320	0.111	0.0900
Round 4		0.00797	0.0320	0.0075
Round 5			0.00762	
Round 6			0.000829	
Round 7			0.000542	
Round 8			0.000243	
Round 9			3.17E-05	

The percentage of primers in the Realigned subpopulation that underwent the given number of rounds of PAR as estimated by the number of rounds of trimming prior to matching the mRNA within ±100 nucleotides of a transcription start site.

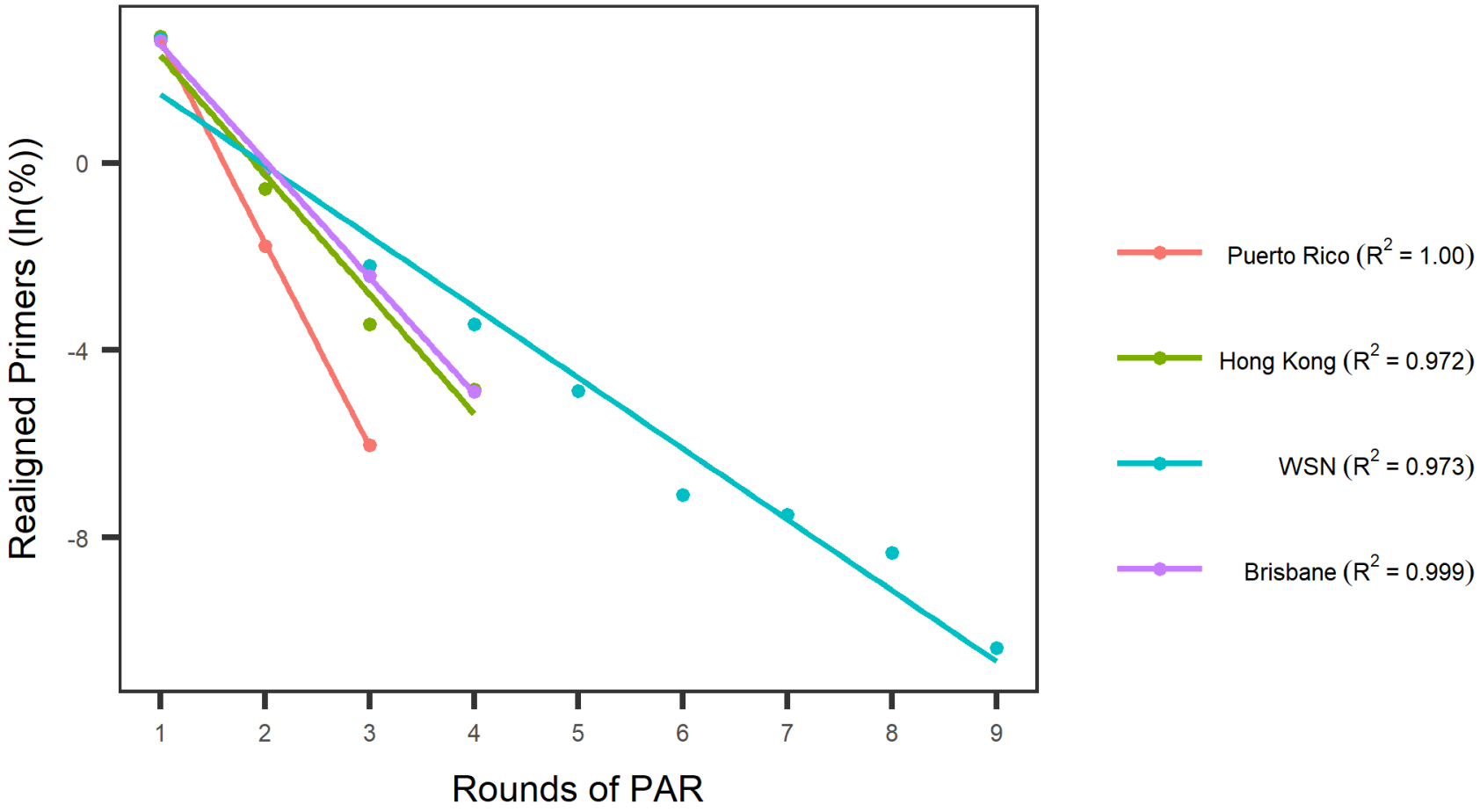
Realigned Primers (ln(%))
0
-4
-8
1
2
3
4
5
6
7
8
9
Rounds of PAR
Puerto Rico (R^2 = 1.00)
Hong Kong (R^2 = 0.972)
WSN (R^2 = 0.973)
Brisbane (R^2 = 0.999)

Figure 3.2.1 Loge Distribution of Realigned Primers at a Given Number of Rounds of PAR. Log_e transformed percentage of mRNA primers that undergo a given number of rounds of PAR or more.

3.3 Complementarity between the Primer and Template Sequences

Others have shown that the PA subunit of RdRp preferentially cleaves nascent pre-mRNA 3' of a guanine residue[6], [7]. This guanine residue has been demonstrated to be directed at 3' C2 of the vRNA template rather than 3' U1[138]. I examined if cleavage at a guanine residue was indeed the case in these data sets and if the presence of this 3' guanine residue was related to the frequency of PAR. The complimentary guanine residue, being directed at 3'C2 of the vRNA, is indistinguishable from the first guanine residue of either the conserved sequences or a nucleotide sequence added by PAR. For example, the primers 5'- GACGGAAGAACA which would prime transcription with a $\begin{matrix} CA \\ UCG \end{matrix}$ duplex and 5'- GACGGAAGAACA**G** which could prime transcription with a $\begin{matrix} \boldsymbol{CAG} \\ UCG \end{matrix}$ duplex would be indistinguishable after a guanine residue directed at 3'C2 has been transcribed on the first primer to give a $\begin{matrix} CAG \\ UCG \end{matrix}$ duplex. Since the matched reads were obtained following removal of the 5'-GC[A/G]AAAGCAGG motif corresponding to the conserved sequence of any nucleotide sequence addition from PAR (G – GC[A/G]AAA) the G would be removed from the second primer to yield 5'- GACGGAAGAACA for both sequences. Simply a guanine residue directed internally is removed by the matching procedure, as it is not directly determinable if this G came from the primer or transcription of a guanine residue directed at 3'C2. The presence of this guanine residue must therefore be extracted from the reference genome as the nucleotide 3' of the end of the matched sequence. For example, the sequence 5'- GACGGAAGAACA was matched to one TSS of the FUCA2 gene. In this gene the nucleotide immediately 3' of this sequence in the reference genome is a guanine residue (5'-…AGA**GACGGAAGAACAG**CG…). To this end, the last three nucleotides of all matched sequences as well as the three nucleotides immediately after this matched sequence in the reference genome, were extracted. In the case of the example primer, the 5'-**ACAG**CG portion

would be extracted; the last three nucleotides of the matched sequence (ACA) and the three nucleotides 3' of this sequence in the reference genome (GCG). This was performed for all cap snatched primers, and those in both the No Realignment and Realigned subpopulations, and the frequency of each nucleotide in these positions is displayed in figure 3.3.1.

These data reveal that a guanine residue was present one nucleotide 3' of the matched sequence for 47.4±4.6% of all cap snatched mRNA. Comparing the enrichment of guanine in both the No Realignment (49.0±4.8%) and Realigned (38.0±4.7) subpopulations, this enrichment is on average 1.30±0.15 fold higher in the No Realignment subpopulation across all four strains. Given the high guanine content of mammalian 5' untranslated regions (~59% in the mini genome of the Puerto Rico, Hong Kong, and WSN data sets; ~53% in the Brisbane data set), it is possible that the guanine residues one nucleotide 3' of the end of the matched sequence is an artifact of this guanine enrichment.

To determine if the observed nucleotide frequencies did indeed relate to a selective process rather than the abundance of the nucleotide residue within the genome, a Monte Carlo procedure was performed. A Monte Carlo procedure allows for the probability of an event, in this case a nucleotide being selected, to be determined from a known sample set. Simply, this procedure simulates random cleavage based on the nucleotide enrichment within the genome. The distribution of these randomized values can then be used to determine if the actual value is explained by the enrichment of a nucleotide within the genome, or if the actual value is due to a selective process. The Monte Carlo procedure was performed by randomly selecting nucleotide residues within the reference genome (100 nucleotides of a transcription start site from which a primer was cap snatched primer) and calculating the frequency of selection of A, C, G, or U/T. This process was repeated 100,000 times to ensure that a robust distribution of randomly selected

nucleotide frequencies could be obtained. The mean and standard deviation for these random distributions were calculated. Using this sample distribution, p-values were calculated as the number of iterations wherein the actual nucleotide enrichment was greater than (if the actual value were higher than the simulated mean) or less than (if the actual value were lower than the simulated mean) the actual value. For example, a cytidine residue is found in the Realigned subpopulation at position +1 (immediately 3' of the end of the matched sequence) in 27.8% of primers. The mean of the 100,000 simulated values was 27.7%. Since the actual value (27.8%) is higher than the simulated mean (27.7%), all iterations where a value higher than the actual value will be used to determine the p-value. A value higher than 27.8% occurred in 46,679 of the 100,000 iterations, therefore the p-value is 0.467 (rounded). The actual enrichment of cytidine at this position was well explained by the abundance of this nucleotide within the genome as 46,679 of the iterations randomly generated a value equal to or greater than the observed enrichment. It can therefore be concluded that this frequency was due to the abundance of cytidine within the genome rather than a selected process. Had this enrichment been from a selective process, very few or none of the 100,000 iterations would have generated a cytidine enrichment similar too, or greater, than what was observed. Since cleavage 3' of a guanine residue is the minimum preference of selected cleavage[6], we first examined the enrichment of a guanine one nucleotide 3' of the end of the matched sequence. The values for the enrichment of a guanine residue at position +1 are presented in table 3.3.1.

The presence of a guanine residue at position +1 (one nucleotide 3' of the end of the matched sequence) is indeed a non-random enrichment ($p < 1/100,000$ for all four strains). It is therefore unlikely that this enrichment was due to the high guanine content of mammalian 5' UTRs (*e.g.*, the reference genome). Instead, it was likely that this enrichment was due to

selective cleavage 3' of a guanine residue as previously reported for the PA endonuclease and RdRp[6]. Overall, RdRp appears to select for transcripts and/or cleavage sites where the transcript can be cleaved 3' of a guanine residue. This selection was highest in the No Realignment subpopulation. A breakdown in this selection was seen in Realigned subpopulation for all four strains, where the enrichment of guanine was lower than in the No Realignment subpopulation and overall population. This difference is particularly apparent in the Puerto Rico data set wherein the Realigned subpopulation has a guanine enrichment at +1 though could be explained by the abundance of guanine in the 5' UTR ($p = 0.317$). It is possible that a breakdown in selection of a complimentary guanine that would allow for priming internally at 3'C2 is related to PAR.

An *in vitro* analysis determined that the capacity to prime internally at 3'C2 with a complimentary guanine was unrelated to PAR[138]. It is currently unknown if priming internally at 3'C2 is associated with a reduction in PAR *in vivo*. I extracted the distribution of primers in the No Realignment and Realignment subpopulation based on the presence or absence of the complimentary guanine. These PAR frequencies are graphed in figure 3.3.1. Fold changes were calculated as the PAR rate for the given variable divided by the strain specific PAR frequency; these values were then averaged.

The presence of the complimentary guanine was associated with a small 0.802±0.078 fold decrease in PAR (11.4±1.3%). Primers that cannot initiate internally undergo PAR 1.18±0.08 fold more frequently (16.8±1.0%). Given the small effect size, it is unlikely that the complimentary guanine is directly associated with PAR. This guanine does, however, modify both the length of the cap snatched primer and the nucleotide sequence used to prime

transcription. It is possible that the complimentary guanine acts in conjunction with one or both of these factors to modify the PAR rate.

RdRp has also been shown to be selective with nucleotides at both the -1 and -2 positions[4], [5], [7]. I examined the abundance of each nucleotide at these positions, and compared these to the simulated values from the Monte Carlo procedure used to generate table 3.3.1. For position -2, only cytidine was examined as it was selected for in all four strains; these data are in table 3.3.2. Nucleotide residues at -1 that had a non-random abundance in three out of four data sets (adenine and uridine) were examined further; the data for adenine at -1 is in table 3.3.3, and the data for uridine at -1 is in table 3.3.4.

The enrichment of cytidine at -2 does appear to be due to a selective process (Table 3.3.2) rather than based on the abundance of this residue within the genome. This abundance is similar in both the Realigned (45.8±12.2%) and the No Realignment subpopulations (48.3±9.1%). It is unlikely that the selection of this cytidine residue is related to PAR despite a slight reduction in PAR rates to 13.5±1.4% when this cytidine was selected for at -2. Since this nucleotide is not used to prime transcription, it is possible that a cytidine at this position may be related to the function of the PA endonuclease.

The nucleotide residue directed at 3'U1 (position -1) has been shown to be related to PAR, with a uridine residue at this position increasing the PAR rate, and an adenine residue at this position decreasing the PAR frequency[138]. There was a selection preference for adenine, and a selection against uridine (tables 3.3.3 and 3.3.4) both overall (33.9±6.1% adenine, 12.3±4.0% uridine) and in the No Realignment subpopulation (36.6±5.8% adenine, 10.3±3.8% uridine). This selection is reversed in the Realigned subpopulation. The enrichment of adenine in the Realigned subpopulation was near or below random in three out of four strains (17.6±7.5%

versus 21.0±1.3% in the genome). The enrichment of uridine was near or above random in the Realigned subpopulation for three out of four strains (24.2±5.8% versus 21.3±1.6% in the genome). In agreement with previous results[138], I found that an uridine residue directed at 3'U1 was associated with an increase in PAR (28.8±3.4%), and that an adenine is associated with a decrease in PAR (7.13±1.79%). These data indicate that PAR may occur due to a breakdown in selective cleavage of a compatible nucleotide that would be the nucleotide directed at 3'U1.

Based on the frequencies of these nucleotides, it would appear as though the cleavage preference for RdRp is to generate a primer with an adenine at position -1 (directed at 3'U1) and guanine at position +1 (directed at 3'C2). Others have shown that IAV has a preference for primers ending in AG or CAG[4], [5], [7]. Given the increase in PAR when the cleavage preference for the nucleotide at -1 (3'U1 directed nucleotide), I next examined if the preference for primers ending in AG or CAG was both enriched in the cap snatched sequences, and if this preference was related to PAR. A 100,000 iteration Monte Carlo procedure was performed, randomly selecting trinucleotide sequences and calculating the frequency of ".AG" and "CAG". The frequency of ".AG" and "CAG" in cap snatched primers both overall and in the Realigned and No Realignment subpopulations were also extracted. These data are in tables 3.3.5 and 3.3.6. Primers ending in an AG dinucleotide occurred 16.4±2.3% in the No Realignment subpopulation. There was a loss of this preference in the Realigned subpopulation wherein 6.07±3.44% of primers ended in an AG dinucleotide, which is less than the abundance of this sequence being chosen randomly (7.42±0.06). Overall, when this preference was met, the PAR frequency was reduced to 5.47±2.44%. Similar results were obtained when examining primers ending in a CAG trinucleotide. Enrichment of the CAG trinucleotide was highest in the No Realignment subpopulation (10.7±1.2%), and near random levels in the Realigned subpopulation

(2.66±1.17% versus the 2.41±0.15% abundance randomly). When this nucleotide sequence preference was met, the PAR frequency was reduced to 4.00±1.86%.

These data indicate that PAR may occur due to a breakdown in sequence selective cleavage. This breakdown may be due to the relative abundance of these nucleotides residues within the genome. To prevent this, RdRp preferentially cleaves at a compatible nucleotide sequence with the compatible length range[4]–[6], at a level above that which can be explained by the nucleotide enrichment of the genome and random selection. When the constraints of abundance lead to a breakdown in this preference, transcripts of low nucleotide sequence compatibility may be cap snatched. These low compatibility sequences were more frequent in the Realigned subpopulation, and appear to be more likely to require PAR prior to transcription.

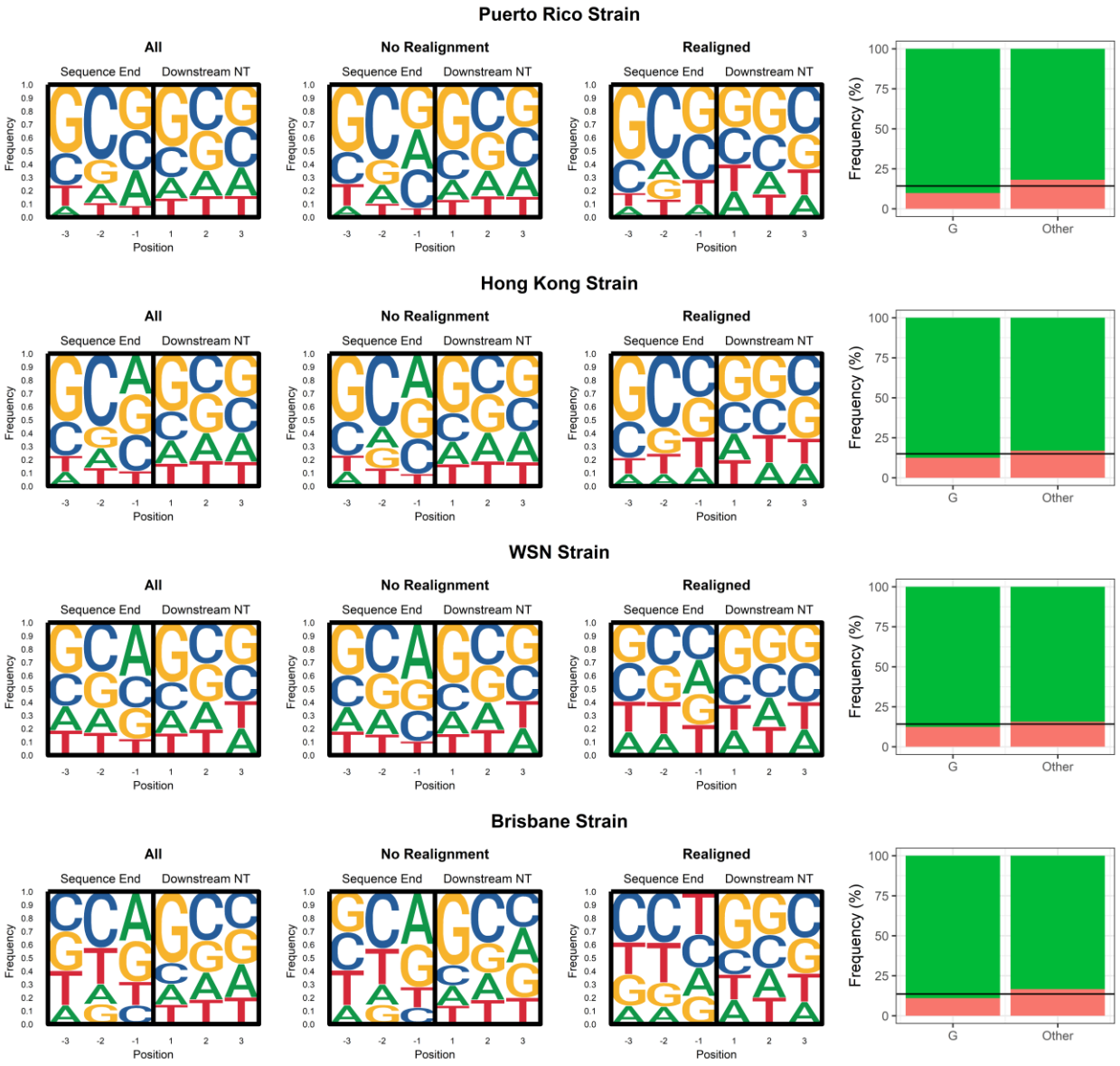
Puerto Rico Strain
All
No Realignment
Realigned
Sequence End
Downstream NT
Frequency
Position
Frequency (%)
G
Other
Hong Kong Strain
WSN Strain
Brisbane Strain

Figure 3.3.1. Nucleotide Enrichment 3' and 5' of the End of the Matched Sequence and Frequency Distributions of PAR relative to presence of a Complimentary Guanine Residue. Columns one to three: the nucleotide enrichment of all sequences and those in both the No Realignment and Realigned subpopulations within three nucleotide residues of the apparent cleavage site (the end of the matched sequence). Column four: the relative distributions of mRNA primers between the Realigned (Red) and No Realignment (Green) subpopulations with or without the complimentary guanine residue; the horizontal black line indicates the PAR frequency for that given strain. Red bars above the black line indicate a bias towards PAR, whereas green bars below the black line indicate a bias against PAR.

Table 3.3.1. Guanine at +1 Enrichment and Randomly Generated Enrichment

	Puerto Rico			Hong Kong			WSN			Brisbane		
	All	No	PAR	All	No	PAR	All	No	PAR	All	No	PAR
G%	47.0	49.5	32.5	43.7	45.0	36.5	45.1	46.0	39.3	53.9	55.6	43.7
Sim%	31.9	31.9	31.9	31.2	31.2	31.2	30.8	30.8	30.8	28.5	28.5	28.5
Sim sd	1.40	1.40	1.40	0.877	0.877	0.877	0.261	0.261	0.261	0.961	0.961	0.961
p-val	0*	0*	0.317	0*	0*	0*	0*	0*	0*	0*	0*	0*
PAR%	9.84	9.84	9.84	12.5	12.5	12.5	12.4	12.4	12.4	11.0	11.0	11.0

Guanine enrichment at position +1 (nucleotide that can be directed at 3'C2 to prime transcription) for All primers (All) and the primers in the No Realignment (No) and Realigned (PAR) subpopulations of all four strains and the simulated values. G% indicates the enrichment (in percent) of the guanine residue 3' of the end of the matched sequence; Sim% indicates the average enrichment of a guanine residue being selected by the 100,000 iteration Monte Carlo procedure, Sim sd is the standard deviation of this percentage. p-value is the number of times a guanine enrichment greater than the observed value was generated randomly, divided by 100,000; *a p-value of 0 indicates that 0/100,000 iterations were above or equal to the observed enrichment. PAR% indicates the PAR frequency when primers have a guanine residue at +1.

Table 3.3.2. Cytidine at -2 Enrichment and Randomly Generated Enrichment

	Puerto Rico			Hong Kong			WSN			Brisbane		
	All	No	PAR	All	No	PAR	All	No	PAR	All	No	PAR
C%	56.8	56.9	56.3	55.2	55.1	55.8	37.3	38.1	32.3	42.6	43.2	38.7
Sim%	27.8	27.8	27.8	27.9	27.9	27.9	27.9	27.9	27.9	24.9	24.9	24.9
Sim sd	1.32	1.32	1.32	0.85	0.85	0.85	0.24	0.24	0.24	0.86	0.86	0.86
p-val	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*
PAR%	14.1	14.1	14.1	15.1	15.1	15.1	12.3	12.3	12.3	12.4	12.4	12.4

Cytidine enrichment at position -2 for All primers (All) and the primers in the No Realignment (No) and Realigned (PAR) subpopulations of all four strains and the simulated values. C% indicates the enrichment (in percent) of the cytidine residue two nucleotides 5’ of the end of the matched sequence; Sim% indicates the average enrichment of a cytidine residue being selected by the 100,000 iteration Monte Carlo procedure, Sim sd is the standard deviation of this percentage. p-value is the number of times a cytidine enrichment greater than the observed value was generated randomly, divided by 100,000; *a p-value of 0 indicates that 0/100,000 iterations were above or equal to the observed enrichment. PAR% indicates the PAR frequency when primers have a cytidine residue at -2.

Table 3.3.3. Adenine at -1 Enrichment and Randomly Generated Enrichment

	Puerto Rico			Hong Kong			WSN			Brisbane		
	All	No	PAR	All	No	PAR	All	No	PAR	All	No	PAR
A%	27.1	30.1	9.2	30.7	33.6	13.9	40.5	42.9	25.8	37.3	39.8	21.5
Sim%	20.2	20.2	20.2	20.4	20.4	20.4	20.6	20.6	20.6	22.9	22.9	22.9
Sim sd	1.25	1.25	1.25	0.792	0.792	0.792	0.244	0.244	0.244	0.919	0.919	0.919
p-val	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0.0536
PAR%	4.84	4.84	4.84	6.78	6.78	6.78	9.06	9.06	9.06	7.84	7.84	7.84

Adenine enrichment at position -1 (the nucleotide directed at 3'U1 to prime transcription) for All primers (All) and the primers in the No Realignment (No) and Realigned (PAR) subpopulations of all four strains and the simulated values. A% indicates the enrichment (in percent) of the adenine residue at the end of the matched sequence; Sim% indicates the average enrichment of a adenine residue being selected by the 100,000 iteration Monte Carlo procedure, Sim sd is the standard deviation of this percentage. p-value is the number of times a adenine enrichment greater than the observed value was generated randomly, or less than the observed value if the observed value was less than the simulated mean, divided by 100,000; *a p-value of 0 indicates that 0/100,000 iterations were above/below or equal to the observed enrichment. PAR% indicates the PAR frequency when primers have an adenine residue at -1.

Table 3.3.4. Uridine at -1 Enrichment and Randomly Generated Enrichment

	Puerto Rico			Hong Kong			WSN			Brisbane		
	All	No	PAR	All	No	PAR	All	No	PAR	All	No	PAR
U%	8.24	6.48	18.8	11.1	9.03	22.7	12.0	10.2	22.6	17.7	15.4	32.5
Sim%	20.2	20.2	20.2	20.5	20.5	20.5	20.8	20.8	20.8	23.6	23.6	23.6
Sim sd	1.23	1.23	1.23	0.790	0.790	0.790	0.249	0.249	0.249	0.938	0.938	0.938
p-val	0*	0*	0.133	0*	0*	0.00374	0*	0*	0*	0*	0*	0*
PAR%	32.5	32.5	32.5	30.7	30.7	30.7	26.8	26.8	26.8	25.0	25.0	25.0

Uridine enrichment at position -1 (the nucleotide directed at 3'U1 to prime transcription) for All primers (All) and the primers in the No Realignment (No) and Realigned (PAR) subpopulations of all four strains and the simulated values. U% indicates the enrichment (in percent) of the uridine residue at the end of the matched sequence; Sim% indicates the average enrichment of a uridine residue being selected by the 100,000 iteration Monte Carlo procedure, Sim sd is the standard deviation of this percentage. p-value is the number of times a uridine enrichment greater than the observed value was generated randomly, or less than the observed value if the observed value was less than the simulated mean, divided by 100,000; *a p-value of 0 indicates that 0/100,000 iterations were above/below or equal to the observed enrichment. PAR% indicates the PAR frequency when primers have an uridine residue at -1.

Table 3.3.5. Enrichment of an AG Dinucleotide at the 5' End of Primers and Randomly Generated Enrichment

	Puerto Rico			Hong Kong			WSN			Brisbane		
	All	No	PAR	All	No	PAR	All	No	PAR	All	No	PAR
AG%	12.3	14.1	1.86	13.8	15.4	4.75	18.0	19.4	9.44	15.7	16.9	8.23
Sim%	7.41	7.41	7.41	7.41	7.41	7.41	7.49	7.49	7.49	7.35	7.35	7.35
Sim sd	0.815	0.815	0.815	0.510	0.510	0.510	0.152	0.152	0.152	0.548	0.548	0.548
p-val	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0.0641
PAR%	2.15	2.15	2.15	5.14	5.14	5.14	7.45	7.45	7.45	7.15	7.15	7.15

AG dinucleotide enrichment at the end of All primers (All) and the primers in the No Realignment (No) and Realigned (PAR) subpopulations of all four strains and the simulated values. AG% indicates the enrichment (in percent) of the AG dinucleotide at the end of the matched sequence; Sim% indicates the average enrichment of an AG dinucleotide being selected by the 100,000 iteration Monte Carlo procedure, Sim sd is the standard deviation of this percentage. p-value is the number of times a AG dinucleotide enrichment greater than the observed value was generated randomly, or less than the observed value if the observed value was less than the simulated mean, divided by 100,000; *a p-value of 0 indicates that 0/100,000 iterations were above/below or equal to the observed enrichment. PAR% indicates the PAR frequency when primers end in a AG dinucleotide.

Table 3.3.6. Enrichment of a CAG Trinucleotide at the 5' End of Primers and Randomly Generated Enrichment

	Puerto Rico			Hong Kong			WSN			Brisbane		
	All	No	PAR	All	No	PAR	All	No	PAR	All	No	PAR
CAG%	9.49	10.9	0.935	10.5	11.9	2.93	9.95	11.1	3.25	8.28	9.04	3.52
Sim%	2.46	2.46	2.46	2.53	2.53	2.53	2.47	2.47	2.47	2.20	2.20	2.20
Sim sd	0.471	0.471	0.471	0.310	0.310	0.310	0.079	0.079	0.079	0.267	0.267	0.267
p-val	0*	0*	0*	0*	0*	0.0990	0*	0*	0*	0*	0*	0.0002
PAR%	1.402	1.402	1.402	4.176	4.176	4.176	4.638	4.638	4.638	5.782	5.782	5.782

CAG trinucleotide enrichment at the end of All primers (All) and the primers in the No Realignment (No) and Realigned (PAR) subpopulations of all four strains and the simulated values. AG% indicates the enrichment (in percent) of the CAG trinucleotide at the end of the matched sequence; Sim% indicates the average enrichment of an AG dinucleotide being selected by the 100,000 iteration Monte Carlo procedure, Sim sd is the standard deviation of this percentage. p-value is the number of times a CAG trinucleotide enrichment greater than the observed value was generated randomly, or less than the observed value if the observed value was less than the simulated mean, divided by 100,000; *a p-value of 0 indicates that 0/100,000 iterations were above/below or equal to the observed enrichment. PAR% indicates the PAR frequency when primers end in a CAG trinucleotide.

3.4 Length of Cap Snatched Primers

The length of cap-snatched primers has been shown to be of importance to the prime-and-realign process in other negative sense single stranded RNA viruses which engage in cap-snatching, with shorter primers being selectively enriched in the Realigned subpopulation[161]–[163]. To examine if PAR was also related to the length of the cap snatched primers for IAV, the length of all matched primers was extracted. For IAV priming may occur at 3'C2 rather than 3'U1[4]–[7], [138]; furthermore, since there was a nonrandom enrichment of a guanine residue one nucleotide 3' of the end of the match sequence, which could have been directed at 3'C2 to prime transcription. For this reason, I included this guanine residue (when present in the reference genome) in the length of primers, effectively increasing the length of matched primers with this G by one nucleotide. I then examined the length of primers in both the No Realignment and Realigned subpopulations. Since PAR adds a nucleotide sequence to the cap-snatched primer, increasing the length, I classified the Realigned subpopulation into two groups: Pre-Realignment – the matched cap snatched primer, and Post-Realignment – the cap snatched primer prior to the iterative trimming process. For example, the sequence 5'-GTTTTCGGCAGCA which was matched as 5'-GTTTTCGGCA would be classified as Realigned, with the trimmed 5'-GTTTTCGGCA being further classified as Pre-Realignment, and the read sequence 5'-GTTTTCGGCAGCA further classified as Post-Realignment. The abundance of cap snatched mRNA in each of these groups was extracted relative to the total number of cap snatched primers. These data are presented in Figure 3.4.1 and Table 3.4.1. Note: the 5' cap structure is not included in these lengths.

Length values for the WSN are not indicative of the actual length of the cap snatched primer. The RNA template switching oligo method used to sequence the RNA and obtain these data added an n-length poly G head group to the extracted RNA[9]. In order to extract the

nucleotide sequence of the cap-snatched primers, all residues 5' of at least three G residues were removed. This removal results in a length decrease for some primers, as all nucleotides between the 5' cap and the first G stretches were be trimmed. For example, if a primer m^7G-GGCAUCGACCA were to be cap snatched, it would be read as 5'-CAUCGACCA. This results in artificially shorter primer lengths. For this reason the WSN values were omitted from further discussion.

The differences in length among the Pre-Realignment and No Realignment subpopulations revealed that cap snatched primers which undergo PAR are indeed of a different length than those that do not undergo PAR. The Pre-Realignment subpopulation were clustered around a length of 10.7±1.2 nucleotides, PAR lengthens these primers by roughly three nucleotides into a subpopulation of primers 14.0±1.7 nucleotides. The No Realignment subpopulation was clustered around 11.9±1.2 nucleotides. There was some overlap between the Pre-Realignment and No Realignment subpopulations particularly at ten and 11 nucleotides of length, though the mean length of primers in the Realigned Subpopulation were approximately 1.2 nucleotides shorter in length.

To further elucidate this process, the percentage of primers in both the Realigned and No Realignment subpopulation for each length between nine and 17 nucleotides was analyzed and compared to the strain specific PAR frequency. These data are displayed in figure 3.4.2. PAR occurred more frequently than the average for that strain with primers ten nucleotides in length and shorter; whereas primers 12 nucleotides or longer underwent PAR less frequently than the average for that strain. 11 nucleotides was the transition point where the frequency of PAR became similar to the strain specific frequency of PAR. This is consistent with the clustering of primers in the Realigned and No Realignment subpopulation on the basis of length. The PAR

frequency appears to be decreasing exponentially as the length of the cap snatched primers increases. These data were $\log_e$ transformed and graphed in Figure 3.4.3. These trends are poorly linear on the $\log_e$ scale between nine and 17 nucleotides of length, with an average R^2 of 0.889±0.075. These trends tend to deviate from linearity after a length of 13 nucleotides has been achieved; these data were re-analyzed between nine and 13 nucleotides in length. These data are displayed in Figure 3.4.4. In this range $\log_e$ transformed PAR rate and length are well correlated with an average R^2 of 0.973±0.035. It is possible that primers 14 nucleotides or longer do not undergo PAR on the basis of length. This is consistent with the results of others where the length of cap snatched primers has been shown to be a major determinant of PAR, but not the sole determinant of this process[138], [161]–[163]. Despite other determinants likely playing a role in this process, length does appear to be a major driving force. Primers less than 12 nucleotides of length being more likely require at least one round of PAR to increase length to within the optimal range.

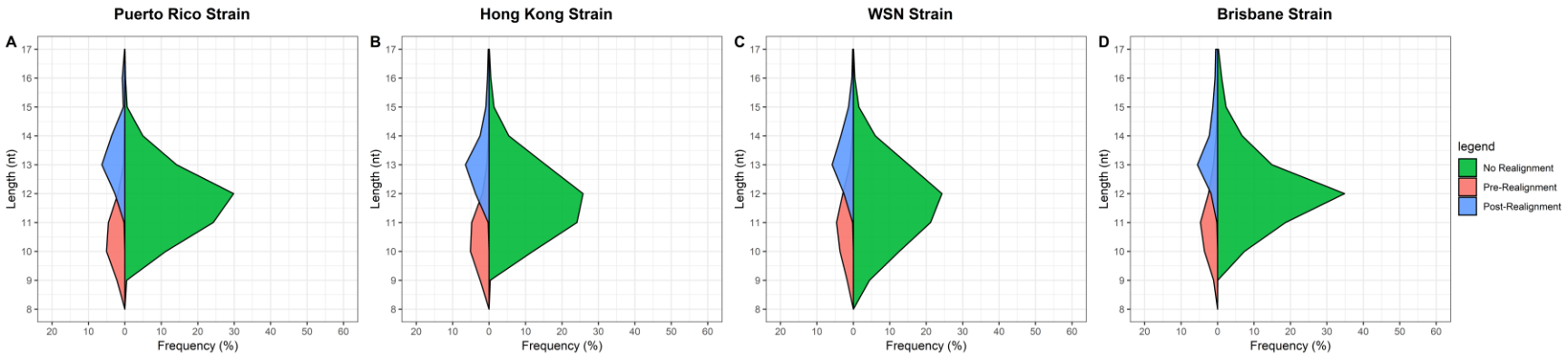
A
Puerto Rico Strain
B
Hong Kong Strain
C
WSN Strain
D
Brisbane Strain
Length (nt)
Frequency (%)
legend
No Realignment
Pre-Realignment
Post-Realignment

Figure 3.4.1. Length of Cap Snatched Primers. The relative abundance of primers based on the length of the cap snatched primer for each subpopulation (Pre-Realignment (red), No Realignment (green)), or following PAR (Post-Realignment (blue)).

Table 3.4.1. Length Distributions and p-values

	Series	Mean Length	Standard Deviation	Lower CI	Upper CI	p-value vs. No Realign	p-value vs. Post-Realign	p-value vs. Pre-Realign
Hong Kong	No Realign	11.8	1.24	9.33	14.3	1	0	0
	Post-Realign	13.7	1.55	10.6	16.8	0	1	0
	Pre-Realign	10.6	1.11	8.35	12.8	0	0	1
Puerto Rico	No Realign	11.7	1.14	9.48	14.0	1	0	0
	Post-Realign	13.6	1.27	11.1	16.2	0	1	0
	Pre-Realign	10.6	1.06	8.44	12.7	0	0	1
WSN	No Realign	11.7	1.39	8.93	14.5	1	0	0
	Post-Realign	13.9	1.64	10.6	17.2	0	1	0
	Pre-Realign	10.9	1.23	8.47	13.4	0	0	1
Brisbane	No Realign	12.1	1.25	9.59	14.6	1	0	0
	Post-Realign	14.6	2.29	10.0	19.1	0	1	0
	Pre-Realign	11.1	1.29	8.51	13.7	0	0	1

The mean length (nucleotides), standard deviation, and confidence intervals for the No Realignment (No Realign), Pre-Realignment (Pre-Realign), and Post-Realignment (Post-Realign) subpopulations. P-values are given along with the other subpopulation they are compared to; p-values of 1 are returned for comparing the same group to the same group; p-values of zero are reported instead of $p < 10^{-308}$.

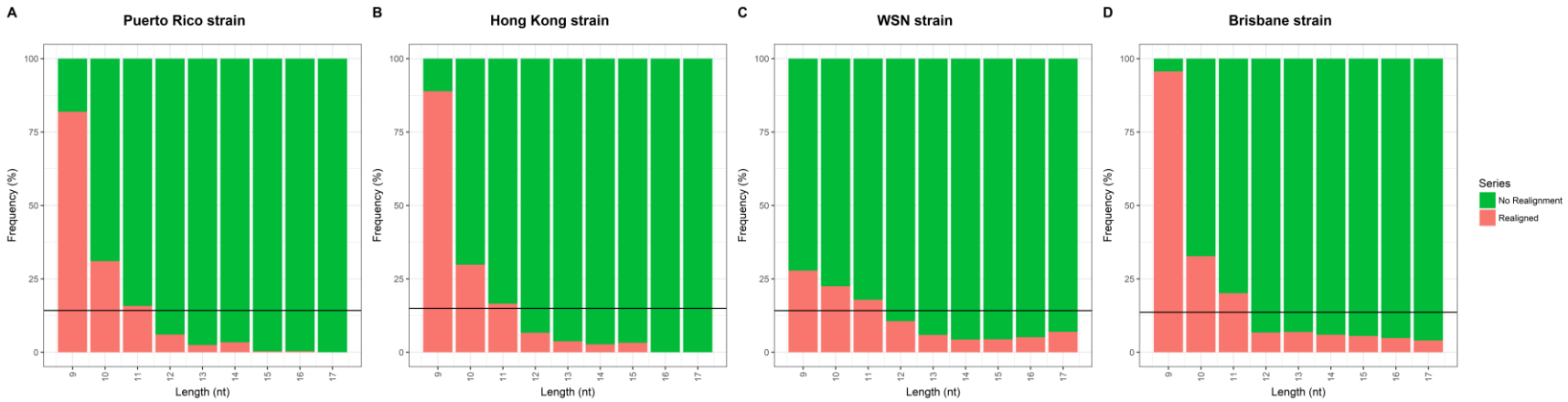
A
Puerto Rico strain
B
Hong Kong strain
C
WSN strain
D
Brisbane strain
Frequency (%)
Length (nt)
Series
No Realignment
Realigned

Figure 3.4.2. PAR Frequency at a given Length. The relative distributions of cap snatched primers between the Realigned (Red) and No Realignment (Green) subpopulations for lengths between nine and 17 nucleotides; the horizontal black line indicates the PAR frequency for that given strain. Red bars above the black line indicate a PAR bias.

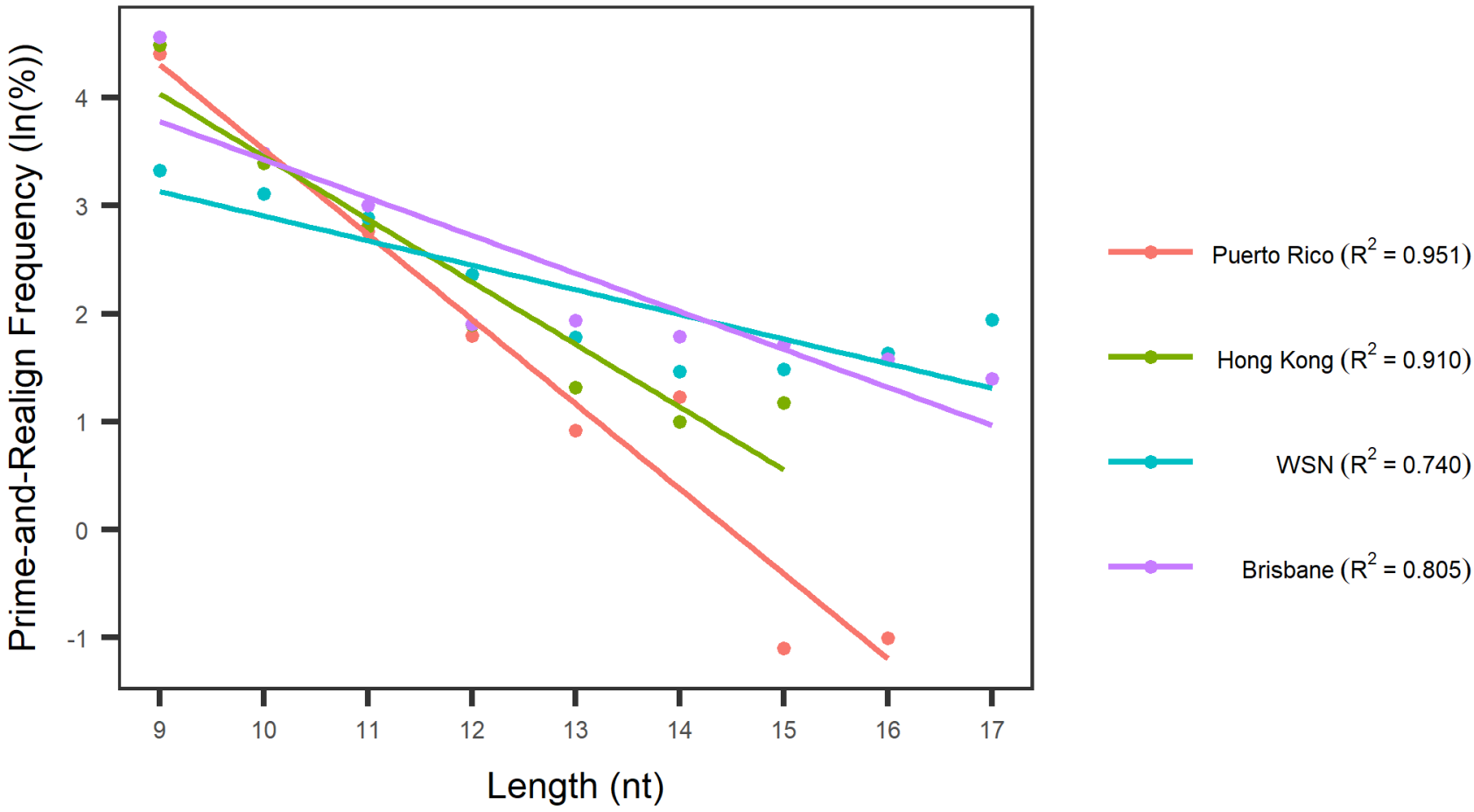
Prime-and-Realign Frequency (ln(%))
4
3
2
1
0
-1
9
10
11
12
13
14
15
16
17
Length (nt)
Puerto Rico (R^2 = 0.951)
Hong Kong (R^2 = 0.910)
WSN (R^2 = 0.740)
Brisbane (R^2 = 0.805)

Figure 3.4.3. Log Linear relation of PAR Frequency and Primer Length. The $\log_e$ transformed PAR frequency for lengths between nine and 17 nucleotides for each of the four IAV strains.

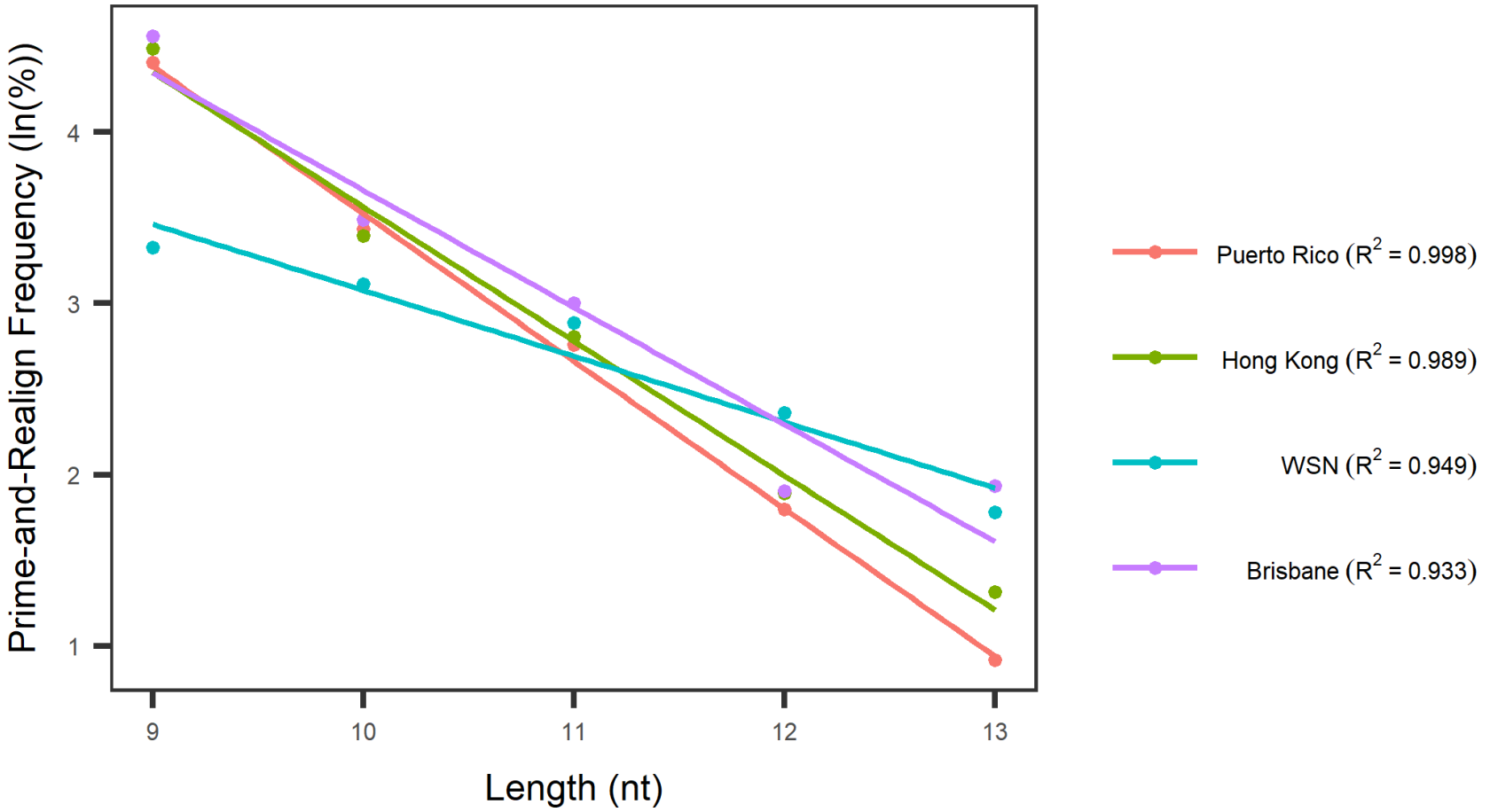
Prime-and-Realign Frequency (ln(%))
4
3
2
1
9
10
11
12
13
Length (nt)
Puerto Rico (R^2 = 0.998)
Hong Kong (R^2 = 0.989)
WSN (R^2 = 0.949)
Brisbane (R^2 = 0.933)

Figure 3.4.4. Log Linear relation of PAR Frequency and Primer Length between Nine and 13 Nucleotides. As in Figure 3.4.3 but only primers between nine and 13 nucleotides in length.

3.5 Length and Nucleotide Sequence Additions

To examine the impact of PAR on the cap snatched primer, I next examined the nucleotide sequence added in each PAR event. Given our interest was in the nucleotide sequences added by PAR, the No Realignment subpopulation was omitted from further analysis. For the Realigned subpopulation, the nucleotide sequence added at each PAR event was extracted. Some sequences underwent multiple rounds of PAR, and each of these PAR events was treated as an independent event. Due to length being a strong determinant of PAR, the nucleotide sequence additions were examined in relation to the length of the primer prior to the nucleotide sequence addition. For example, a primer originally nine nucleotides in length which underwent two rounds of realignment having three nucleotides and four nucleotides added for a final length of 16 nucleotides will be examined as a primer of nine nucleotides lengthened to 12 nucleotides, and again as a primer of 12 nucleotides lengthened to 16 nucleotides. These data were then analyzed relative to all PAR events. These data are displayed in figures 3.5.1.

Figure 3.5.1 shows that nucleotide sequences generally consist of a GCA addition (56.6±10.1%; 61.6±1.2% omitting WSN) across all four IAV strains for a net lengthening process of three nucleotides. This was consistent with the net change in Realigned primer length from 10.7±1.2 nucleotides to 14.0±1.7 following PAR. For all strains except WSN, GCAA was the second most common nucleotide sequence associated with PAR (12.6±0.3% omitting WSN). This process is similar among the Puerto Rico, Hong Kong, and Brisbane strains. For the WSN strain a GC nucleotide sequence addition was the second most common PAR addition (31.9% versus 3.33±1.64% for the other three strains). The WSN data set has an enrichment of GC PAR events when transcribing the M Segment. 4,489,171 of the 4,938,415 PAR events (90.9%) when transcribing the M subunit are a GC nucleotide sequence addition, with these 4,489,171 GC

additions being 86% of the 5,214,292 PAR events of GC for the entire WSN data set. Furthermore, PAR events from transcribing the M subunit (4,938,415) are 30.3% of the total PAR events in the WSN data set (16,323,819).

The segmented genome of IAV possesses two different conserved sequences, which vary at position 3' Y4. The 3' U4 containing conserved sequences (5/8ths of the IAV genome for Puerto Rico, Brisbane, and WSN; 6/8ths for Hong Kong) would be associated with PAR nucleotide sequence additions of GCA, and the 3' C4 containing conserved sequences would be associated with PAR nucleotide sequence additions of GCG. Despite at least 25% of the IAV genome having a cytidine residue at position 3' Y4, very few PAR events were associated with a GCG containing nucleotide sequence additions. GCG containing additions were relatively most common in WSN data set (9.65% GCG+, 51.3% GCA+), followed by the Puerto Rico (8.73 GCG+, 79.2% GCA+), Hong Kong (4.70% GCG+, 79.0% GCA+), and Brisbane (1.10% GCG+, 86.9% GCA+). These data indicate the PAR might be biased towards nucleotide sequence additions containing GCA (81.7±4.5%) over the equivalent GCG containing addition (4.84±3.81%), or that transcribing mRNA from a 3'C4 containing template results in less PAR, or both.

The number of PAR additions of a given nucleotide sequence length appears to decrease proportionally between additions of three (GC[A/G]) and six (GC[A/G]AAA) nucleotides. The frequency of nucleotide sequence additions of a given length, were $\log_e$ transformed and graphed in Figure 3.5.2. The frequency of PAR additions does not follow a clear trend between additions of one (G) to three (GC[A/G]) nucleotides. The frequency of additions does decrease in a linear $\log_e$ manner between three (GC[A/G]) and six (GC[A/G]AAA) nucleotides ($R^2 = 0.962\pm0.040$). PAR typically occurred following a three nucleotide sequence addition (64.8±3.9% GCA and

GCG), which is consistent with the dimension estimate of the RdRp active site[110]. This increase in the length of the addition may reflect a stretching of the active site to accommodate more base pairs prior to the factors that lead to PAR being applied. These factors would include separation of the 3' 5' base pair region, the separation of the template and product duplex, and their direction into the respective exit tunnels.

Given the high consistency of a GCA addition being the nucleotide sequence added by PAR, it is possible that this nucleotide sequence addition occurs irrespective of length of the cap snatched primer. I repeated the analysis for figure 3.5.1 and displayed these data relative to the number of PAR events at a given length, rather than the total number of PAR events. These data are in figure 3.5.3.

While there are some changes in the nucleotide sequence addition length at a given primer length, PAR appeared to remain primarily a three nucleotide addition (GCA or GCG) at all lengths. To determine if the sequence added was changing in response to an increase in length, I correlated the length of the primer prior to PAR with the lengths of the additions. These data revealed a poor correlation between primer length and nucleotide sequence addition length (Pearson = -0.0647±0.182 omitting WSN). Given the weak correlation and the consistent observed trend of three nucleotide long nucleotide sequence additions (64.8±3.9% GCA and GCG), these data suggest that the nucleotide sequence addition is fairly irrespective of the primer length prior to PAR.

Taken together, these data suggest that the nucleotide sequence associated with PAR (GCA) is fairly similar across all primer lengths. This process lengthened a primer by three nucleotides, generally by a GCA addition, allowing for transcription to be re-primed with a three nucleotide longer primer. Additions of GCA are more common than would be expected given the

distribution of the IAV genome between 3'U4 and 3'C4 templated segments, and may indicate that this process is either more common on 3'U4 templated mRNA or that PAR is typically a GCA containing addition regardless of vRNA template.

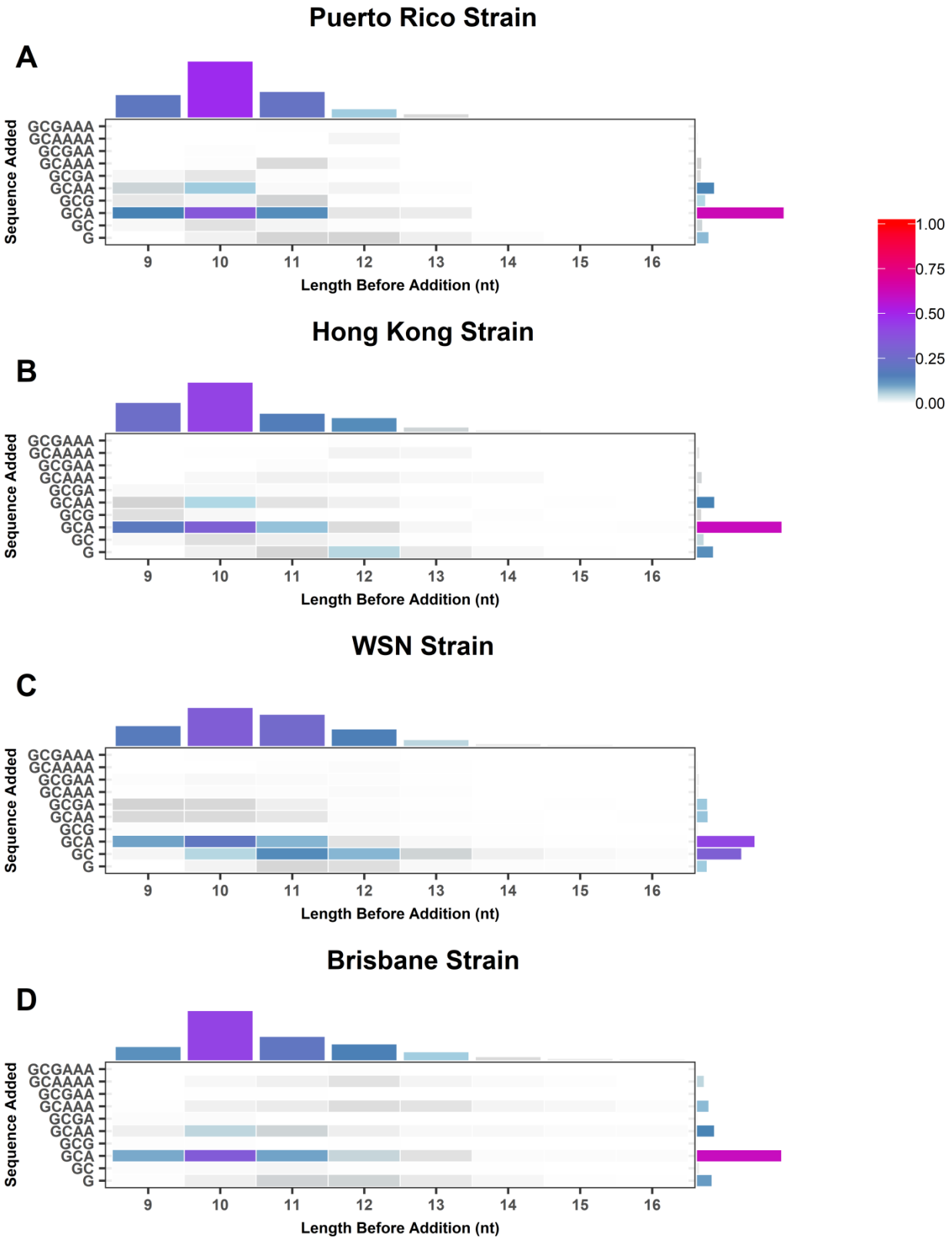
Puerto Rico Strain
A
Sequence Added
GCGAAA
GCAAAA
GCGAA
GCAAA
GCGA
GCAA
GCG
GCA
GC
G
9
10
11
12
13
14
15
16
Length Before Addition (nt)
Hong Kong Strain
B
Sequence Added
GCGAAA
GCAAAA
GCGAA
GCAAA
GCGA
GCAA
GCG
GCA
GC
G
9
10
11
12
13
14
15
16
Length Before Addition (nt)
WSN Strain
C
Sequence Added
GCGAAA
GCAAAA
GCGAA
GCAAA
GCGA
GCAA
GCG
GCA
GC
G
9
10
11
12
13
14
15
16
Length Before Addition (nt)
Brisbane Strain
D
Sequence Added
GCGAAA
GCAAAA
GCGAA
GCAAA
GCGA
GCAA
GCG
GCA
GC
G
9
10
11
12
13
14
15
16
Length Before Addition (nt)
1.00
0.75
0.50
0.25
0.00

Figure 3.5.1. Sequence Additions at all Start and End Lengths. The length of the primer prior to the addition and the frequency of a given addition. Vertical bars show the frequency of lengthening events from that length; horizontal bars show the frequency of lengthening events with the specific nucleotide sequence addition.

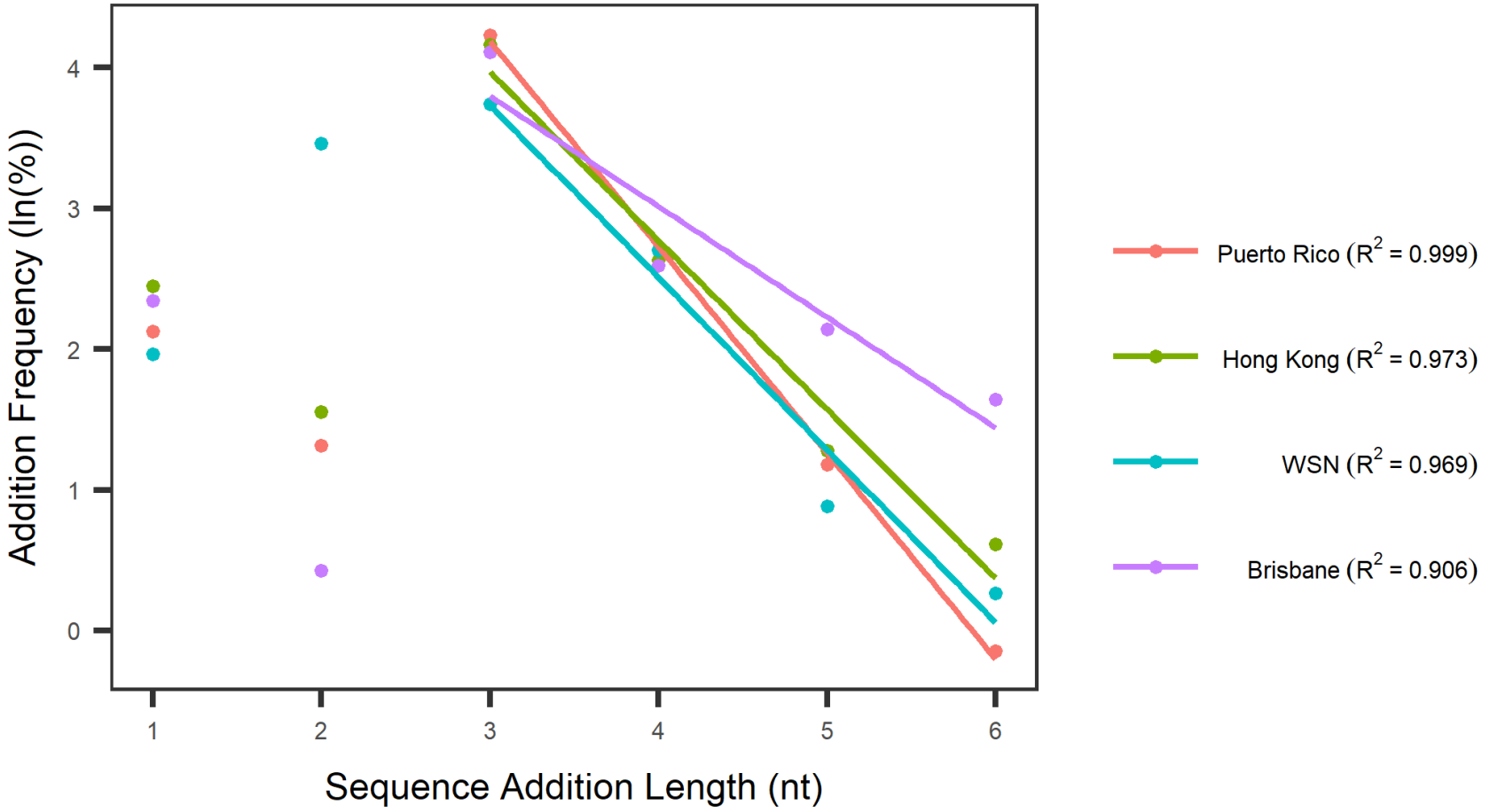
Addition Frequency (ln(%))
0
1
2
3
4
1
2
3
4
5
6
Sequence Addition Length (nt)
Puerto Rico (R^2 = 0.999)
Hong Kong (R^2 = 0.973)
WSN (R^2 = 0.969)
Brisbane (R^2 = 0.906)

Figure 3.5.2. Loge Transformed Frequency of Nucleotide Sequence Additions of a Given Length. The $\log_e$ frequency of a given addition length. R^2 values are obtained from the frequency of additions between three (GC[A/G]) and six (GC[A/G]AAA) additions.

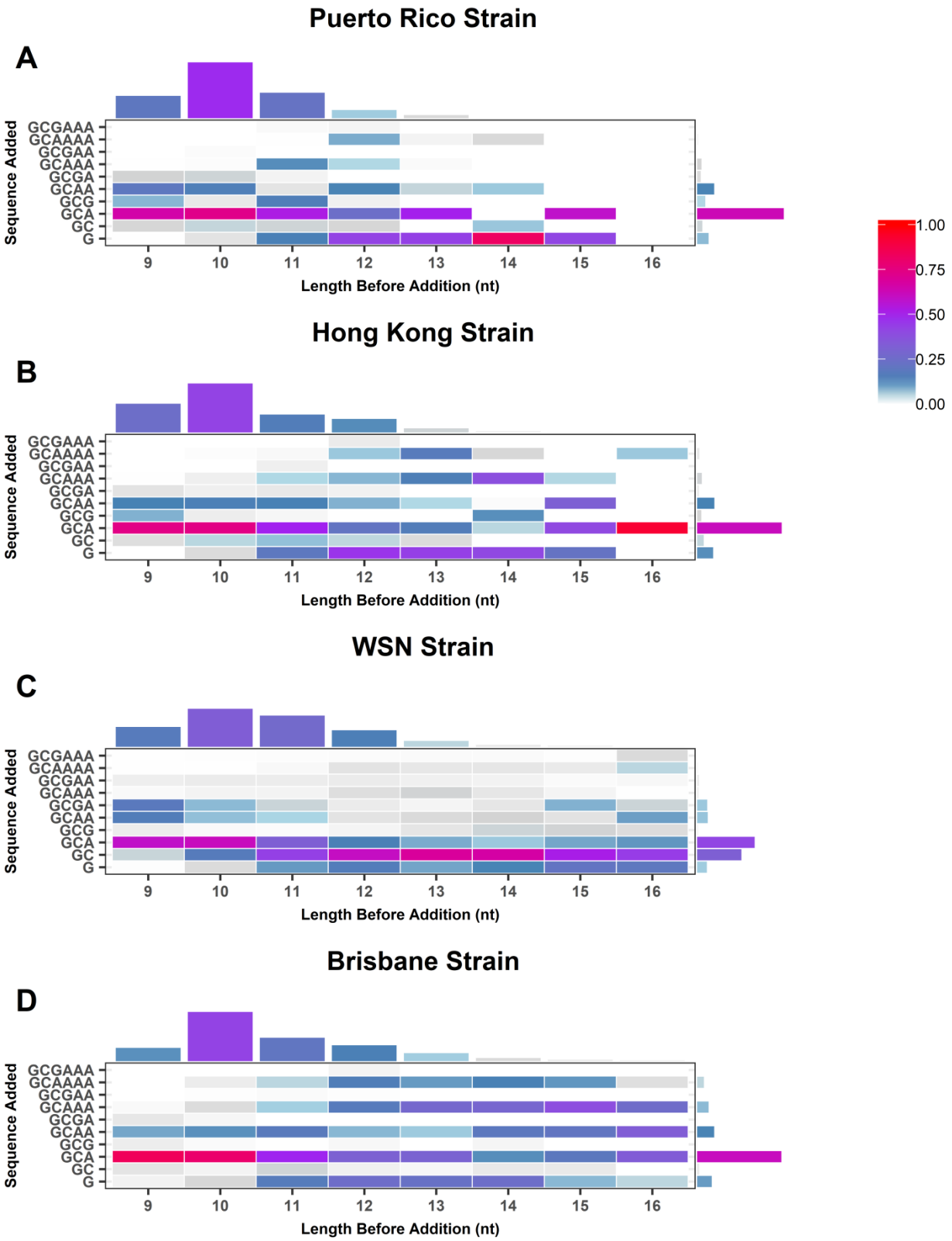
Puerto Rico Strain
A
Hong Kong Strain
B
WSN Strain
C
Brisbane Strain
D
Sequence Added
GCGAAA
GCAAAA
GCGAA
GCAAA
GCGA
GCAA
GCG
GCA
GC
G
9
10
11
12
13
14
15
16
Length Before Addition (nt)
1.00
0.75
0.50
0.25
0.00

Figure 3.5.3. Sequence Additions at all Start and End Lengths at a Given Length. The length of the primer prior to the addition and the frequency of a given addition at that length. Vertical bars show the frequency of lengthening events from that length; horizontal bars show the frequency of lengthening events with the specific nucleotide sequence addition.

3.6 Duplex Stability from Bond Energy and Stacking Interactions

RdRp exerts a stabilizing effect on the RNA duplex during transcription. This duplex would also be stabilized by the hydrogen bonds and stacking interactions of the nucleotides in this transcription duplex. Given that PAR has been shown to be related to the nucleotide directed at 3'U1 to prime transcription[138] and that there is less PAR when transcribing genes with a cytidine rather than an uridine at 3'Y4[1], [7]–[9], [138], it may be possible that these observations are related to free energy differences. Thus it is possible that the subpopulation of cap snatched primers which underwent PAR were also from a higher energy (less stable) subpopulation. Since 64.8±3.9% of PAR (omitting WSN) occurs following a three nucleotide long sequence addition, these differences in free energy was examined after three nucleotides of transcription (addition of a GCA or GCG). To determine if this was indeed the case, the free energy of all possible nucleotide sequence additions when transcribing the UCG[C/U]U portion of the vRNA conserved sequences was calculated using the Nearest Neighbor method[124], [210], [211] which takes into account hydrogen bonds and stacking interactions of base pairs. When a range of values was given, the highest free energy (least stable) values were used for these calculations, and an initiation penalty was omitted; an example calculation can be found with equation 1 in the Methods section. Some sequences underwent multiple rounds of PAR, and each of these PAR events was treated as independent transcription event. For example, a primer that underwent two rounds of realignment having a GCAA nucleotide sequence addition, followed by a GCA nucleotide sequence addition, then the transcription of the conserved sequences without PAR, will be treated as three separate transcription attempts (PAR1, PAR2, and transcription). The No Realignment and transcription subpopulation of the Realigned subpopulation were then combined into the Transcribed population. Given PAR may happen anywhere between one to six nucleotides of transcription and differences in free energy may be

related to the length of the nucleotide sequence addition, the Realigned population was subdivided into two categories: Will Realign (those that undergo PAR after four or more nucleotides have been transcribed); Realigns (those that realign after three nucleotides have been transcribed). For example, a primer in the Realigned population which underwent PAR after the addition of a GCA would be classified as Realigns; a primer in the Realigned population which underwent PAR after the addition of a GCGAA would be classified as Will Realign. These data are displayed in figure 3.6.1.

mRNA that underwent PAR had a higher average free energy than those which underwent transcription, and those which underwent PAR following a nucleotide sequence addition of three or more had an intermediate free energy. The magnitude of these differences is similar for the Puerto Rico, Brisbane, and WSN strains. The free energy of the subpopulation which underwent PAR following a three nucleotide sequence addition for these three strains was clustered around -6.54±0.03 kcal/mol. The subpopulation which undergoes complete transcription is clustered around -6.95±0.04 kcal/mol. This was a free energy difference of 0.41 kcal/mol between transcription that proceeds through the conserved sequence and transcription that resulted in PAR. While the trend was the same for the Brisbane strain, the energy difference between the Realigns and Transcribed subpopulations for the Brisbane strain was larger. The subpopulation which underwent transcription had an average free energy of -7.22 kcal/mol; whereas the subpopulation which underwent PAR had a mean free energy of -6.38 kcal/mol. The difference between these two subpopulations for the Brisbane strain was 0.84 kcal/mol. At an addition of three nucleotides, differences in the free energy come from three sources: the variation in the vRNA conserved sequences at position four (3'-UCG[C/U]UUUCGUCC), the nucleotide sequence addition (GC[A/G]), and the nucleotide directed at 3'U1. It is possible these

factors may differentially affect the Brisbane strain in comparison to the other three strains. These data suggest that primers that undergo PAR had a higher free energy RNA duplex on the basis of hydrogen bonds and stacking interactions.

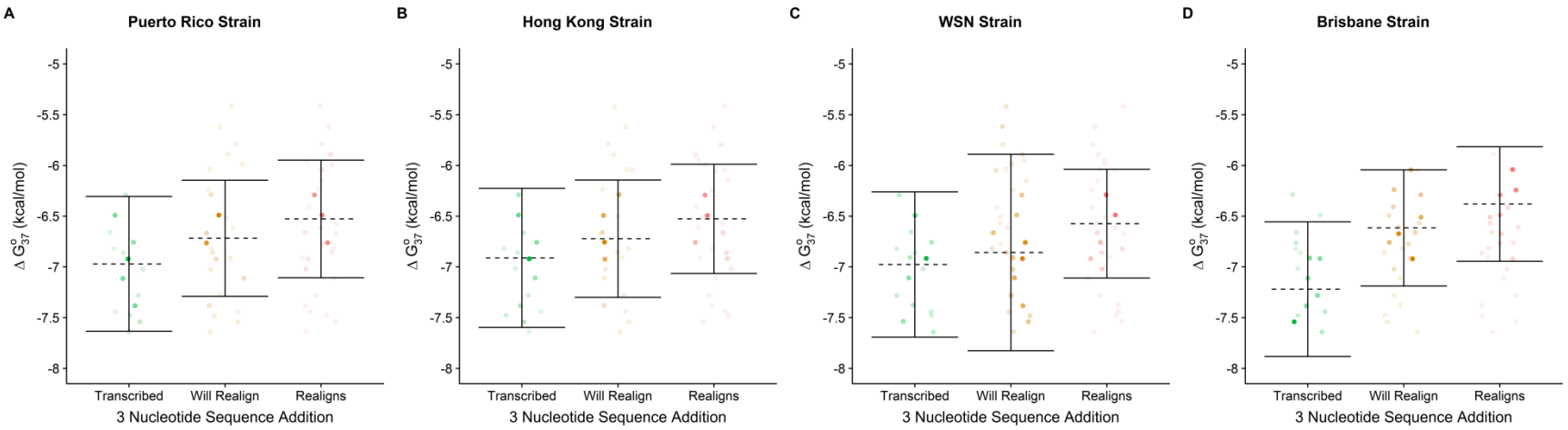
A
Puerto Rico Strain
B
Hong Kong Strain
C
WSN Strain
D
Brisbane Strain
Δ G°37 (kcal/mol)
-5
-5.5
-6
-6.5
-7
-7.5
-8
Transcribed
Will Realign
Realigns
3 Nucleotide Sequence Addition

Figure 3.6.1. Free Energy (ΔG^{o}_{37}) of Duplex Structure formed after a Three Nucleotide Sequence Addition Length. The free energy of the RNA duplex after three nucleotides have been transcribed. "Transcribed" refers to primers which will undergo complete transcription; "Will Realign" refers to the subpopulation of mRNA which will undergo PAR after four or more nucleotides have been transcribed; "Realigns" refers to the subpopulations of mRNA which undergoes PAR after a three nucleotide sequence addition. Dashed line represents the mean of each population, error bars are two standard deviations from this mean value.

Table 3.6.1. Free Energy (ΔG^{o}_{37}) Distributions and p-values

Strain	Series	Mean	Standard Deviation	Lower CI	Upper CI	p-value vs Transcribed	p-value vs Will Realign	p-value vs Realigns
Puerto Rico	Transcribed	-6.97	0.332	-6.31	-7.63	1	0	0
	Will Realign	-6.72	0.286	-6.15	-7.29	0	1	0
	Realigns	-6.53	0.290	-5.95	-7.11	0	0	1
Hong Kong	Transcribed	-6.91	0.343	-6.23	-7.60	1	0	0
	Will Realign	-6.72	0.289	-6.14	-7.30	0	1	0
	Realigns	-6.53	0.269	-5.99	-7.06	0	0	1
WSN	Transcribed	-6.98	0.358	-6.26	-7.69	1	0	0
	Will Realign	-6.86	0.484	-5.89	-7.83	0	1	0
	Realigns	-6.57	0.268	-6.04	-7.11	0	0	1
Brisbane	Transcribed	-7.22	0.332	-6.56	-7.88	1	0	0
	Will Realign	-6.62	0.286	-6.04	-7.19	0	1	0
	Realigns	-6.38	0.282	-5.82	-6.94	0	0	1

The mean free energy, standard deviation, and confidence intervals of the RNA duplex after three nucleotides have been transcribed. “Transcribed” refers to primers which will undergo complete transcription; “Will Realign” refers to the subpopulation of mRNA which will undergo PAR do not do so at that length of nucleotide sequence addition; “Realigns” refers to the subpopulations of mRNA which undergoes PAR at that length of nucleotide sequence addition. P-values are given along with the other subpopulation they are compared to; p-values of 1 are returned for comparing the same group to the same group; p-values of zero are reported instead of $p < 10^{-308}$.

3.7 3’ vRNA Conserved Sequences

One possible factor affecting RNA duplex stability during transcription of the 3’ vRNA conserved sequence is the natural variation in the pyrimidine residue position four of the vRNA template (3’-UCG[**U/C**]UUUCGUCC). This difference modifies the free energy from hydrogen bonds and stacking interactions of the RNA duplex during transcription. First I calculated the free energy for transcription from 3’C2 to 3’Y4 for the two conserved sequence variations using the Nearest Neighbor method[124]; these calculations took into account the base pair and stacking interactions. To focus solely on the affect of the conserved sequence, the free energy of the nucleotide residue directed at 3’U1 was omitted.

There was a free energy difference conferred by the pyrimidine residue present at 3’Y4. A GCA addition to the CGUU portion of the vRNA conserved sequence has a free energy of -5.09 kcal/mol; A GCG addition to the CGCU portion of the vRNA conserved sequence has a free energy of -5.71 kcal/mol. The net difference between these two conserved sequences from the differences in hydrogen bonds and stacking interactions was 0.62 kcal/mol. This free energy difference suggests that transcription of 3’C4 containing conserved sequences should result in less frequent PAR particularly for the Puerto Rico, Hong Kong, and WSN strains. This 0.62 kcal/mol free energy difference is larger than the between the sequences that undergo PAR and those that do not was ~0.41 kcal/mol for these three strains (Figure 3.6.1). This free energy difference alone may not be as important for the Brisbane strain which has a 0.84 kcal/mol energy difference (Figure 3.6.1).

I next examined if this free energy, from hydrogen bonds and stacking interactions conferred by the variation in the vRNA conserved sequences, difference did indeed affect the PAR frequency. I classified primers based on the vRNA segment from which they were transcribed, and then extracted the PAR frequency. I also analyzed these primers on the basis of

the conserved sequence transcribed for each segment. mRNA transcribed from the PB2, PB1, and PA segments were grouped into the 3'C4 templated mRNA, while the other five segments were grouped into the 3'U4 templated mRNA. For the Hong Kong strain, the PA segment has the 3'U4 template rather than the 3'C4 template, so it was classified with the 3'U4 genes. These PAR frequencies were graphed relative to the strain specific PAR rate. These data are in figure 3.7.1. Fold change values were calculated based as PAR frequency divided by PAR frequency for the specific strain, and then averaged among all four strains.

mRNA transcribed from 3'U4 genes were 1.43±0.13 fold more likely to undergo PAR (20.3±1.1%), whereas those transcribed from 3'C4 genes were 0.480±0.239 fold less likely to undergo PAR (6.77±3.10%). There was some background variation of PAR rates between mRNA transcribed from genome subunits; however on average, all segments follow the trend of their conserved sequence. This difference was conferred by the vRNA conserved sequences rather than another factor. mRNA transcribed from the PA gene underwent PAR more frequently solely for the Hong Kong strain (1.04 fold; 15.7% versus 0.74±0.35 fold; 10.6±5.3% for the other three strains). This difference for the PA gene was likely due to the difference in template between the Hong Kong strain (3'U4) and the other three strains (3'C4). While the PAR rate was increased for 3'U4 segments in the Brisbane strain data set, the PAR rate for the 3'C4 segments was higher than in the other three strains. This is in agreement with the larger free energy difference between the subpopulation that underwent PAR and the subpopulation that did not undergo PAR, as seen in the Brisbane strain (Figure 3.6.1). The 0.62 kcal/mol decrease in free energy is larger than the difference for the other three strains (0.41 kcal/mol), but not as large as the difference for the Brisbane strain (0.84 kcal/mol). While the reduced PAR rate when transcribing the 3'C4 template is consistent across all four strains, it is likely that other factors

may be of more importance to the Brisbane strain, such as the nucleotide sequence addition to the template.

An alternative explanation for these variations is the difference in free energy from hydrogen bonds and stacking interactions in the 3' 5' base pair region of the vRNA promoter. This region contains five base pairs: 11 to 15 of the 5' end pair with nucleotides ten to 14 of the 3' end[23]. These base pair regions have three base pairs between the shared conserved sequences, and two base pairs from the segment specific untranslated regions forming a $\begin{matrix} 5' - AGGNN \\ 3' - UCCNN \end{matrix}$ duplex. This duplex has variable free energy ranging from -7.65 kcal/mol to -9.61 kcal/mol. I calculated the free energy for each of these 3' 5' base pair regions using the sequences submitted to the NCBI Influenza Virus Resource database[213] for the Hong Kong strain, and compared this to the PAR frequencies observed for each subunit. These data are in table 3.7.1.

It is unlikely that the differences seen between mRNA transcribed from each gene segment arise from the differences in free energy from base pairs and stacking interactions. PAR frequency is not correlated with the energy differences in the 3' 5' base pair region ($R^2 = 0.043$), even when factoring in the differences at 3'Y4 ($R^2 = 0.020$), or the ratio between the 3' 5' region and the energy differences at 3'Y4 ($R^2 = 0.138$). If the energy differences from this region were a key determinant of PAR, mRNA transcribed from the PB1 segment (-9.61 kcal/mol) should undergo PAR more frequently than the HA segment (-8.91 kcal/mol), even when factoring in the difference conferred by the 4^{th} nucleotide of the conserved sequence (-3.90 kcal/mol for PB2, and -3.82 kcal/mol for HA). Despite the 0.08 kcal/mol difference in these two regions after factoring in the template, mRNA transcribed from the HA segment underwent PAR 3.20 fold more frequently than those transcribed from the PB1 segment. While it is still possible that this

region acts as an energy barrier promoting PAR, similar to the tight 5’ end hold leading to polyadenylation[126], PAR is not related to the energy differences between the base pairs in this region.

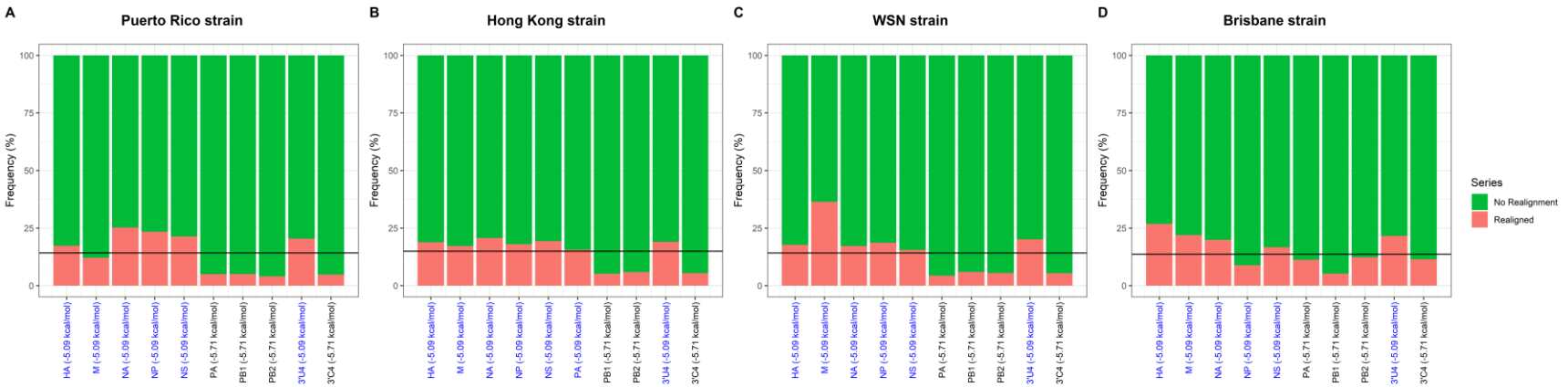
A
Puerto Rico strain
B
Hong Kong strain
C
WSN strain
D
Brisbane strain
Frequency (%)
100
75
50
25
0
HA (-5.09 kcal/mol)
M (-5.09 kcal/mol)
NA (-5.09 kcal/mol)
NP (-5.09 kcal/mol)
NS (-5.09 kcal/mol)
PA (-5.71 kcal/mol)
PB1 (-5.71 kcal/mol)
PB2 (-5.71 kcal/mol)
3'U4 (-5.09 kcal/mol)
3'C4 (-5.71 kcal/mol)
Series
No Realignment
Realigned

Figure 3.7.1. Frequency Distributions of PAR relative to IAV Subunit Transcribed and vRNA Conserved Sequences. The distributions of primers between the Realigned (Red) and No Realignment (Green) subpopulations; the horizontal black line indicates the PAR frequency for that given strain. Red bars above the black line indicate a PAR bias. Blue coloured segments indicate that this segment has the 3'U4 template, whereas black coloured segments indicate that this segment has the 3'C4 template.

Table 3.7.1. 3’ 5’ Base Pair Region, Free Energy, and PAR Frequency for the Hong Kong Strain

Segment	PAR	3' Seq	5' Seq	3' 5' Energy	Energy 3'C2-3'Y4	Diff.	Ratio
HA	18.9	UCCCC	AGGGU	-8.91	-5.09	-3.82	1.75
M	17.2	UCCAU	AGGUA	-7.65	-5.09	-2.56	1.50
NA	20.7	UCCUC	AGGAG	-9.03	-5.09	-3.94	1.77
NP	18.0	UCCCA	AGGGU	-9.60	-5.09	-4.51	1.89
NS	19.3	UCCCA	AGGGU	-9.60	-5.09	-4.51	1.89
PA	15.6	UCCAU	AGGUA	-7.65	-5.09	-2.56	1.50
PB1	5.19	UCCGU	AGGCA	-9.61	-5.71	-3.90	1.68
PB2	5.91	UCCAG	AGGUC	-9.19	-5.71	-3.48	1.61

Column one lists the vRNA, column two lists the prime and realign rate in percent. Columns three and four list the nucleotides from the 3’ and 5’ end that form the base pairs in the 3’ 5’ base pair region. Column five lists the free energy from hydrogen bonds and stacking interactions for the 3’ 5’ base pair region. Column six lists the free energy when transcribing from 3’C2 to 3’Y4 of the vRNA conserved sequences. Column seven is the difference between columns five and six, column eight is the ratio of column six to column five.

3.8 Conserved Sequences and Nucleotide Sequence Additions

RdRp has been observed to transcribe both guanine and adenine residues at the 4th nucleotide in the vRNA conserved sequence regardless of the pyrimidine at this position[1], [7]–[9]. Simply, both $\frac{GCA}{CGCU}$ and $\frac{GCG}{CGUU}$ duplexes have been observed rather than the Watson Crick $\frac{GCG}{CGCU}$ and $\frac{GCA}{CGUU}$ duplexes. The nucleotide sequence added would modify the free energy from hydrogen bonds and stacking interactions. To investigate this difference I calculated the free energy from stacking interactions and hydrogen bonds for all nucleotide sequence additions of three nucleotides or longer added by PAR to each vRNA conserved sequences, as well as the free energy. To focus on the free energy of the nucleotide sequence addition, the free energy of the nucleotide residue directed at 3'U1 was omitted. These data are in table 3.8.1.

Free energy is indeed modified by the purine residue directed at pyrimidine residue at position four of the vRNA conserved sequence. The most stable free energies are obtained when the correct Watson Crick base pair is transcribed. When the A-C mismatch or G-U wobble pair is transcribed forming a $\frac{GCA}{CGCU}$ or $\frac{GCG}{CGUU}$ duplex respectively the free energy is increased by 0.87 kcal/mol. Incorporation of the A-C mismatch into the duplex from transcription of four or more nucleotides (transcription of $\frac{GCAA}{CGCUU}$ or longer) results in a very large increase in free energy (increase > 5kcal/mol). These A-C mismatches or bulge loops would also likely disrupt the structure of the RNA duplex in the RdRp active site, which may result in a further loss of stability. These data suggest that transcription errors at 3'Y4 would result in a net loss of RNA duplex stability.

In these data sets, GCA containing additions (81.7±4.5%) occur more frequently than the equivalent GCG addition (4.84±3.81%), and it is possible that PAR consists of a GCA

containing addition regardless of conserved sequence transcribed. I examined the frequency of these additions on the basis of conserved sequence transcribed. Given our interest was in the nucleotide sequences added by PAR, the No Realignment subpopulation was omitted from further analysis. For the Realigned subpopulation, the nucleotide sequence added at each PAR event was extracted. Some sequences underwent multiple rounds of PAR and each of these PAR events was treated as an independent event. For example, a primer, which underwent PAR twice, with a GCA and a GCAA addition, would be considered twice as both a GCA addition and a separate GCAA addition. I analyzed these PAR events on both templates (All), and on either the 3'U4 or 3'C4 templates. These data were then analyzed relative to all PAR events on the given template(s), and are displayed in figure 3.8.1. Since PAR typically occurs following a three nucleotide sequence addition, these data were also analyzed omitting shorter G or GC additions.

For the Brisbane strain, PAR is similar regardless of conserved sequence transcribed. There was a clear preference for GCA nucleotide sequence additions (59.2% on the 3'U4 template and 61.4% on the 3'C4 template). Overall, 84.2% of PAR additions on the 3'C4 template were GCA containing additions, whereas 1.61% of PAR additions were the correct GCG containing additions. This increase in transcription of the higher free energy AC mismatch (0.87 kcal/mol difference) is consistent with the 0.84 kcal/mol free energy difference seen between the subpopulation of primers which underwent PAR at three nucleotides and those which underwent complete transcription for the Brisbane strain (Figure 3.6.1). Transcription of the AC mismatch likely explains the difference in free energy between these subpopulations for the Brisbane strain.

For all strains except Brisbane, the vRNA conserved sequence transcribed modifies the nucleotide sequence added by PAR. Transcription of the 3'U4 template leads to PAR typically

after transcription of a GCA (three nucleotides; 61.7±10.3%), for the 3'C4 template this typically occurred before transcription of three nucleotides (before the free energy decrease from the GC pair at 3'C4; G 34.0±13.3% or GC 25.2±1.6). For the 3'C4 template, PAR of three or more nucleotides typically occurs after an adenine residue has been transcribed either by the longer GCGA (22.8±13.0%) addition or by transcription of the AC mismatch (GCA 7.55±3.13%). The correct GCG addition on the 3'C4 template occurred fairly infrequently (4.28±3.58%), as is the transcription of a GCG addition on the 3'U4 template (2.87±2.48%). It is possible that when re-priming transcription following a GCG addition, that the primer may be able to prime transcription internally at 3'C2, initiating transcription with a $\frac{GCG}{UC}$ duplex, as has been observed for cap snatched primers ending in G[110], [138]. If all GC additions were GCG additions that were used to prime transcription internally, the frequency of GCG additions on the 3'C4 template would be 30.7±1.7% and 11.2±14.9% when transcribing 3'C4 and 3'U4 templated genes respectively. This may explain the increased frequency of GC additions on the 3'C4 template versus the 3'U4 template for both the Puerto Rico and Hong Kong strains, though GC additions occur more frequently when transcribing 3'U4 templated mRNA for the WSN strain (32.8% 3'U4 and 27.0% 3'C4). Even with the combination of GC and GCG additions, PAR typically does not occur following a three nucleotide sequence addition when transcribing the 3'C4 template.

Aside from 3'C4 templated mRNA of the Brisbane data set, transcription of the incorrect purine at 3'Y4 tends to be an infrequent event. When transcribing the 3'U4 template, GCG containing additions occurred as 2.75±2.14% of PAR additions whereas the correct GCA containing addition occurred as 82.3±15.1% of PAR additions among all four strains. For the 3'C4 template, GCA containing additions occurred as 7.94±3.01% of PAR additions, whereas

the correct GCG containing additions occurred as 32.9±14.1% of PAR additions. It is important to note that PAR did occur more frequently when transcribing 3'U4 templated mRNA versus 3'C4 templated mRNA (Figure 3.8.1), so these incorrect additions associated with PAR may be similar when examining all transcription of the conserved sequences rather than just PAR additions.

Table 3.8.1 Free Energy from Hydrogen Bonds and Stacking Interactions for PAR additions of Three Nucleotides or Longer on both vRNA Conserved Sequences

Template	Addition	GCA	GCG	Difference
CGUU	GC[A/G]	-5.09	-4.22	0.87
CGUUU	GC[A/G]A	-5.99	-5.21	0.78
CGUUUU	GC[A/G]AA	-6.89	-6.11	0.78
CGUUUUC	GC[A/G]AAA	-7.69	-6.91	0.78
CGCU	GC[A/G]	-4.84	-5.71	0.87
CGCUU	GC[A/G]A	-1.76	-7.61	5.86
CGCUUU	GC[A/G]AA	-2.92	-8.51	5.59
CGCUUUC	GC[A/G]AAA	-3.99	-9.31	5.33

Template is the nucleotide sequence of the 3'U4 (grey) or 3'C4 template from 3'C2. Addition indicates the nucleotide sequence transcribed with the variation at the 3rd (3'Y4 directed) nucleotide. GCA indicates the free energy from GCA containing additions, GCG indicates the free energy from GCG containing additions, both in kcal/mol. Difference indicates the free energy difference between the GCA and GCG containing addition.

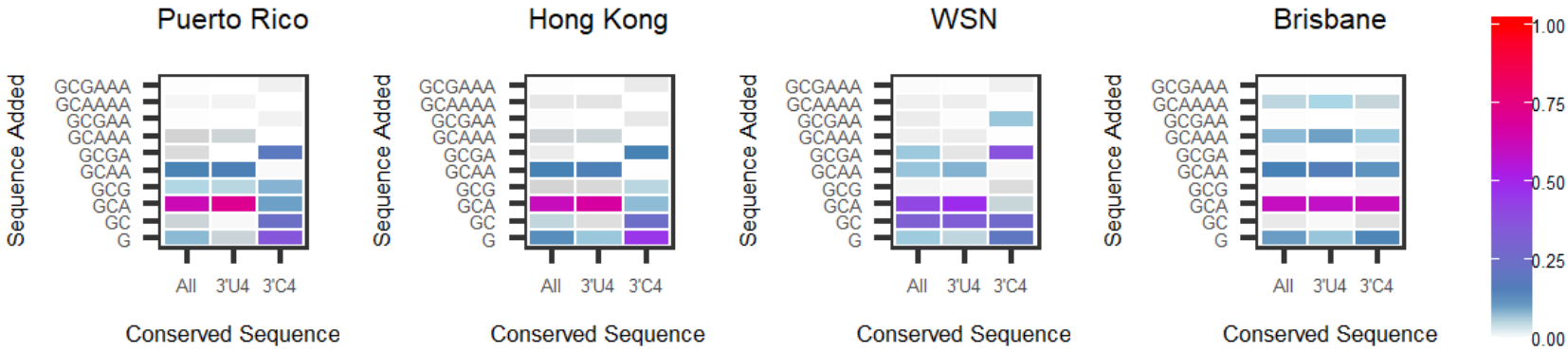
Puerto Rico
Hong Kong
WSN
Brisbane
Sequence Added
GCGAAA
GCAAAA
GCGAA
GCAAA
GCGA
GCAA
GCG
GCA
GC
G
All
3'U4
3'C4
Conserved Sequence
1.00
0.75
0.50
0.25
0.00

Figure 3.8.1. Nucleotide Sequence Additions on a Given Viral Conserved Sequence. Shows the nucleotide sequence transcribed prior to each PAR event either over-all (All), or on the 3'U4 or 3'C4 template as indicated.

3.9 Primer Sequence End

Cap-snatching viruses have been shown to selectively cleave host mRNA, for use of primers, such that the 3' end of the cap snatched primer is compatible with the viral conserved sequences[4]–[7], [156], [157], [160]–[168], [170]–[172]. Furthermore, the frequency of an mRNA being cap snatched increases as the nucleotide sequence compatibility of the mRNA with the viral conserved sequence increases[4], [5], [172], [174]. This has been shown for IAV which preferentially cap snatches primers with a CAG in the compatible length range to form a $\frac{CAG}{UC}$ duplex at initiation[4]–[7]. Despite this preference, IAV may cleave non-compatible primers and prime transcription without base pairing[1], [6]–[9], [138]. Figure 3.3.1 shows that the 3'U1 directed nucleotide, while different in enrichment among strains, was primarily an adenine residue (33.9±6.1%) followed by guanine (30.0±4.5%), cytidine (24.0±7.2%), and uridine (12.3±4.0%). It is possible that the nucleotide directed at 3'U1 was selected for on the basis of free energy when paired with 3'U1. I next compared the relative abundance of the 3' most nucleotide of each matched sequence both for each strain and the average of all four strains with the free energy from stacking interactions and hydrogen bonds when directing this nucleotide at 3'U1. Free energy was calculated using the Nearest Neighbor method[124]. These data are in figure 3.9.1.

For each strain, the nucleotide to be directed at 3'U1 (the 3' most nucleotide of the matched sequence) does not appear to be selected for on the basis of free energy (R^2 = 0.727±0.154); however, the average selection of all four strains is well correlated to the free energy (R^2 = 0.963). Sequence specific cleavage requires the compatible sequence to be within the compatible length range[4]–[6], and that mRNA with this sequence be at least moderately

abundant[4], [5], [172], [174]. It is possible the deviance from this overall correlation seen in each strain is due to the abundance of mRNA with a compatible sequence within the optimal length range. Divergence in cap snatched mRNA has been reported between different plants infected with rice stripe and rice grassy stunt viruses[163]. An overall selection of adenine above the enrichment within the reference genome (33.9±6.1% vs. 21.0±1.3%) for each strain, as well as a selection against uridine for each strain (12.3±4.0% vs. 21.3±1.6%) was observed. Guanine and cytidine selected for near their abundance in the reference genome (29.9±4.5% vs. 30.6±1.4% and 24.0±7.2% vs. 27.1±1.5%). It is possible that this process overall tends to select for a 3'U1 directed nucleotide that will lead to the lowest free energy from hydrogen bonds and stacking interactions as best as possible within the constraints of abundance.

Others have shown that the identity of the nucleotide residue directed at 3'U1 is an important determinant of PAR[138], it is possible that this relationship is mediated by the free energy from hydrogen bonds and stacking interactions rather than the identity of the base pair. To determine if the nucleotide directed at 3'U1 of the conserved sequences could be related to PAR, I compared the free energy from hydrogen bonds and stacking interactions and the PAR frequency for all 3'U1 directed nucleotides. These data are displayed in figure 3.9.2. Fold change values were calculated based as PAR frequency divided by PAR frequency for the specific strain, and then averaged among all four strains.

There was a good correlation between PAR frequency and the free energy of the nucleotide directed at 3'U1 to prime transcription for all strains (R^2 = 0.915±0.059), which is stronger when the PAR frequencies of all four strains are averaged and then correlated (R^2 = 0.992). This correlation is weakest in the Brisbane strain (R^2 = 0.831), indicating that the nucleotide directed at 3'U1 alone may not be as important for determining PAR for this strain,

and other factors such as the length or the vRNA template may be more important. Priming with a U-U mismatch pair had the highest free energy (-1.2 kcal/mol) and was 2.01±0.20 fold (28.8±3.5%) more likely to undergo PAR. Priming with a C-U mismatch pair had the second highest free energy (-1.4 kcal/mol) and was 1.33±0.31 fold (18.8±3.8%) more likely to undergo PAR. Priming with a guanine residue (maximum of -1.57 kcal/mol), which forms a G-U wobble pair, was close to the strain average of PAR (13.4±2.8%) and was 0.94±0.18 fold more likely to undergo PAR. Priming with the Watson-Crick A-U base pair (maximum of -1.73 kcal/mol) resulted in a 0.5±0.13 fold reduction in PAR (7.13±1.79%). The base directed at 3'U1 is important in determining PAR *in vivo*, which is in agreement with previous *in vitro* experiments[138]. This relationship was determined by the free energy from hydrogen bonds and stacking interactions rather than the identity of the nucleotide.

RdRp has been reported to prime transcription internally with primers ending in a guanine residue that is directed at 3'C2[138]. In this data set there was a non-random enrichment of guanine one nucleotide 3' of the matched sequence (Figure 3.3.1), which may have been directed at 3'C2 to prime transcription internally. I also saw that this guanine residue was associated with a modest reduction in the frequency of PAR. To determine if initiating internally or terminally impacted the PAR frequency on the basis of the nucleotide directed at 3'U1, I reanalyzed the data making a distinction between initiating internally or terminally. These data are displayed in figure 3.9.3.

Free energy and PAR rates are better correlated when no distinction is made for the complimentary guanine (R^2 = 0.915±0.059 without distinction, R^2 = 0.908±0.052 when priming terminally, and R^2 = 0.871±0.112 when priming internally). Similar to the overall data, this correlation is better when the PAR frequencies for each strain are averaged and then correlated

(R^2 = 0.958 terminally, and R^2 = 0.983 internally), however the correlation is higher when no distinction is made (R^2 = 0.992). It is possible that other factors that influence PAR (length and vRNA conserved sequence transcribed) disrupt this correlation on the reduced data set when this distinction is made.

For all 3'U1 directed nucleotide residues, priming internally with a complimentary guanine residue resulted in a 0.690±0.206 fold reduction in PAR in comparison to priming terminally at 3'U1. Without the complimentary guanine, only priming terminally with an adenine residue had a PAR frequency lower than the strain average. The ability to prime transcription internally reduced the PAR frequency for primers with a 3'U1 directed guanine and, for the Puerto Rico and Hong Kong strains, cytidine residues to below the strain PAR frequency (see Figure 3.9.3). Priming with an uridine residue remained associated with PAR, though at a reduced rate, when priming internally with a complimentary guanine residue. It would appear as though priming internally rather than terminally results in a reduction in free energy (more stable duplex) that is not directly apparent in the free energy calculations for stacking interactions and hydrogen bonds. Using the difference between the two linear models to predict the free energy from the average PAR rates when priming terminally, it can be estimated that priming terminally resulted in a PAR rate reduction that was consistent with a -0.13 kcal/mol energy reduction.

These data suggest that PAR was related to the nucleotide directed at 3'U1. Both the selection of this nucleotide and the PAR frequency are related to the free energy from hydrogen bonds and stacking interaction. The complimentary guanine residue reduces the PAR rate regardless of 3'U1 directed nucleotides with a similar effect of a -0.13 kcal/mol energy reduction. In agreement with previous publications I find that the PAR rate is more dependent on the nucleotide directed at 3'U1 than the presence of absence of the complimentary guanine[138].

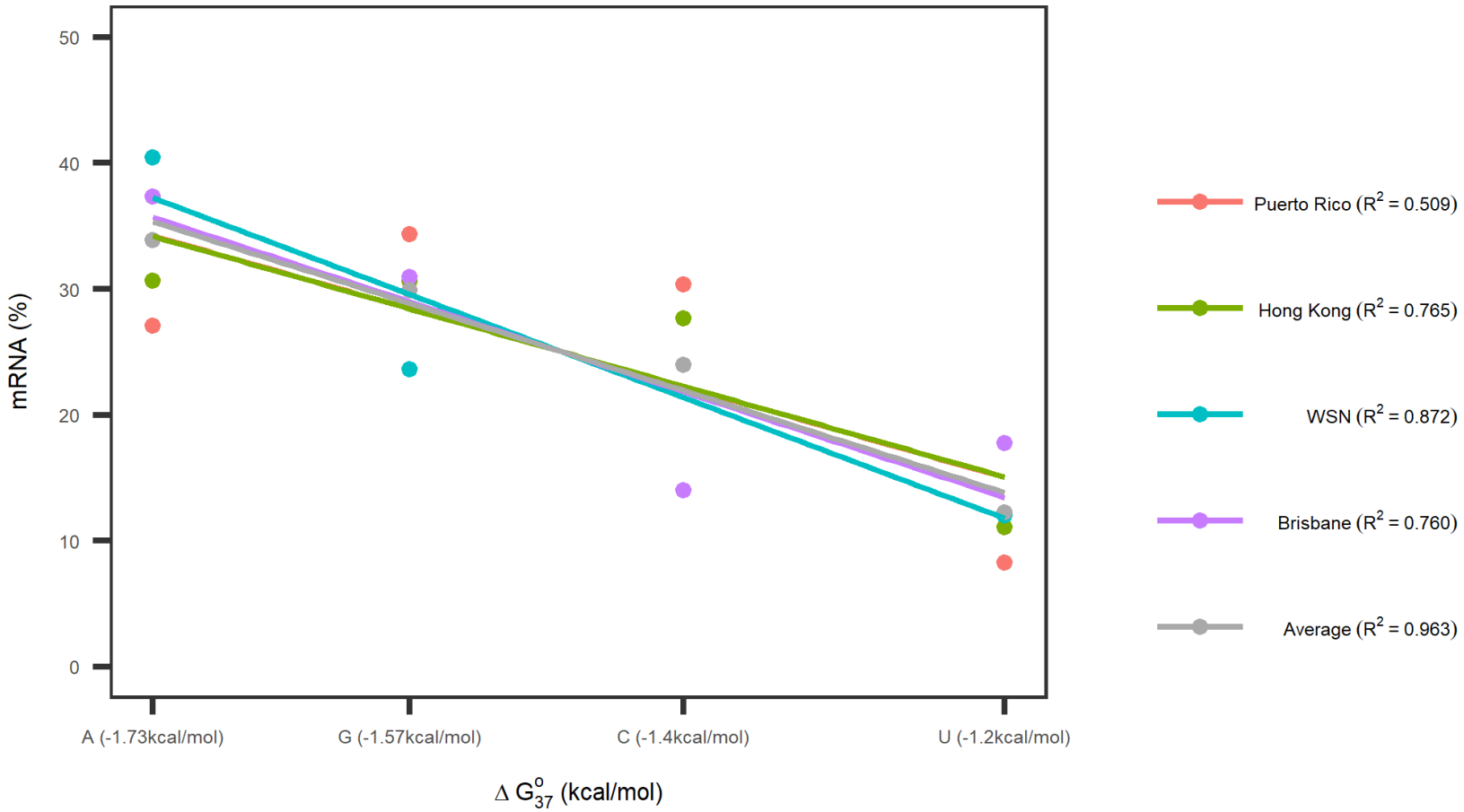
50
40
30
20
10
0
mRNA (%)
A (-1.73kcal/mol)
G (-1.57kcal/mol)
C (-1.4kcal/mol)
U (-1.2kcal/mol)
Δ G°37 (kcal/mol)
Puerto Rico (R2 = 0.509)
Hong Kong (R2 = 0.765)
WSN (R2 = 0.872)
Brisbane (R2 = 0.760)
Average (R2 = 0.963)

Figure 3.9.1. Nucleotide Enrichment and Free Energy When Priming Transcription. The relative abundance of the 3' most nucleotide of the matched sequence (the nucleotide directed at 3'U1 to prime transcription) and the free energy when this nucleotide is directed at 3'U1 for each strain as well as the average value for all four strains.

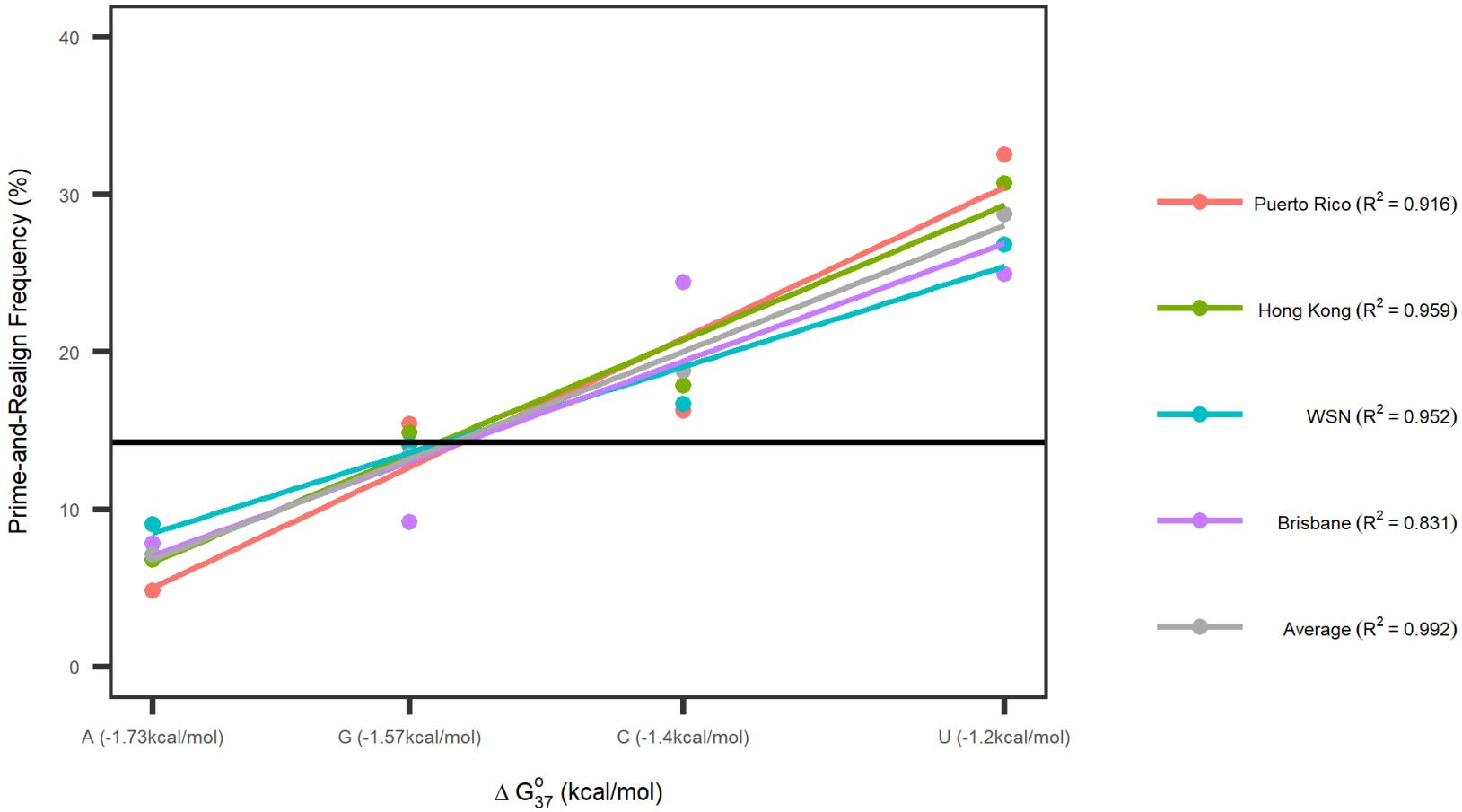
Prime-and-Realign Frequency (%)
40
30
20
10
0
A (-1.73kcal/mol)
G (-1.57kcal/mol)
C (-1.4kcal/mol)
U (-1.2kcal/mol)
Δ G°37 (kcal/mol)
Puerto Rico (R2 = 0.916)
Hong Kong (R2 = 0.959)
WSN (R2 = 0.952)
Brisbane (R2 = 0.831)
Average (R2 = 0.992)

Figure 3.9.2. Free Energy of the 3'U1 Directed Nucleotide and PAR Frequencies. The nucleotide directed at 3'U1 and the associated PAR frequency for each of the four strains.

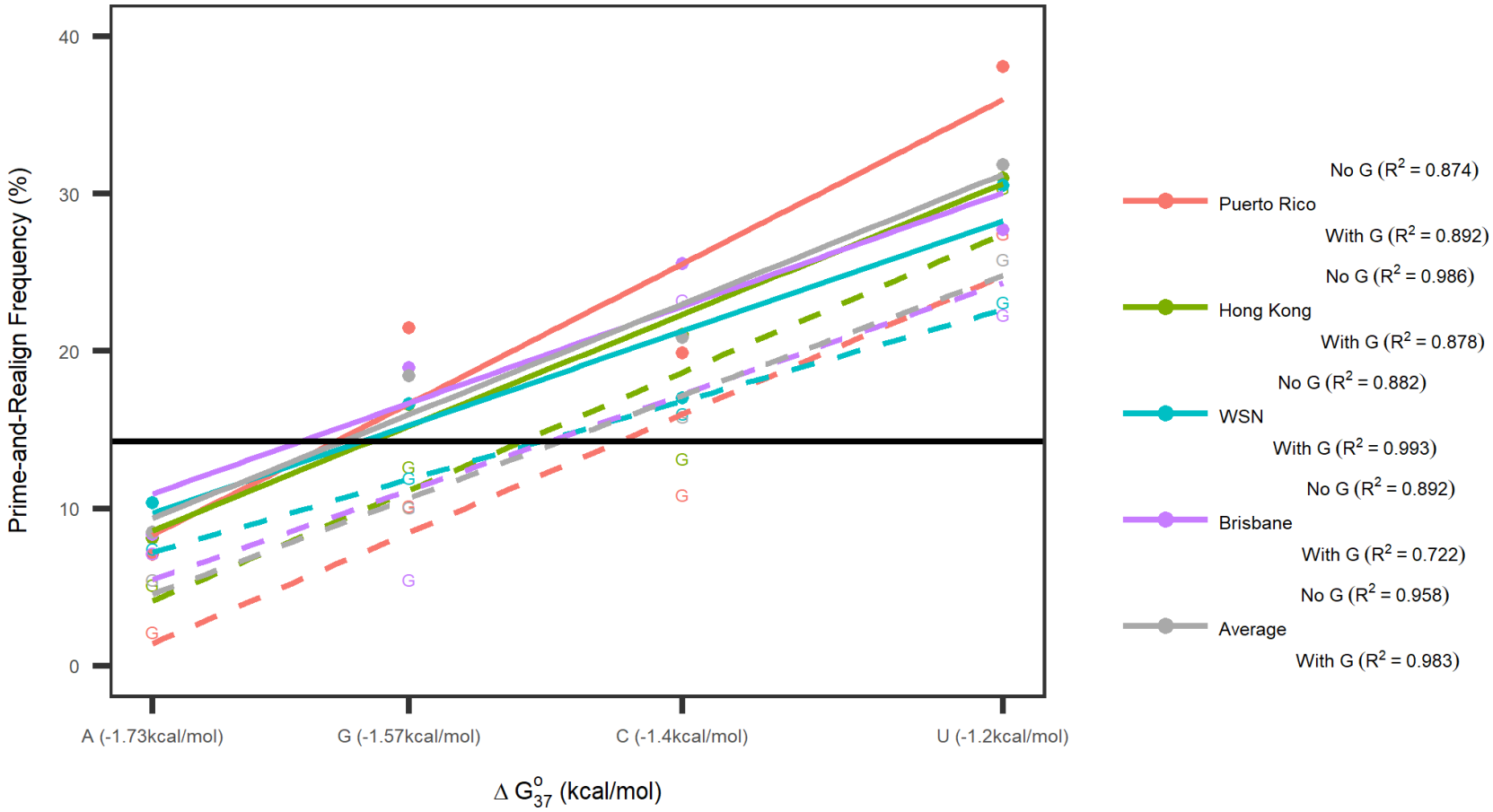
Prime-and-Realign Frequency (%)
40
30
20
10
0
A (-1.73kcal/mol)
G (-1.57kcal/mol)
C (-1.4kcal/mol)
U (-1.2kcal/mol)
Δ G°37 (kcal/mol)
No G (R2 = 0.874)
Puerto Rico
With G (R2 = 0.892)
No G (R2 = 0.986)
Hong Kong
With G (R2 = 0.878)
No G (R2 = 0.882)
WSN
With G (R2 = 0.993)
No G (R2 = 0.892)
Brisbane
With G (R2 = 0.722)
No G (R2 = 0.958)
Average
With G (R2 = 0.983)

Figure 3.9.3. Free Energy of the 3'U1 Directed Nucleotide and PAR Frequencies Factoring in the 3'C2 directed Guanine. PAR frequencies making the distinction for the presence of the complimentary guanine residue that is directed at 3'C2. For the right figure, circles represent priming at 3'U1 only, "G" represents priming with a guanine directed at 3'C2; solid line indicates linearization of nucleotides without the G, dashed line represents priming with a 3'C2 directed G.

3.10 Relative Guanine and Cytidine Content

High guanine and cytidine (G+C) content is an important feature of mammalian 5' untranslated regions (UTRs), and may be related to the breadth of mRNA expression[214], [215]. PAR adds a G to GC[A/G]AAA nucleotide sequence which can potentially modify the G+C content of the IAV mRNA 5' UTR. PAR may increase the relative G+C content. For example, the matched primer 5'-CUUCGCCCAC had 70% relative G+C content prior to PAR; after a PAR addition of GCG, this primer (5'-CUUCGCCCACGCG) had 76.9% G+C content. PAR may also lower the relative G+C content. For example, the matched primer 5'-ACAGGGCCCGCA had 75% G+C content; after a PAR addition of GCAAAA, this primer (5'- ACAGGGCCCGCAGCAAAA) had 61.1% G+C content. I next examined these data to see if G+C content was related to PAR. Primers were classified into the No Realignment or Realigned subpopulations; the Realigned subpopulation was subdivided into the Pre Realignment and Post Realignment groups. The amount of guanine and cytidine residues within these mRNA sequences was extracted and taken relative to the length of the mRNA primer. The G+C content of the reference genome (within 100 nucleotides of the transcription start sites) was also extracted. The distribution of these populations as well as their mean and 95% confidence intervals are displayed in figure 3.10.1.

These data reveal that all subpopulations are all highly similar within each individual strain. The Brisbane strain has a lower G+C content in comparison to the other three strains; this may be due to extraction of mRNA at 12 hours post infection for the Brisbane data set[8], as IAV infection is known to alter the cellular transcriptome[197]–[199]. The No Realignment subpopulation is clustered around 56.2±6.7%, the Pre-Realignment subpopulation is of a slightly lower G+C content clustered around 53.8±5.2%, and the Post-Realignment subpopulation is clustered around 56.6±5.3%. While the Pre-Realignment subpopulation is indeed on average of lower G+C content versus the No Realignment subpopulation the confidence intervals of these

populations have a very high degree of overlap, and the subpopulations themselves are highly heterogeneous. It is unlikely that PAR occurs to raise or lower G+C content as the majority of primers are within the same relative G+C content range.

To assess the effects of the relative G+C content of cap snatched primers on the frequency of PAR, I examined the enrichment of primers in both the No Realignment and Realigned subpopulations on the basis of relative G+C content. G+C content distributions were divided into ten groups representing the percentages between 0% and 100%. The relative distribution of the two groups for each G+C content range was then extracted and plotted relative to the strain frequency of PAR. These data are in figure 3.10.2. There is no clear trend in PAR frequencies seen on the basis of G+C content. Most values remain close to the strain PAR frequency for that specific strain (~14.3%). Due to the lack of a distinct trend, and low number of points of deviation from the strain PAR frequency, it is unlikely that relative G+C content is related to PAR.

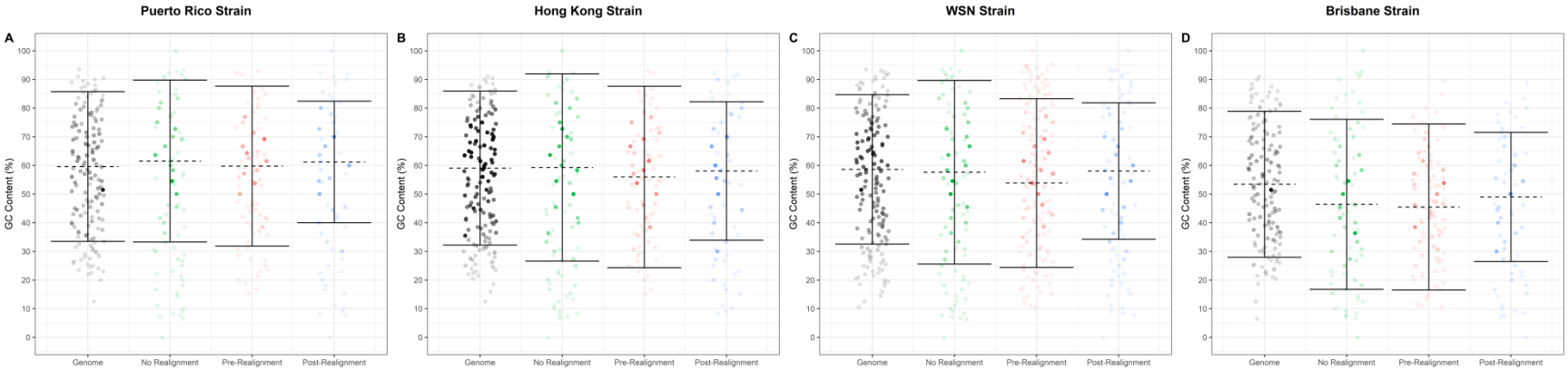
A
Puerto Rico Strain
B
Hong Kong Strain
C
WSN Strain
D
Brisbane Strain
GC Content (%)
0
10
20
30
40
50
60
70
80
90
100
Genome
No Realignment
Pre-Realignment
Post-Realignment

Figure 3.10.1. Relative Guanine and Cytidine Residue Enrichment of the Cap-Snatched Genome, No Realignment, Pre-Realignment, and Post-Realignment Subpopulations. The relative G+C content of primers in each subpopulation. Genome (black) values are based on the window wherein cap-snatched mRNA were matched (±100 nucleotides of known transcription start sites). No Realignment (green), Pre-Realignment (red), and Post-Realignment (blue) subpopulations are relative to primer length.

Table 3.10.1. Relative G+C Content Distributions and p-values

Strain	Series	Mean	Standard Deviation	Lower CI	Upper CI	p vs. No Realign	p vs post-realign	p vs. pre-realign
Hong Kong	No Realign	59.3	16.3	26.6	92.0	1	0	0
	Post-Realign	58.1	12.1	33.9	82.2	0	1	0
	Pre-Realign	56.0	15.8	24.3	87.6	0	0	1
Puerto Rico	No Realign	61.5	14.1	33.3	89.8	1	0	0
	Post-Realign	61.2	10.6	40.0	82.4	0	1	0
	Pre-Realign	59.8	14.0	31.9	87.7	0	0	1
WSN	No Realign	57.6	16.0	25.6	89.7	1	0	0
	Post-Realign	58.0	11.9	34.2	81.8	0	1	0
	Pre-Realign	53.9	14.7	24.4	83.3	0	0	1
Brisbane	No Realign	46.4	14.8	16.7	76.1	1	4.41E-304	5.35E-40
	Post-Realign	49.0	11.3	26.5	71.5	4.41E-304	1	0
	Pre-Realign	45.5	14.5	16.5	74.5	5.35E-40	0	1

The mean relative G+C content (% composition of primers), standard deviation, and confidence intervals for the No Realignment (No Realign), Pre-Realignment (Pre-Realign), and Post-Realignment (Post-Realign) subpopulations. P-values are given along with the other subpopulation they are compared to; p-values of 1 are returned for comparing the same group to the same group; p-values of zero are reported instead of $p < 10^{-308}$.

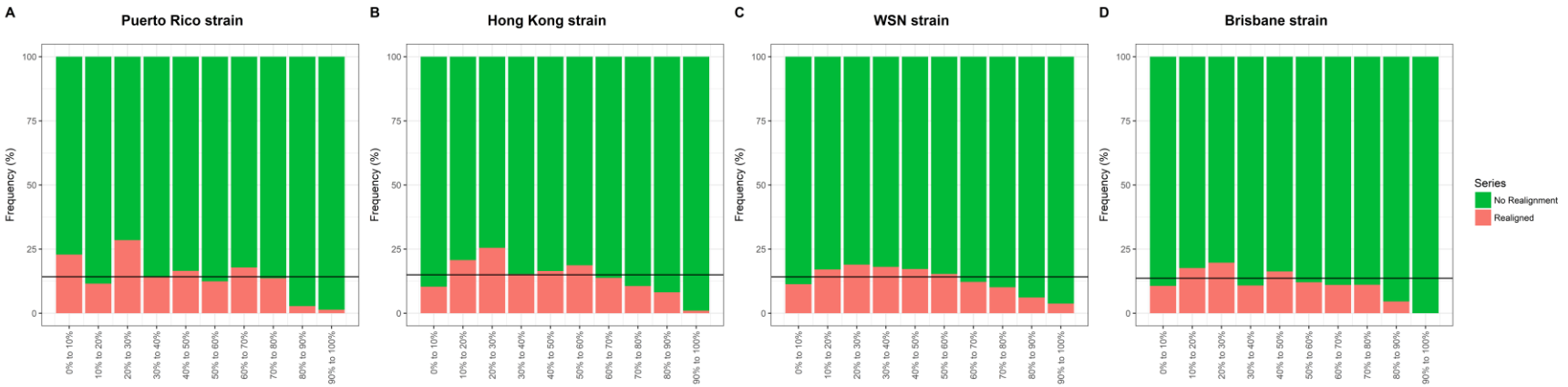
A
Puerto Rico strain
B
Hong Kong strain
C
WSN strain
D
Brisbane strain
Frequency (%)
100
75
50
25
0
0% to 10%
10% to 20%
20% to 30%
30% to 40%
40% to 50%
50% to 60%
60% to 70%
70% to 80%
80% to 90%
90% to 100%
Series
No Realignment
Realigned

Figure 3.10.2. Frequency Distributions of PAR Relative to Guanine and Cytidine Content. The relative distributions of mRNA primers between the Realigned (Red) and No Realignment (Green) subpopulations; the horizontal black line indicates the PAR frequency for that given strain. Red bars above the black line indicate a PAR bias. G+C content is divided into ten percentage increments.

3.11 Effect Modeling

Given the strong associations between prime and realign probability and the length of the primer, the nucleotide directed at 3'U1, the vRNA conserved sequence, and the weak association between PAR and the guanine directed at 3'C2, I next examined the effects of all variable together on this process. Unfortunately transcription of the incorrect purine, which is of particular importance for the Brisbane data set, had to be omitted as both the final transcription of the conserved sequence and exact sequence of the vRNA template were both absent from the retrieved data. I extracted the length (factoring in the complimentary guanine), the nucleotide directed at 3'U1 and the presence of a G directed at 3'C2, and the vRNA conserved sequence, for primers in the No Realignment and Realigned subpopulations. Both a logistic regression and examining the frequency of PAR for the given variables generate the same probability for a primer to undergo PAR on the basis of a given length, sequence end, and conserved sequence. Since the latter requires significantly less computation than the former, I extracted the frequency of PAR for all primers on the basis of length, sequence end, and conserved sequence transcribed. These data did have some noise, as some lengths and sequences ends are poorly represented in the data sets. To reduce noise, a PAR frequency was only included if the data point consisted of at least 1/6400 of the total reads for that data set for all data (1% in each of the 64 possible variable combinations). These thresholds were modified to $1/1200^{th}$ of the data for the 3'U4 data set and to $1/12800^{th}$ of the 3'C4 data set to reflect the differences in PAR events for these templates. Due to the length of primers in the WSN data set not being representative of the actual primer length[9], this data set was omitted from this analysis. PAR frequencies represent the average of Puerto Rico, Hong Kong, and Brisbane strains. These data are in Figure 3.11.1.

PAR is modified by both the length of the primer and the nucleotide sequence used to prime transcription in a manner similar to the overall observed trends. Priming transcription with

an 11 nucleotide long primer ending in AG is associated with an average PAR rate of 10.1%, which is similar to the ~10% PAR rate for a primer of similar sequence end and length used to previously estimate the PAR frequency[138]. Priming with an uridine residue alone remained associated with PAR regardless of the length or template transcribed. Both the vRNA template and the nucleotide residues used to prime transcription changed the impact of length on PAR. As length increased the PAR rates tend to decrease. Overall, a length of 12 nucleotides was less associated with PAR than the average of the three strains (14.3%), except when priming with an uridine residue. For the 3'C4 template a length of at least 12 nucleotides was less associated with PAR; for the 3'U4 template PAR was less associated with a length 13 nucleotides. Priming internally with a guanine residue directed at 3'C2 was related to a higher PAR frequency than priming terminally with the same 3'U1 directed nucleotide for primers 11 nucleotides, or 12 nucleotides if priming with UG, and shorter in length. At a length of 12 nucleotides or more, the ability to prime internally is associated with a reduction in PAR rates in comparison with the same terminally directed nucleotide.

The 3'C4 template showed a reduction in PAR rates save for when priming with an uridine residue alone. Priming terminally with cytidine or guanine residues showed reduced PAR at a length of 11 nucleotides. PAR rates when priming terminally on the 3'U4 template with a cytidine or guanine residue was only partially decreased with length, with the PAR rates remaining close to the strain average of 14.3%. PAR rates lower than 5% were observed 11 times with the 3'U4 template, primarily when priming internally with a primer of at least 13 nucleotides and a 3'U1 directed guanine or adenine. For the 3'C4 template PAR rates of 5% or less were observed 35 times, particularly when priming internally with any residue and a primer 12 nucleotides long.

From these data, it can be concluded that an ideal primer is one that is 13 nucleotides long ending in AG to allow for a more stable Watson-Crick pair at 3'U1 and internal initiation at 3'C2. Increasing the length or transcribing a 3'C4 templated gene will decrease the probability of PAR further.

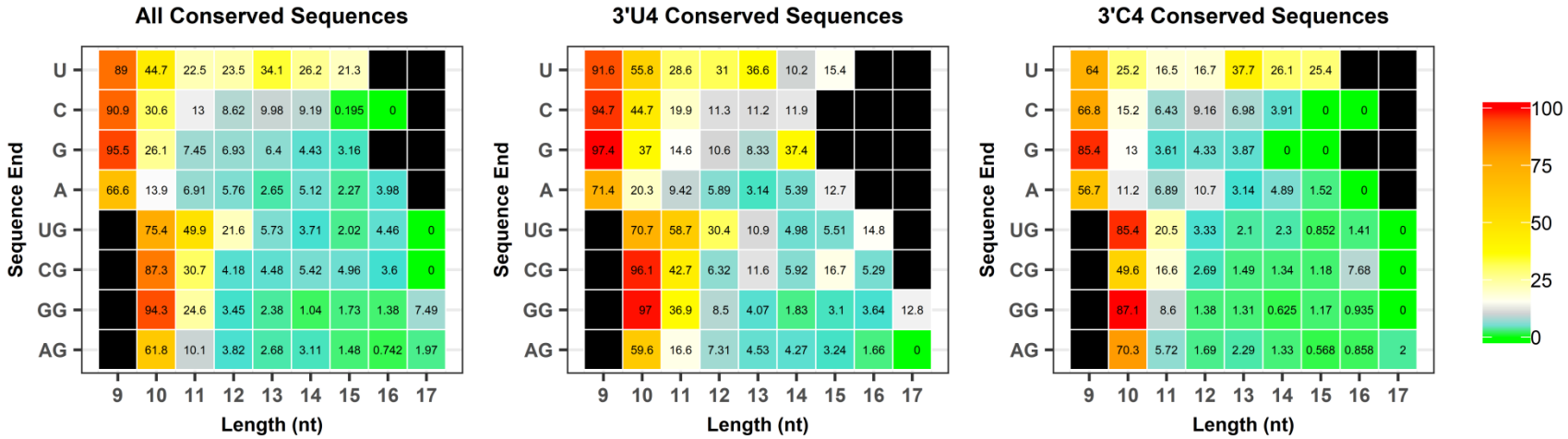
All Conserved Sequences
3'U4 Conserved Sequences
3'C4 Conserved Sequences
Sequence End
U
C
G
A
UG
CG
GG
AG
89 44.7 22.5 23.5 34.1 26.2 21.3
90.9 30.6 13 8.62 9.98 9.19 0.195 0
95.5 26.1 7.45 6.93 6.4 4.43 3.16
66.6 13.9 6.91 5.76 2.65 5.12 2.27 3.98
75.4 49.9 21.6 5.73 3.71 2.02 4.46 0
87.3 30.7 4.18 4.48 5.42 4.96 3.6 0
94.3 24.6 3.45 2.38 1.04 1.73 1.38 7.49
61.8 10.1 3.82 2.68 3.11 1.48 0.742 1.97
91.6 55.8 28.6 31 36.6 10.2 15.4
94.7 44.7 19.9 11.3 11.2 11.9
97.4 37 14.6 10.6 8.33 37.4
71.4 20.3 9.42 5.89 3.14 5.39 12.7
70.7 58.7 30.4 10.9 4.98 5.51 14.8
96.1 42.7 6.32 11.6 5.92 16.7 5.29
97 36.9 8.5 4.07 1.83 3.1 3.64 12.8
59.6 16.6 7.31 4.53 4.27 3.24 1.66 0
64 25.2 16.5 16.7 37.7 26.1 25.4
66.8 15.2 6.43 9.16 6.98 3.91 0 0
85.4 13 3.61 4.33 3.87 0 0
56.7 11.2 6.89 10.7 3.14 4.89 1.52 0
85.4 20.5 3.33 2.1 2.3 0.852 1.41 0
49.6 16.6 2.69 1.49 1.34 1.18 7.68 0
87.1 8.6 1.38 1.31 0.625 1.17 0.935 0
70.3 5.72 1.69 2.29 1.33 0.568 0.858 2
9 10 11 12 13 14 15 16 17
Length (nt)
100
75
50
25
0

Figure 3.11.1. PAR Frequencies Based on Length and Nucleotide Sequence used to Prime Transcription. Left panel shows the average PAR frequency of the Puerto Rico, Hong Kong, and Brisbane data sets. Middle panel shows PAR frequencies for these strains when transcribing 3'U4 templated genes. Right panel shows PAR frequencies for these strains when transcribing 3'C4 templated genes.

4. Discussion

A prime-and-realign process (PAR) occurred at least once for 14.3±0.6% of cap snatched primers. Higher PAR rates are seen when the length of a cap snatched primer is less than 12 nucleotides (Figure 3.4.2 and Figure 3.4.3) and/or the RNA duplex has a higher free energy after three nucleotides have been transcribed (Figure 3.7.1 and Figure 3.9.2). PAR rescues these primers typically by converting the primer into one at least three nucleotides longer ending in an adenine residue (Figure 3.5.1, Figure 3.8.1). This increase in length and sequence compatibility was sufficient to reduce the frequency of a subsequent round of PAR to 4.44±2.47% (Table 3.2.1). This reduction is similar to the 5.76% frequency of a subsequent PAR event after lengthening a nine nucleotide long primer by a GCA addition (Figure 3.11.1). In a singular round, PAR typically addresses both of the factors (length and RNA duplex stability) that are associated with PAR.

4.1 PAR Frequency

This PAR rate of 14.3±0.6% was similar in order of magnitude to the ~10% estimation of PAR frequency obtained from an 11 nucleotide long primer ending in an AG dinucleotide[138]. The PAR rate for a similar primer in these data was 10.1% (Figure 3.11.1), which is consistent with the previously estimated frequency. PAR rates for other viruses, which use a cap-snatching process to prime transcription are generally higher than that of IAV. Rice grassy stunt virus (RGSV) has a PAR frequency of 12.8%[163], 27% for Toscana virus[167], 35.2% for rice stripe virus (RSV)[163], and approximately 70% for Schmallenberg virus[166]. This may be due to differences in the structure or key functional regions of these RdRp, as this is known to affect PAR frequency[138]. Additionally, the free energy through a three nucleotide region of the 3’

conserved sequences of these other viruses is higher (less stable) than the CGU region of the IAV conserved sequences. These data indicate that the IAV RNA-dependent RNA polymerase (RdRp) and/or the IAV conserved sequences may be better able to prevent the factors that cause PAR.

4.2 PAR was not Related to the G+C Content of Cap Snatched mRNA

G+C content in mammalian 5' untranslated regions (UTR) may be related to the breadth of gene expression[214], [215]. PAR did modify the relative G+C of cap snatched primers (Figure 3.10.1, Pre-Realignment versus Post-Realignment subpopulations). Cap snatched primers with a broad range of relative G+C content did not require G+C content modification through PAR prior to transcription (Figure 3.10.1, No Realignment subpopulation). Furthermore, the relative G+C content of these subpopulations had a very high degree of overlap (Figure 3.10.1), and the PAR frequency remained close to the strain PAR frequency across 10% intervals of relative G+C content (Figure 3.10.2). These data suggest that PAR does not occur in response to relative G+C content, and that transcription of the conserved sequences occurred irrespective of relative G+C content.

IAV employs multiple mechanisms to selectively translate IAV mRNA both by disrupting host gene expression and selectively expressing IAV genes. IAV disrupts host gene expression through methods including cap-snatching[1], [7], disruption of enzymatic polyadenylation[35], [79], and retention of polyadenylated mRNA in the nucleus[78], [81]. NS1 mediates selective translation of IAV mRNA in the cytoplasm through interactions with host translation factors and the viral conserved sequences[74]–[77]. Modification of G+C content in the 5' UTR of IAV mRNA through PAR is likely unnecessary. PAR being unrelated to the relative G+C content in the context of mammalian, cells wherein the relative G+C content may

be important[214], [215], makes sense as selective expression can be mediated by other factors[1], [7], [35], [74]–[79], [81].

4.3 Length of Cap Snatched Primers

Compatible sequences are cleaved within a length range determined by the distance from the cap binding region to the endonuclease active site. For IAV this distance has been estimated to be approximately 50 angstroms[130], which is consistent with cleavage of primers between ten and 13 nucleotides[6]. Others have shown that the length distribution of cap snatched primers increases above background at approximately ten nucleotides and peaks at 11 and 12 nucleotides depending on the gene transcribed[1], [6]–[8], [14]. We found that the length of cap snatched primers with the No Realignment subpopulation of primers was consistent with this length estimation, as this subpopulation was clustered around 11.9±1.2 nucleotides (Figure 3.5.1). The length of the cap snatched primers within Realigned subpopulation prior to PAR was found to be clustered around 10.7±1.2 nucleotides (Figure 3.5.1). While there is overlap of these two populations, the Realigned subpopulation is indeed of a shorter length. These data suggest that PAR may occur due to a breakdown in the length selective cleavage process.

Once a primer has been cleaved, it must pass through the product exit tunnel to reach the RdRp active site to prime transcription[130]. The distance from the PB2 cap binding region through the product exit tunnel to the PB1 active site was estimated to be 54 Angstroms (estimated to be 12 to 15 nucleotides[130]); however, RdRp selects for primers of a shorter length range (50 angstroms; estimated to be ten to 13 nucleotides[130]) than is required to reach the active site. In agreement with the 12 nucleotide estimated distance, we observed that a length

of 12 or more nucleotides, including the guanine residue directed at 3'U1, is associated with less frequent PAR than the strain average (Figure 3.4.2 and Figure 3.11.1). Primers shorter than 12 nucleotides are still used to prime transcription. In such cases PAR is favored, with the PAR frequency following an exponential increase for every nucleotide in length decreased (Figure 3.4.4). This length of 12 nucleotides is similar to that proposed as the minimum length to avoid PAR for RSV, RGSV, and Maize Stripe Virus. Each of these viruses tend to undergo PAR more frequently when transcription is primed with a cap snatched primer shorter than 12 nucleotides in length[160]–[162]. Length is therefore likely a key determinant of PAR shared among all Group V cap-snatching viruses, though the optimal length range may be different.

It is currently unknown how these shorter primers are used to prime transcription; however, these data suggest that RdRp has a mechanism to shorten the distance from the PB2 cap binding region to the PB1 active site. Cap snatched primers shorter than nine nucleotides have been observed in full-length IAV mRNA without nucleotide sequence repeats associated with PAR[1], [7], [8]. It is likely that a conformation change of RdRp is involved in this process, which would cause transcription to not be energetically favorable. It is also possible that a unique mechanism is used to prime transcription with these shorter primers. Given the hold of PB2 on the m^7G cap[107], [109], [110], [185], [186] and the flexibility of the PB2 subunit[130] a conformation change of RdRp to a strained conformation to allow for priming of transcription seems more plausible.

Internal initiation, with a complimentary guanine, was less favorable than terminal initiation for primers 11 nucleotides or shorter. This suggests that 3'C2, which had been estimated to be at position -1 of the RdRp active site[110], may not be optimal for priming transcription with these shorter primers. These data suggest that priming transcription with an 11

nucleotide or shorter primer has a preference for terminal initiation and that 3'U1 may occupy position -1 of this conformation. In agreement with this, the PAR rates for primers which initiate internally at a length of 11 or less have a similar PAR frequency to those that initiate terminally at a length of ten nucleotides or less (Figure 3.11.1). Internal initiation at 3'C2 was associated with a lower PAR frequency for primers 12 nucleotide in length or longer. These data indicate that internal initiation, which is favored over terminal initiation, may function properly only when the distance between the cap binding region and active site does not require compression. The structure of the IBV RdRp where 3'C2 was shown to be at the active site was crystallized using a 13 nucleotide long primer ending in 5'-AGC[110], which is the equivalent to a 12 nucleotide long primer ending in a guanine residue. Since a primer of 12 nucleotides is the minimum length within the estimated nucleotide length equivalence of 54 angstroms[130], it may be that this length is the minimum requirement for efficient internal initiation on the vRNA template.

Despite a possible conformation change of RdRp based on the length of the primer, the nucleotide sequence addition associated with PAR occurs irrespective of the length of the primer (Figure 3.5.3). The length of the cap snatched primer and the nucleotide sequence addition length are poorly correlated (Pearson's R = -0.0647±0.182). If there is a conformation change to decrease length to the active site, this change does not affect the ability of RdRp to transcribe three nucleotides (or form four base pairs). These data suggest that RdRp can stabilize the transcription for at least a three nucleotide sequence addition prior to the events that would normally trigger PAR on the basis of length, even in a potentially strained conformation.

PAR frequencies follow an exponential decrease particularly between nine and 13 nucleotides in length (Figure 3.4.4). The PAR frequency of primers between 14 and 17

nucleotides in length plateaus rather than continuing to decrease further (Figure 3.5.2). This 14 nucleotide length switching point is similar to the upper limit of 12-14 nucleotide length estimate by Reich *et al*.,[130] for the 54 angstrom distance. At this point, PAR rates may no longer be driven by length of the cap snatched primer. Length alone does not explain PAR, and there are other key determinants to this process. This is consistent with results obtained for RSV and RGSV as the length of cap snatched primers also was concluded not to be the sole determinant of PAR[160]–[162]. When a cap snatched primer is 11 nucleotides shorter PAR was favored regardless of other determinants, to a magnitude higher than that of other determinants (Figure 3.12.1).

We can conclude that this process is primarily driven by length of the cap snatched primers, with primers shorter than the 12 nucleotide distance from the PB2 cap binding region to the PB1 active site likely requiring PAR. These shorter primers occur through either a breakdown in length selective cleavage and/or the PA endonuclease to PB2 cap binding region leading to cleavage of a population of primers that are shorter than optimal length (50 Angstroms versus 54 Angstroms) [130].

4.4 Free Energy of the RNA Duplex

Others have shown that PAR is related to the nucleotide directed at 3'U1, and the variation in vRNA conserved sequences[1], [7], [9], [138]. These factors modulate the free energy from hydrogen bonds and stacking interactions of the RNA duplex during transcription. Following three nucleotides of transcription, there was a 0.41 kcal/mol difference between transcription that results in PAR and transcription that does not result in PAR (Figure 3.6.1). This difference was larger in the Brisbane data set (0.84 kcal/mol), though the trend remained similar; higher free energy was associated with primers that underwent PAR following an addition of three

nucleotides. These data suggest that the free energy from the 3'U1 directed nucleotide, and the variation at position four of the conserved sequence may be important in determining the probability of PAR for a given primer.

4.5 Free Energy and Sequence Specific Cleavage

The overall range of the free energy difference from the nucleotide directed at 3'U1 is 0.53 kcal/mol. This energy difference corresponds with the 0.41 kcal/mol free energy difference between transcription which did or did not result in PAR (Figure 3.6.1). A 3'U1 directed adenine residue (-1.73 kcal/mol) was associated with a reduction in PAR to below the strain average when it was ten nucleotides or more 3' of the 5' cap (Figure 3.11.1A); PAR on the basis of length alone is not reduced below this average for primers less than 12 nucleotides in length (Figure 3.5.2). Primers initiating terminally with an uridine residue remained associated with more PAR than the strain PAR frequency regardless of the primer length (Figure 3.12.1A). These results are in agreement with previous *in vitro* results, where PAR was shown to be related to the identity of the nucleotide directed at 3'U1; primers ending in an adenine residue were less likely to undergo PAR than those ending in an uridine residue[138].

The observed PAR frequency for each 3'U1 directed nucleotide was more related to the free energy from hydrogen bonds and stacking interactions than the ability to form a cis Watson Crick/Watson Crick base pair. Base pairing may occur on any of the 3 sides of the roughly triangular nucleotides, indicating 12 possible combinations[216]. Typically, cis Watson Crick/Watson Crick pairs are associated with elongation reactions (transcription) as they allow for proper positioning of the 3' end of the growing strand for the incoming nucleotide triphosphate to be ligated. It is possible that RdRp may allow for a non cis Watson Crick/Watson

Crick base pair at 3'U1, for which the affinity of the paired nucleotide is similar to that of the free energy values. The structure of the base pair at 3'U1 has not been resolved in recent crystal structures (only the phosphate of 3'U1 in the Flu B primer bound structure has electron density)[110], and it is possible the orientation of this nucleotide in the active site is such that base pairs other than the cis Watson Crick/Watson Crick base pair are possible. The identity of the 3' most nucleotide as an uridine residue appears to be important to this flexibility. Mutation of 3'U1 to 3'A1 allowed for priming transcription with a primer ending in an uridine residue, however, priming with a primer ending in an adenine residue lead to a large decrease in transcription[138]. It is tempting to speculate that the isostericity of an AA pair is less compatible with transcription while the pairs with the uridine residue remain compatible with transcription. The isostericity of the UU pair would also likely be somewhat different, as, while compatible with transcription[138], it remains associated with PAR regardless of primer length (Figure 3.11.1). Despite the possibility of base pairs other than cis Watson Crick/Watson Crick base pairs at 3'U1, the PAR frequency is well correlated to the free energy of the nucleotide directed at 3'U1 ($R^2 = 0.915 \pm 0.059$; Figure 3.9.2). This indicates that this PAR is likely related to the free energy from hydrogen bonds and stacking interactions at 3'U1 rather than the identity of the base pair at 3'U1.

Given the correlation between free energy and PAR rates and the previously demonstrated selective cleavage[4]–[6], it could be expected that RdRp would attempt to select for primers with an adenine to direct at 3'U1. It is surprising that an adenine at position -1 did not occur as frequently as a cytidine at position -2 or a guanine at position +1 (Figure 3.3.1). The minimum cleavage preference is 3' of a guanine residue that can be directed at 3'C2 to prime transcription internally rather than terminally[6], [110], [138]. PA does preferentially cleave

primers at a sequence with multiple bases of compatibility[4], [5], with this selective cleavage being able to at least partially overcome abundance. These data did show an enrichment of nucleotide residues at positions -2 through +1 of the cap snatched primer that could not be explained by random selection or the abundance of these nucleotide residues within the genome (Table 3.3.1, Table 3.3.2, Table 3.3.3, and Table 3.3.4).

A cytidine residue was enriched at the 5' dangling end position (nucleotide residue at -2) in 48.0±9.5% of all reads (48.3±9.1% and 45.8±12.2% of the No Realignment and Realigned subpopulations respectively; Table 3.3.2). This enrichment of cytidine was selective and non-random for all four strains, with the cytidine composition of the cap snatched genome being 27.1±1.5% (i.e., cytidine had a 27.1±1.5% abundance in the cap snatched genome). It is unknown what role, if any a cytidine plays at this position in the cap-snatching process. It is unlikely that this cytidine is selected for on the basis of free energy when priming transcription. A cytidine 5' dangling end of an A-U or G-U pair only decreases the free energy of the duplex by -0.3 kcal/mol[124]. Guanine would be better from an energy perspective (-0.4 kcal/mol) at this position, with adenine being the same (-0.3 kcal/mol), and uridine being worse (-0.2kcal/mol) [124]. Furthermore, there was no association between PAR and the presence of this cytidine residue. The 13.5±1.4% PAR frequency when this cytidine residue was present was similar to the strain average of 14.3±0.6%. It is possible that this cytidine may interact with the PD-(D/E)XK endonuclease to direct the localization of the cleavage site one to two nucleotides 3' of this cytosine.

From a free energy perspective, cleavage of primers so that an adenine residue could be directed at 3'U1 should be the preference of the endonuclease. As seen in these data and the results of others[6], [7], the preference of the endonuclease is to cleave transcripts 3' of a

guanine residue (Figure 3.3.1, and Table 3.3.1). Primers cleaved 3' of a guanine residue may prime transcription internally at 3'C2 rather than terminally at 3'U1[6], [110], [138]. The cytidine at position two of the vRNA has been estimated to occupy position -1 of the RdRp active site to allow for internal initiation[110]. This process has been previously shown to only occur when the cap snatched primer ends in a residue that can base pair with the penultimate vRNA nucleotide[138]. Since a guanine directed at 3'C2 is indistinguishable from a guanine transcribed at 3'C2, we extracted the presence of this guanine from the reference genome as the nucleotide 3' of the end of the matched sequence (Figure 3.3.1). Overall, 47.4±4.6% of cap snatched primers had a guanine residue immediately 3' of the last matched nucleotide (the one directed at 3'U1) of the cap snatched primer (Figure 3.3.1). This enrichment is indeed non-random (Table 3.3.1), and likely relates to the sequence specific cleavage of transcripts 3' of a guanine residue previously reported[6]. We also see that this enrichment is higher in the No Realignment (49.0±4.8%) than the Realigned (38.0±4.7%) subpopulation, and a breakdown in cleavage 3' of a guanine residue may be related to PAR.

Cleavage 3' of a guanine residue may not always result in internal initiation at 3'C2, and this guanine may prime transcription terminally at 3'U1[138]. Terminal initiation with a guanine residue, without a guanine residue that may have been directed at 3'C2, occurred for ~41% and ~60% of all G-U initiation pairs in the No Realignment and Realigned subpopulations (respectively). This indicates that cleavage 3' of a G residue may have been 12.4±4% and 16.9±5.8% more frequent in the No Realignment and Realigned subpopulations. This would increase the frequency of cleavage 3' of a guanine residue to 61.4±3.5% and 54.9±2.1% respectively. This suggests that primers ending in a guanine residue failed to prime transcription internally ~20% or ~31% of the time for the No Realignment and Realigned subpopulations

respectively. When this internal initiation process fails, PAR occurs more frequently than the strain average (Figure 3.9.3).

Internal initiation with a guanine residue is associated with a modest reduction (0.802±0.078 fold; 11.4±1.3%) in PAR (Figure 3.3.1, and Figure 3.9.3); however, internal initiation was more associated with PAR than priming terminally when the primer was 11 nucleotides in length or less (Figure 3.11.1). For primers 12 nucleotides or longer, priming internally was associated with less frequent PAR than priming terminally (Figure 3.11.1); simply internal initiation with a complimentary guanine was rescued by length. A previous *in vitro* analysis of PAR concluded that the 3'C2 directed guanine residue was unrelated to PAR[138]; however, the primers analyzed were 11 nucleotides in length – the length at which the complimentary guanine does not yet have an effect on PAR rates. For primers 12 nucleotides or longer, PAR rates are much lower when priming internally with a guanine residue directed at 3'C2, than terminally for all residues. The combination of length and the complimentary guanine residue was also capable of rescuing primers with an uridine directed at 3'U1, which the combination of length and 3'C4 template cannot achieve (Figure 3.11.1).

These data suggest that the presence of this complimentary guanine affects the free energy in a way that was not directly evident in these free energy calculations. Others have suggested that there is a free energy benefit of priming transcription internally rather than terminally[137]. The decreased PAR frequency seen between priming terminally and internally (Figure 3.9.3) is consistent with a reduction in free energy of 0.13 kcal/mol. Simply, on average, priming internally is 0.13 kcal/mol more stable than priming terminally with the equivalent nucleotide residue. These data suggest that this complimentary guanine residue was associated with a free energy reduction that was not apparent in the mathematically obtained value. It is also

possible that internal initiation may afford RdRp the option of ignoring 3'U1. Ignoring 3'U1 this seems unlikely as PAR rates do still followed the trend of the free energy of the 3'U1 directed nucleotide, with some noise (Figure 3.11.1). Regardless, priming internally was of benefit, likely due to a decrease in free energy, which was not directly visible in the mathematically obtained value.

We also see selection of nucleotides at position -1 (directed at 3'U1; Figure 3.3.1, Table 3.3.3 and Table 3.3.4). Overall, an adenine residue was selected for at position -1 (33.9±6.1%). This enrichment of adenine could not be explained by the 21.0±1.3% abundance of adenine within the cap snatched genome. An uridine residue was also selected against at this position 12.2±4.0% despite uridine having an abundance of 21.3±1.6% within the genome. Both guanine (29.9±4.5%) and cytidine (24.0±7.2%) occurred at -1 at a frequency similar to that of these residues within the genome. Simply, the selection of the nucleotide at position -1, while unable to fully overcome the lower abundance of adenine within the cap snatched genome, selects nucleotide residues on the basis of their free energy. This trend is not observed in the Realigned subpopulation; uridine residues at position -1 occurred more frequently, and adenine residues occurred less frequently (Table 3.3.3 and Table 3.3.4). While this trend is not observed in any individual strain, the average abundance amongst all four strains of each possible 3'U1 directed nucleotide residue was well correlated to the free energy of that nucleotide when directed at 3'U1 (R^2 = 0.963; Figure 3.9.1). The lack of correlation on the individual strain level likely arises due to differences in the abundance of transcripts that could serve as cap donors. The transcripts cap snatched by RSV and RGSV differed considerable between infected plants, with 45% and 49% of the cap snatched transcripts occurring in less than three data sets[163]. It is

likely that the differential abundance of genes may have led to the loss of correlation between free energy and nucleotide enrichment seen on the strain level.

IAV has been shown to preferentially cap snatch primers with a CAG motif within the compatible length range to form a forming a $\begin{matrix} CAG \\ UCG \end{matrix}$ RNA duplex when priming transcription[4], [5], [7]. This preference is observed in all four data sets; primers ending in CAG comprised 9.6±0.9% of all cap snatched primers, or 10.7±1.2% of the No Realignment and 2.66±1.17% of the Realigned subpopulation (Table 3.3.6). CAG was the most enriched trinucleotide sequence motif seen in the data set, and occurred more frequently than the abundance in the cap snatched genome (2.41±0.15%). This ~4.0 fold enrichment likely indicated a clear preference for cleavage at this sequence, which is consistent with what was previously reported[4], [5], [7]. Overall, we see that when cleavage preferences are met there was less PAR; both AG and CAG showed less frequent PAR (5.47±2.44% and 4.00±1.85%; Table 3.3.5 and Table 3.3.6) than the strain average of 14.3±0.6%. Cleavage of a CAG or AG would result in the highest complementarity with the vRNA, and this complementarity resulted in a decreased PAR frequency.

RdRp will deviate from cap-snatching on the basis of abundance for primers with a compatible nucleotide sequence within the compatible length range[4], [5], [7]. Due to both the subtractive nature of the cap-snatching process and the association with paused Pol II[2], a primer with a compatible nucleotide sequence, within the length range, may not always be available to RdRp. Association of RdRp with Pol II transcribing a pre-mRNA lacking a compatible sequence within the optimal length range may result in cap-snatching of a primer fated for PAR. Abundance versus sequence specificity has not been examined for IAV, though this process has been examined for Sin Nombre virus (SNV). For SNV, multiple nucleotides of complementarity increased the enrichment of a cap snatched primer; however if this primer had

low abundance, it would not be enriched in the cap snatched pool despite the complementarity[172]–[174]. Similarly, we saw an enrichment of cap snatched primers with a compatible nucleotide sequence that exceeds abundance (Figure 3.3.1, Table 3.3.1, Table 3.3.3, Table 3.3.5, and Table 3.3.6); however, low compatibility sequences are also cap snatched. Deviation from the strain preference in both the nucleotide directed at 3'U1 or 3'C2 result in an increase in PAR. These data suggest that PAR was related to a breakdown in sequence specific cleavage. 15.0±2.5% of all cap snatched primers ended in an AG dinucleotide (Table 3.3.5), 18.9±3.7% ended in an adenine residue, and 32.5±5% ended in a guanine residue that was likely directed at 3'C2. In total, 66.4±6.9% of all cap snatched primers had at least one compatible nucleotide with the vRNA ultimate or penultimate 3'-UC residues. Only 33.6±6.9% of cap snatched primers were cleaved without either an adenine to direct at 3'U1 or a guanine to direct at 3'C2; however, 50.4±8.2% of the primers within the Realigned subpopulation lacked both residues. There was also increased frequency of uridine and cytidine directed at 3'U1, and less cleavage 3' of a guanine residue in the Realigned subpopulation – all of which are associated with an increase in PAR. If a primer is not cleaved at a compatible sequence with the vRNA PAR is more likely to occur. RdRp guards against, given the selection against both uridine and cytidine at position -1, which would have been directed at 3'U1. When sequence specific cleavage failed, RdRp can rescue these primers through PAR, which generates a primer ending typically ending in an adenine residue (Figure 3.5.1 and Figure 3.8.1) to be directed at 3'U1. Due to the limited number of transcripts available for cap-snatching, rescuing the cap snatched primer is likely preferable to releasing the primer and finding a primer with a more compatible nucleotide sequence. Association of RdRp with Pol II transcribing a pre-mRNA lacking a compatible sequence within the optimal length range may result in cap-snatching of a primer

fated for PAR; however, PAR affords RdRp the ability to utilize poor primers that are cap snatched due to abundance despite low sequence compatibility.

4.6 Free Energy and the vRNA Conserved Sequences

IAV genes are similar at the 3' end through the first 12 nucleotides, with one natural variation at the 4th nucleotide[123]. These conserved sequences are 3'-UCG**[U/C]**UUUCGUCC-5', with the cytidine residue at position 3'Y4 being present in the PB2, PB1, and (usually) the PA genes. Others have shown that 3'C4 templated genes undergoing less PAR[1], [7], [9], [138]. The data examined herein were consistent with previous findings (Figure 3.7.1). mRNA transcribed from 3'U4 genes underwent PAR 1.43±0.13 fold more frequently (20.3±1.1%), whereas those transcribed from 3'C4 genes underwent PAR 0.480±0.239 fold less frequently (6.77±3.10%) than the strain average (14.3±0.6%). This difference was associated with the residue at 3'Y4, as transcription of PA genes underwent PAR 1.04 fold more frequently in the Hong Kong data set wherein the PA gene possess the 3'U4 conserved sequence. mRNA transcribed from the PA gene underwent PAR 0.74±0.35 fold less frequently than the strain average for the other three strains. An alternative possibility for the difference in PAR frequencies between subunits was the differences in free energy in the 3' 5' base pairing region (Table 3.7.1). While the failure of this region to separate may lead to PAR, the differences in the free energy of the 3' 5' base pair region does not appear to be cause of the differences observed in PAR rates among the gene segments. PAR frequency is not correlated with the energy differences in the 3' 5' base pair region (R^2 = 0.043), even when factoring in the differences at 3'Y4 (R^2 = 0.020), or the ratio between the 3' 5' region and the energy differences at 3'Y4 (R^2 = 0.138). If this region did indeed lead to PAR, either the M or PA segment of the Hong Kong strain should have the lowest PAR frequency. Both the M and PA segments of the Hong Kong strain had a PAR frequency

similar to the strain average (Figure 3.7.1), which was consistent with the PAR frequency on the basis of the conserved sequences transcribed. These data suggest that indicates that the differences in PAR frequencies were conferred by the conserved sequence transcribed.

The sole difference between the two conserved sequences is the nucleotide at position four. When transcribing a GCA addition on the 3'U4 template the free energy was -5.09 kcal/mol, and when transcribing a GCG addition on the 3'C4 template the free energy was -5.71 kcal/mol. This resulted in a net difference of 0.62 kcal/mol between these conserved sequences which is greater than 0.41 kcal/mol free energy difference between transcription which resulted in PAR or complete transcription (Figure 3.7.1). Figure 3.11.1 show the effects of the vRNA conserved sequence in relation to length and the nucleotide sequence used to prime transcription. A PAR rate less than 5% (an approximate three fold reduction in PAR) was observed 35 times for the 3'C4 template, yet only 11 times for the 3'U4 template. Simply transcription of a gene with the 3'C4 template results in less PAR once a minimum length of 11 nucleotides has been achieved. The combination of the 3'C4 template and a ten nucleotide long primer ending in an adenine residue was associated with a reduction in PAR below the strain average; PAR on the basis of length alone is not rescued until a length of 12 nucleotides. The 3'C4 template also rescues terminal initiation with cytidine and guanine residues (Figure 3.12.1B versus 3.12.1C). Priming terminally with a 12 nucleotide or longer primer ending in a guanine or cytidine residue on the 3'U4 template resulted in a PAR rate which dropped below, but remained near, the strain PAR frequency. Priming terminally with a primer of 11 nucleotides or longer ending in a cytidine or guanine residue on the 3'C4 template was associated with a reduction well below the strain PAR frequency (14.3±0.6%). Transcribing 3'C4 templated genes versus 3'U4 genes allows for more flexibility in both length and priming nucleotide sequence without requiring PAR.

PAR primarily consists of GCA containing (GCA to GCAAAA) rather than GCG containing (GCG to GCGAAA) additions (Figure 3.5.1). GCA containing additions were transcribed 81.7±4.5% that a primer underwent PAR, with 61.6±1.2% being a GCA nucleotide sequence addition. For all strains, the frequency of different nucleotide sequences added by PAR was similar in the 3'U4 data set, and the overall data set (Figure 3.8.1, All vs. 3'U4). For all strains except the Brisbane strain, transcription of the 3'C4 conserved sequences changed the frequency of the nucleotide sequence additions. For the Brisbane strain, PAR remained associated with a GCA containing addition rather than a GCG containing addition when transcribing the 3'C4 conserved sequences. When three nucleotides have been transcribed (a GCA addition), the AC mismatch pair results in a reduction of free energy from -5.71 kcal/mol (GC pair) to -4.84 kcal/mol (AC mismatch). This 0.87 kcal/mol difference is consistent with the 0.84 kcal/mol free energy difference observed for the Brisbane strain between primers which underwent PAR and those which did not following three nucleotides of transcription (Figure 3.6.1). For the other three strains, PAR typically occurred either prior to the transcription of the 3rd nucleotide (G or GC addition) or after an adenine residue had been transcribed (GCGA or GCA; Figure 3.8.1). When PAR occurs prior to transcription at 3'C4 (a $\frac{GC}{CGC}$ or $\frac{G}{CGC}$ duplex) the conserved sequence transcribed does not affect the free energy. When three or more nucleotides are transcribed on the 3'C4 template, PAR will occur either after four or more nucleotides have been transcribed (GCGA addition or longer) or after transcribing an AC mismatch (GCA addition), rather than occurring after a GCG addition. PAR adds a nucleotide sequence ending in adenine residue such that the subsequent round of transcription will be less likely to undergo PAR due to the decrease in free energy.

It is possible that the increased frequency of GC additions seen with PAR that occurred when transcribing the 3'C4 template were GCG additions wherein transcription was primed at 3'C2 following PAR. At a length of 12 nucleotides or more (a nine nucleotide primer undergoing PAR following three nucleotides of transcription), priming transcription with a $\frac{CG}{UCGC}$ duplex had a lower PAR frequency than priming transcription terminally with an adenine or guanine residue (Figure 3.11.1). This difference is conferred by the 3'C4 template, as terminal initiation with an adenine residue on the 3'U4 template had a lower PAR frequency than priming transcription internally with a $\frac{CG}{UCGU}$ duplex (Figure 3.11.1). The 3'C4 template confers a 0.62 kcal/mol free energy decrease versus the 3'U4 template, which is larger than the 0.33 kcal/mol difference between priming with an adenine versus a cytidine at 3'U1. A subsequent round of PAR is therefore unlikely, as the length of the primer has been addressed by the nucleotide sequence addition, with the template itself addressing the free energy.

Transcription of the incorrect purine at 3'Y4 occurred with PAR additions when transcribing both 3'U4 (82.3±15.1% GCA, 2.75±2.14% GCG) and the 3'C4 (32.9±14.1% GCG, 7.94±3.01% GCG; omitting Brisbane) templated genes. The mutation rate for IAV transcription has been estimated between 10^{-4} to 10^{-6} substitutions per nucleotide per strand copied[16]–[18]. This suggests that transcription of the G-U wobble pair and the A-C mismatch may occur more frequently than expected even when factoring in the PAR frequencies of 20.3±1.1% and 6.77±3.10% when transcribing the 3'U4 and 3'C4 templates (Figure 3.7.1). It is important to note that either adenine or guanine at this position are both recognized as IAV due to the natural variation at position 3'Y4, so these additions, while incorrect on the basis of the template, are not true transcription errors. It is also possible that what appears to be an A-C mismatch or a G-U wobble pair is actually correct transcription of the conserved sequence. The nucleotide sequence

of the conserved sequence transcribed is not available in the data set, and it is possible that a subpopulation of each gene has mutated to acquire the other conserved sequence. If these are indeed G-U wobble pairs and AC mismatches, then the free energy would be increased (Table 3.8.1), and PAR would be more likely on the basis of free energy.

Another possibility is that G-U wobble pairs are allowed for when transcribing from 3'U4-U7. A PAR transcribed nucleotide sequence ending in a guanine residue (ex., GCG, GCAG) may allow for internal initiation at 3'C2 following PAR. If, following PAR, RdRp returns to the conformation proposed by Reich *et al*.,[110], 3'C2 would occupy position -1 allowing for priming to occur at this nucleotide. In order to prime terminally, as the current PAR mode suggests, the vRNA would have to be shifted to allow for 3'U1 to be at the -1 position. There is some evidence for this in these data, as PAR events consisting of a single guanine addition often occur following a previous PAR event. What we classify as a separate PAR event may in fact be G-U wobble pairs transcribed in this uridine track to allow for internal initiation following PAR. Furthermore, when transcribing the 3'C4 conserved sequences PAR events of GCG are infrequent; however, GC additions are more frequent on the 3'C4 template. These GC additions when transcribing the 3'C4 template may be GCG additions wherein the ultimate G residue was directed at 3'C2 to prime transcription following PAR. IAV is one of few cap-snatching viruses that lacks conserved sequences with a dinucleotide or trinucleotide repeat on the 3' end. For example, *Nairoviridae* has a 3'-AGAGUUUCU sequence, *Phenuiviridae* has a 3'-UGUGUUUC conserved sequence, and *Hantaviridae* has a 3-AUCAUCAUCUG conserved sequence[133], and IBV has a 3'-UCGUCUU conserved sequence[217]. For IAV, transcribing a GCAG nucleotide sequence, with a G-U wobble pair at 3'U5 would allow for internal initiation

following PAR. This would allow for IAV to have a pseudo nucleotide repeat sequence for priming following PAR with two nucleotide of compatibility rather than one.

It is tempting to speculate that the GCA addition might be transcribed when PAR will occur rather than PAR occurring due to the GCA addition. The factors leading to PAR may lead to transient misalignment of the primer and the template, which has been shown to increase transcription errors, including errors in which a nucleotide is skipped[218], [219]. In such cases, the cytidine residue at position four might be skipped, and an adenine residue may be paired with 3'U5 of the conserved sequences. Transcription of the AC mismatch or a longer nucleotide sequence addition (ex., GCGA) prior to PAR would allow for an adenine residue to be directed at 3'U1 decreasing the free energy on the subsequent round of transcription.

Allowing for this AC mismatch may be important late in infection, as the Brisbane data set, collected 12 hours post infection[8], had a very high frequency of AC mismatches. The relation of this time period to the AC mismatch is unknown, however, there is a reduction of both guanine and cytidine in the cap snatched genome for the Brisbane strain (Table 3.3.1, Table 3.3.2, and Figure 3.10.1). It is possible that guanine has a lower availability later in infection (four hours versus 12 hours), and nucleotide availability has been shown to affect PAR[14]. It is also possible that mutations of 3'C4 to 3'U4 may have been allowed to accumulate for the PB2, PB1, and PA subunits due to the longer infection time prior to the harvesting of mRNA (12 hours[8]). This mutated 3'U4 templated subpopulation would likely undergo more PAR (Figure 3.7.1) than the 3'C4 templated subpopulation, leading to an increase in the observation of GCA additions. The 3'U4 template has also been shown to be more transcriptionally active in regards to mRNA synthesis (the 3'C4 template is more active in replication)[220], which would further contribute to the observation of GCA additions on what should be a 3'C4 template. Simply, the

3’U4 mutants of PB2, PB1, and PA would transcribe more mRNA, which would undergo more frequent PAR (GCA additions), leading to an increase in the observation of GCA additions.

Generally, as free energy decreases, the probability of PAR decreases. Priming transcription with an adenine residue (-1.73 kcal/mol) of a 3’C4 templated gene (-5.71 kcal/mol), was capable of reducing the PAR frequency below the strain average at ten nucleotides of length (11.2% versus 14.3±0.6%). While length still drives this process, increased stability of the RNA duplex can partially compensate for a shorter primer in the ten to 11 nucleotide range. Furthermore, priming transcription terminally with a uridine, cytidine, or guanine residue on the 3’U4 template, retains a higher probability of PAR even with increasing length. Internal initiation with a guanine residue directed at 3’C2 in combination with a length of at least 12 (3’C4) or 13 (3’U4) nucleotides is capable of reducing the PAR rate below the strain average for primers with a 3’U1 directed uridine residue. This complimentary guanine residue likely modifies the free energy in a manner that is not visible in calculation, as it reduces the PAR frequencies of the 3’U1 directed nucleotide in a manner consistent with a -13 kcal/mol change. Together with length these factors that modify the free energy also modify the probability of PAR.

4.7 The Prime and Realign Process

PAR generally occurred following the addition of three nucleotides (or two nucleotides if priming internally) 64.8±3.9% of the time. This nucleotide sequence addition length is not influenced by the length of the primer (Figure 3.5.1 and Figure 3.5.3). These data suggest that RdRp is capable of stabilizing transcription through a four nucleotide long range, even if other factors are sufficient to cause the polymerase to stutter. The active site of IBV has been estimated to allow for approximately three base pairs[110], though my data suggests that it may

allow for four base pairs. Further evidence that the active site may allow for four base pairs comes from an examination of polymerase stuttering on the polyuridine track. This process requires the uridine track to be at least five nucleotides in length[129]; a length of four nucleotides had 4% polyadenylation activity, and a length of three nucleotides had no activity[129]. This would suggest that the RdRp active site can accommodate three or four base pairs (one at which to prime transcription, and then transcription of three nucleotides). Another source of evidence is polymerase stuttering on this uridine track during replication. Insertions of three to four nucleotide long sequences (UCUU and UCU) from polymerase stuttering were observed[128] on NA genes with a shortened 5' gene specific region (this shortening allows for insertion mutations to accumulate via polymerase stuttering[99], [128]). These taken together with our results that PAR followed the addition of three nucleotides (four base pairs) 64.8±3.9% of the time, it can be concluded that RdRp may stabilize four base pairs prior to polymerase stuttering. This suggest that all factors affecting PAR, namely the length of the primer and free energy of the RNA duplex, are evaluated following a three nucleotide sequence addition. If the sum affects of these factors falls below a given threshold, PAR likely occurs (Figure 4.1).

This threshold is likely calculated on the basis of the summation of free energy from the RNA duplex, conformational strain, and other affecting factors. If the distance from the PB2 cap binding region to the PB1 active site was compressed to allow for priming with a shorter (<12 nucleotide) primer, this conformational strain would be at least partially applied to the RNA duplex. Simply, the primer would be pulled toward the PB2 cap binding region in a manner similar to compressing a spring, with this compression being proportional to the length of the primer. If this primer is too short (nine nucleotides or less) and/or the free energy of the RNA duplex is too high (such as when priming terminally with a uridine, priming terminally on the

3’U4 template, or a GCA addition to the 3’C4 template), PAR will likely occur. If the free energy is sufficiently low, such as when priming transcription with an adenine on the 3’C4 template, the effects of short length (ten nucleotides to the 3’U1 directed residue) can be overcome. As length increases to 12 or more nucleotides, the free energy requirement to avoid PAR increases, and most nucleotide sequences used to prime transcription become less associated with PAR than the strain average. Free energy does still affect PAR frequency for primers 12 nucleotides or longer as priming internally with a guanine residue directed at 3’C2, with an adenine residue directed at 3’U1, and/or transcribing the 3’C4 template all result in a decreased PAR frequency for the given length of primer. These factors act in addition to the stabilizing affects of RdRp, likely through the priming loop[137]–[139], to mitigate the frequency of PAR.

Despite the high frequency of a three nucleotides sequence addition, RdRp can transcribe further down the conserved sequences prior to a polymerase stuttering event. These data suggest that RdRp may defer the application of this threshold; however the frequency of doing so decreases exponentially with each additional nucleotide beyond three transcribed (R^2 = 0.962±0.040; Figure 3.5.2). It is possible that the polymerase may be able to over track down the template while stabilizing the duplex and transcribe additional nucleotides prior to the evaluation of factors that cause PAR. Nucleotide residues up to 3’U7 are not tightly held by RdRp. 3’6-UUCG-9 interact with the polymerase via their phosphate backbone[23]. 3’C8 and 3’G9 are involved in promoter secondary structure, as there is a sharp turn between them[23], and they may not be able to be pulled into the active site without breaking of the 3’ 5’ base pair region. It may be possible for nucleotides up to 3’U7 to enter the active site without breaking the 3’ 5’ base pair region, though conformational strain to allow for this is likely. This exponential decrease in

addition frequency may be indicative of stretching of the active site to accommodate more base pairs. This subpopulation that undergoes PAR after more than four nucleotides was transcribed was also of an intermediate free energy (Figure 3.6.1). It is possible that this modest decrease in free energy may allow for more transcription prior to PAR in a similar manner as to how decreased free energy allows for transcription in the absence of PAR despite a shorter primer length. Regardless of the mechanism of these additions, PAR becomes exponentially less likely to occur after three nucleotides have been transcribed.

The combination of the dimensions of the active site and the limited length of the conserved sequence which may enter the active site force the factors that contribute to typically be evaluated after three nucleotides of transcription. At this point the 3' 5' base pair region must be broken, and the primer and the template must split and exit through separate channels[23].

It has been hypothesized that RdRp uses free energy from the template product duplex to offset the free energy from the 3' 5' base pairing region, with this region serving as a barrier to further transcription[110]. While failure of the 3' 5' base pair region to melt may promote PAR, the PAR frequency is not related to the difference in free energy of the 3' 5' base pair region (Table 3.7.1). While free energy from the RNA duplex may be important in breaking the 3' 5' base pair region, it is not the sole determinant of this process. That said, failure of the 3' 5' base pair region to separate would lead to conformational strain on the RNA duplex during transcription, and may result in PAR. While the hold on the 3' end through base pairing is weaker, this hold would be similar to the tight hold on the 5' end which forces polyadenylation of mRNA[125]. Loss of this tight 5' end hold by mutation of key residues results in a reduction or loss of polyadenylation[23], [127], [147]. Over all this energy threshold likely increases when RdRp is forced to adopt a strained conformation, such as when priming transcription with a cap

snatched primer shorter than 12 nucleotides in length, however this threshold is not the free energy of the 3' 5' base pair region duplex.

The other possible explanation could be a failure of the template product duplex to separate and exit through the respective product and template exit tunnels[140]. A similar polymerase architecture with the template and product exit tunnels has been observed for other cap-snatching Group V viruses[23], [132]–[134], [193]. Members of *Arenaviridae* and Bunyavirales appear to lack a 3' 5' base pairing region in their RdRp[132], [221], despite these viruses requiring a region capable of forming base pairs for transcription and replication[222]–[224]. It has been proposed for such viruses that PAR may be related to failure of the template product duplex to separate[167]. The vRNA may follow the cap snatched primer into the template exit tunnel. Given the RdRp active site appears to be able to accommodate at least three base pairs, PAR may occur at this point as the template and product would need to be separated for more transcription to take place. In such cases the RNA duplex may be broken without net movement of the vRNA through the RdRp active site and out of the template exit tunnel, and the 3' end of the vRNA would return to the active site.

Regardless of the mechanism, once the factors leading to PAR cause the template to dissociate with the product, RdRp would likely return to a lower energy state. For the polymerase, that would involve the re-association of the 3' 5' base pair region, and returning of 3'U1 or 3'C2 to position -1 of the active site. For the primer, this would result in it moving down the template exit tunnel, with the 3' most nucleotide residue being used to once again prime transcription at 3'U1 of the vRNA.

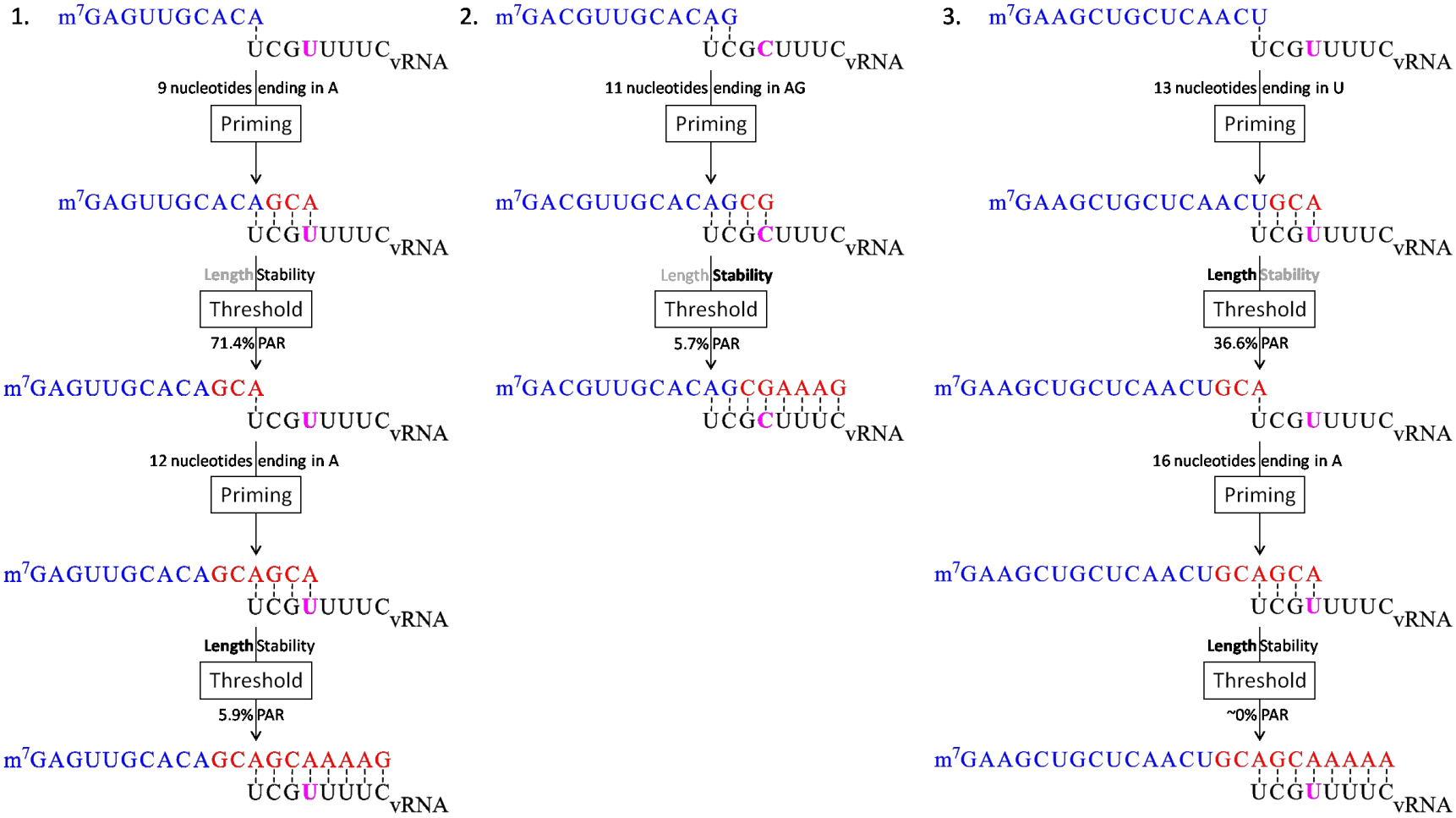
1.
m7GAGUUGCACA
UCGUUUUC vRNA
9 nucleotides ending in A
Priming
m7GAGUUGCACAGCA
UCGUUUUC vRNA
Length Stability
Threshold
71.4% PAR
m7GAGUUGCACAGCA
UCGUUUUC vRNA
12 nucleotides ending in A
Priming
m7GAGUUGCACAGCAGCA
UCGUUUUC vRNA
Length Stability
Threshold
5.9% PAR
m7GAGUUGCACAGCAGCAAAAG
UCGUUUUC vRNA
2.
m7GACGUUGCACAG
UCGCUUUC vRNA
11 nucleotides ending in AG
Priming
m7GACGUUGCACAGCG
UCGCUUUC vRNA
Length Stability
Threshold
5.7% PAR
m7GACGUUGCACAGCGAAAG
UCGCUUUC vRNA
3.
m7GAAGCUGCUCAACU
UCGUUUUC vRNA
13 nucleotides ending in U
Priming
m7GAAGCUGCUCAACUGCA
UCGUUUUC vRNA
Length Stability
Threshold
36.6% PAR
m7GAAGCUGCUCAACUGCA
UCGUUUUC vRNA
16 nucleotides ending in A
Priming
m7GAAGCUGCUCAACUGCAGCA
UCGUUUUC vRNA
Length Stability
Threshold
~0% PAR
m7GAAGCUGCUCAACUGCAGCAAAAA
UCGUUUUC vRNA

Figure 4.1. Graphical Model of the Determinants of Prime-and-Realign. Cap-snatched primers (blue) prime transcription either terminally at 3'U1 or internally at 3'C2 of the vRNA conserved sequences (black). Following transcription (red) of 3'Y4 (pink) the threshold for PAR is calculated on the basis of the length of the primer, and the stability of the RNA duplex. For the threshold, black letters indicate the primer is sufficient for transcription on the basis of that variable, grey letters indicate that the primer is insufficient for transcription on the basis of that variable; bolding these letters indicates the strength of the (in)sufficiency. PAR frequencies were obtained from Figure 3.11.1. The cap-snatched primer is listed in the 5' to 3' order; the 5' cap (m^7G) is not factored into the primer length of the primer in nucleotides. The vRNA conserved sequences are listed in the 3' to 5' order; vRNA indicates the remainder of the gene segment. **1.** A nine nucleotide long primer ending in an adenine residue is used to prime transcription of a 3'U4 templated vRNA segment. Following transcription of three nucleotides (GCA) the threshold is evaluated. This primer is strongly insufficient on the basis of length, yet is sufficient on the basis of stability; such a primer has a 71.4% chance to undergo PAR. If PAR occurs, transcription is re-primed with a 12 nucleotide long primer ending in an adenine residue. Following transcription of three nucleotides (GCA) the threshold is evaluated. This primer is strongly sufficient on the basis of length and sufficient on the basis of stability; such a primer has a 5.9% chance to undergo PAR. Transcription will most likely occur. **2.** An 11 nucleotide long primer ending in an AG dinucleotide is used to prime transcription of a 3'C4 templated vRNA segment. Following transcription of three nucleotides (GCG) the threshold is evaluated. This primer is insufficient on the basis of length, yet is strongly sufficient on the basis of stability; such a primer has a 5.7% change to undergo PAR. Transcription will most likely occur. **3.** A 13 nucleotide long primer ending in an uridine residue is used to prime transcription of a 3'U4 templated vRNA segment. Following transcription of three nucleotides (GCA) the threshold is evaluated. This primer is strongly sufficient on the basis of length, yet is strongly insufficient on the basis of stability; such a primer has a 36.6% chance to undergo PAR. If PAR occurs, transcription is re-primed with a 16 nucleotide long primer ending in an adenine residue. Following transcription of three nucleotides (GCA) the threshold is evaluated. This primer is strongly sufficient on the basis of length and sufficient on the basis of stability; such a primer has an ~0% chance to undergo PAR. Transcription will most likely occur.

5. Conclusion

I have found that primers are biased towards PAR on the basis of length and RNA duplex stability. Shorter primer lengths and lower RNA duplex stability arise due to breakdowns in length selective and/or sequence selective cleavage. PAR allows for both of these factors to be addressed in a single process. PAR typically adds a GCA addition (61.6±1.2% of observed additions) or longer nucleotide sequence ending in adenine regardless of either primer length (Figure 3.6.1) or conserved sequence transcribed (Figure 3.8.1). The net result following PAR is a primer three nucleotides longer ending in a compatible nucleotide sequence with 3'U1. By addressing the length, the PAR rate is exponentially decreased (Figure 3.4.4), and by increasing the sequence compatibility, the PAR frequency is further decreased (Figure 3.9.2). For example, a nine nucleotide long primer ending in an uridine residue has a 89% chance of undergoing PAR. One round of PAR, adding a GCA nucleotide sequence, converts this example primer to one 12 nucleotides long ending in an adenine residue; such a primer has a 5.76% change of undergoing PAR, and thus 94.24% chance of being used to primer IAV transcription (Figure 3.11.1). For this reason, PAR is a highly efficient rescue, rescuing 95.6±2.5% which underwent PAR (14.3±0.6% of all cap snatched primers) in a single round, with the number of primers requiring an additional round of PAR dropping exponentially (Figure 3.2.1). Prime-and-realign converts poor primers on the basis of length and sequence compatibility with the 3' end of the vRNA into one that can efficiently undergo transcription of the critical conserved sequence without errors. Prime-and-realign, therefore, affords tremendous flexibility to RdRp in cap snatched primer length and sequence compatibility.

6. References

[1] D. Sikora, L. Rocheleau, E. G. Brown, and M. Pelchat, “Influenza A virus cap-snatches host RNAs based on their abundance early after infection,” *Virology*, vol. 509, no. June, pp. 167–177, 2017.

[2] M. Lukarska *et al.*, “Structural basis of an essential interaction between influenza polymerase and Pol II CTD,” *Nature*, vol. 541, no. 7635, pp. 117–121, 2017.

[3] M. Martínez-alonso and N. Hengrung, “Polymerases Directly Bind the Serine-5-Phosphorylated Carboxyl- Terminal Domain of Host RNA Polymerase II,” vol. 90, no. 13, pp. 6014–6021, 2016.

[4] C. Geerts-Dimitriadou, R. Goldbach, and R. Kormelink, “Preferential use of RNA leader sequences during influenza A transcription initiation in vivo,” *Virology*, vol. 409, no. 1, pp. 27–32, 2011.

[5] C. Geerts-Dimitriadou, M. P. Zwart, R. Goldbach, and R. Kormelink, “Base-pairing promotes leader selection to prime in vitro influenza genome transcription,” *Virology*, vol. 409, no. 1, pp. 17–26, 2011.

[6] K. Datta, A. Wolkerstorfer, O. H. J. Szolar, S. Cusack, and K. Klumpp, “Characterization of PA-N terminal domain of Influenza A polymerase reveals sequence specific RNA cleavage,” *Nucleic Acids Res.*, vol. 41, no. 17, pp. 8289–8299, 2013.

[7] D. Sikora, L. Rocheleau, E. G. Brown, and M. Pelchat, “Deep sequencing reveals the eight facets of the influenza A/HongKong/1/1968 (H3N2) virus cap-snatching process,” *Sci. Rep.*, vol. 4, 2014.

[8] W. Gu *et al.*, “Influenza A virus preferentially snatches noncoding RNA caps,” *Rna*, vol. 21, no. 12, pp. 2067–2075, 2015.

[9] D. Koppstein, J. Ashour, and D. P. Bartel, “Sequencing the cap-snatching repertoire of H1N1 influenza provides insight into the mechanism of viral transcription initiation,” *Nucleic Acids Res.*, vol. 43, no. 10, pp. 5052–5064, 2015.

[10] J. K. Taubenberger and D. M. Morens, “Pandemic influenza--including a risk assessment of H5N1,” *Rev Sci Tech*, vol. 28, no. 1, pp. 187–202, 2009.

[11] A. D. Iuliano *et al.*, “Estimates of global seasonal influenza-associated respiratory mortality: a modelling study,” *Lancet*, vol. 391, no. 10127, pp. 1285–1300, 2018.

[12] G. A. Poland, “Vaccines against Avian Influenza — A Race against Time,” *N. Engl. J. Med.*, vol. 354, no. 13, pp. 1411–1413, 2006.

[13] T. K. W. Cheung and L. L. M. Poon, “Biology of influenza a virus,” *Ann. N. Y. Acad. Sci.*, vol. 1102, pp. 1–25, 2007.

[14] S. J. Plotch, M. Bouloy, I. Ulmanen, and R. M. Krug, “A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription,” *Cell*, vol. 23, no. 3, pp. 847–858, 1981.

[15] J. S. Robertson, M. Schubert, and R. A. Lazzarini, “Polyadenylation sites for influenza virus mRNA.,” *J. Virol.*, vol. 38, no. 1, pp. 157–63, 1981.

[16] J. W. Drake, “Rates of spontaneous mutation among RNA viruses.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 9, pp. 4171–5, 1993.

[17] E. Nobusawa and K. Sato, “Comparison of the Mutation Rates of Human Influenza A and B Viruses,” *J. Virol.*, vol. 80, no. 7, pp. 3675–3678, 2006.

[18] M. D. Pauly, M. C. Procario, and A. S. Lauring, “A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses,” *Elife*, vol. 6, pp. 1–18, 2017.

[19] E. Ghedin *et al.*, “Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution,” *Nature*, vol. 437, no. 7062, pp. 1162–1166, 2005.

[20] M. Selman, S. K. Dankar, N. E. Forbes, J. J. Jia, and E. G. Brown, “Adaptive mutation in influenza A virus non-structural gene is linked to host switching and induces a novel protein by alternative splicing,” *Emerg. Microbes Infect.*, vol. 1, no. 000, p. 0, 2012.

[21] L. Simonsen *et al.*, “The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance,” *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1811–1820, 2007.

[22] X. Xu *et al.*, “Reassortment and evolution of current human influenza A and B viruses,” *Virus Res.*, vol. 103, no. 1–2, pp. 55–60, 2004.

[23] A. Pflug, D. Guilligay, S. Reich, and S. Cusack, “Structure of influenza A polymerase bound to the viral RNA promoter,” *Nature*, vol. 516, no. 7531, pp. 355–360, 2014.

[24] S. M. Carr, E. Carnero, A. García-Sastre, G. G. Brownlee, and E. Fodor, “Characterization of a mitochondrial-targeting signal in the PB2 protein of influenza viruses,” *Virology*, vol. 344, no. 2, pp. 492–508, 2006.

[25] K. M. Graef *et al.*, “The PB2 Subunit of the Influenza Virus RNA Polymerase Affects Virulence by Interacting with the Mitochondrial Antiviral Signaling Protein and Inhibiting Expression of Beta Interferon,” *J. Virol.*, vol. 84, no. 17, pp. 8433–8445, 2010.

[26] J. C. D. Long and E. Fodor, “The PB2 Subunit of the Influenza A Virus RNA Polymerase Is Imported into the Mitochondrial Matrix,” *J. Virol.*, vol. 90, no. 19, pp. 8729–8738, 2016.

[27] S. Yamayoshi, M. Watanabe, H. Goto, and Y. Kawaoka, “Identification of a Novel Viral Protein Expressed from the PB2 Segment of Influenza A Virus,” *J. Virol.*, vol. 90, no. 1, pp. 444–456, 2016.

[28] W. Chen *et al.*, “A novel influenza A virus mitochondrial protein that induces cell death,” *Nat. Med.*, vol. 7, no. 12, pp. 1306–1312, 2001.

[29] Z. T. Varga *et al.*, “The influenza virus protein PB1-F2 inhibits the induction of type i interferon at the level of the MAVS adaptor protein,” *PLoS Pathog.*, vol. 7, no. 6, 2011.

[30] H. M. Wise *et al.*, “A Complicated Message: Identification of a Novel PB1-Related Protein Translated from Influenza A Virus Segment 2 mRNA,” *J. Virol.*, vol. 83, no. 16, pp. 8021–8031, 2009.

[31] J. J. Sanz-Ezquerro, T. Zürcher, S. de la Luna, J. Ortín, and a Nieto, “The amino-terminal one-third of the influenza virus PA protein is responsible for the induction of proteolysis.,” *J. Virol.*, vol. 70, no. 3, pp. 1905–1911, 1996.

[32] A. Rodriguez, A. Perez-Gonzalez, and A. Nieto, “Influenza Virus Infection Causes Specific Degradation of the Largest Subunit of Cellular RNA Polymerase II,” *J. Virol.*, vol. 81, no. 10, pp. 5315–5324, 2007.

[33] B. W. Jagger *et al.*, “An overlapping protein-coding region in influenza A virus segment 3 modulates the host response,” *Science (80-.).*, vol. 337, no. 6091, pp. 199–204, 2012.

[34] T. Hayashi, L. A. MacDonald, and T. Takimoto, “Influenza A Virus Protein PA-X Contributes to Viral Growth and Suppression of the Host Antiviral and Immune Responses,” *J. Virol.*, vol. 89, no. 12, pp. 6442–6452, 2015.

[35] D. A. Khaperskyy, S. Schmaling, J. Larkins-Ford, C. McCormick, and M. M. Gaglia, “Selective Degradation of Host RNA Polymerase II Transcripts by Influenza A Virus PA-X Host Shutoff Protein,” *PLoS Pathog.*, vol. 12, no. 2, pp. 1–25, 2016.

[36] Y. Muramoto, T. Noda, E. Kawakami, R. Akkina, and Y. Kawaoka, “Identification of Novel Influenza A Virus Proteins Translated from PA mRNA,” *J. Virol.*, vol. 87, no. 5, pp. 2455–2462, 2013.

[37] J. Stevens, A. L. Corper, C. F. Basler, J. K. Taubenberger, P. Palese, and I. A. Wilson, “Structure of the Uncleaved Human H1 Hemagglutinin from the Extinct 1918 Influenza Virus,” *Science (80-.).*, vol. 303, no. 5665, pp. 1866–1870, 2004.

[38] J. Stevens, O. Blixt, T. M. Tumpey, J. K. Taubenberger, J. C. Paulson, and I. A. Wilson, “Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus,” *Science (80-.).*, vol. 312, no. 5772, pp. 404–410, 2006.

[39] T. Horimoto, K. Nakayama, S. P. Smeekens, and Y. Kawaoka, “Proprotein-processing endoproteases PC6 and furin both activate hemagglutinin of virulent avian influenza viruses.,” *J. Virol.*, vol. 68, no. 9, pp. 6074–8, 1994.

[40] D. P. Nayak, R. A. Balogun, H. Yamada, Z. H. Zhou, and S. Barman, “Influenza virus morphogenesis and budding,” *Virus Research*, vol. 143, no. 2. pp. 147–161, 2009.

[41] B. Y. L. M. Staudt and W. Gerhard, “GENERATION RESPONSE OF ANTIBODY DIVERSITY IN THE IMMUNE VIRUS OF BALB / c MICE TO INFLUENZA HEMAGGLUTININ * I . Significant V a r i a t i o n in R e p e r t o i r e Expression b e t w e e n I n d i v i d u a l M i c e Much has been learned about the diver,” vol. 157, no. February, 1983.

[42] K. S. Matlin, H. Reggio, A. Helenius, and K. Simons, “Infectious Entry Pathway of Influenza-Virus in A Canine Kidney-Cell Line,” *J. Cell Biol.*, vol. 91, no. 3, pp. 601–613, 1981.

[43] R. J. Connor, Y. Kawaoka, R. G. Webster, and J. C. Paulson, “Receptor Specificity in Human, Avian, and Equine H2 and H3 Influenza Virus Isolates,” *Virology*, vol. 205, no. 1, pp. 17–23, 1994.

[44] H. Marjuki *et al.*, “Membrane accumulation of influenza A virus hemagglutinin triggers nuclear export of the viral genome via protein kinase Cα-mediated activation of ERK signaling,” *J. Biol. Chem.*, vol. 281, no. 24, pp. 16707–16715, 2006.

[45] G. Winter and S. Fields, “The structure of the gene encoding the nucleoprotein of human influenza virus A/PR/8/34,” *Virology*, vol. 114, no. 2, pp. 423–428, 1981.

[46] F. Baudin, C. Bach, S. Cusack, and R. W. Ruigrok, “Structure of influenza virus RNP. I. Influenza virus nucleoprotein melts secondary structure in panhandle RNA and exposes the bases to the solvent.,” *EMBO J.*, vol. 13, no. 13, pp. 3158–65, 1994.

[47] K. Yamanaka, A. Ishihama, and K. Nagata, “Reconstitution of influenza virus RNA-nucleoprotein complexes structurally resembling native viral ribonucleoprotein cores,” *J. Biol. Chem.*, vol. 265, no. 19, pp. 11151–11155, 1990.

[48] D. Elton, L. Medcalf, K. Bishop, D. Harrison, and P. Digard, “Identification of amino acid residues of influenza virus nucleoprotein essential for RNA binding.,” *J. Virol.*, vol. 73, no. 9, pp. 7357–7367, 1999.

[49] A. K.-L. Ng *et al.*, “Structure of the influenza virus A H5N1 nucleoprotein: implications for RNA binding, oligomerization, and vaccine design,” *FASEB J.*, vol. 22, no. 10, pp. 3638–3647, 2008.

[50] R. W. Compans, J. Content, and P. H. Duesberg, “Structure of the ribonucleoprotein of influenza virus.,” *J. Virol.*, vol. 10, no. 4, pp. 795–800, 1972.

[51] A. Labaronne, C. Swale, A. Monod, G. Schoehn, T. Crï¿½pin, and R. W. H. Ruigrok, “Binding of RNA by the nucleoproteins of influenza viruses A and B,” *Viruses*, vol. 8, no. 9, pp. 1–14, 2016.

[52] E. C. Hutchinson and E. Fodor, “Nuclear import of the influenza A virus transcriptional machinery,” *Vaccine*, vol. 30, no. 51, pp. 7353–7358, 2012.

[53] W. W. H. Wu, Y. H. B. Sun, and N. Panté, “Nuclear import of influenza A viral ribonucleoprotein complexes is mediated by two nuclear localization sequences on viral nucleoprotein,” *Virol. J.*, vol. 4, pp. 1–12, 2007.

[54] F. T. Vreede, T. E. Jung, and G. G. Brownlee, “Model suggesting that replication of influenza virus is regulated by stabilization of replicative intermediates.,” *J. Virol.*, vol. 78, no. 17, pp. 9568–72, 2004.

[55] G. I. Shapiro and R. M. Krug, “Influenza virus RNA replication in vitro: synthesis of viral template RNAs and virion RNAs in the absence of an added primer,” *J. Virol.*, vol. 62, no. 7, pp. 2285–2290, 1988.

[56] J. Hausmann, E. Kretzschmar, W. Garten, and H. D. Klenk, “Biosynthesis, intracellular transport and enzymatic activity of an avian influenza A virus neuraminidase: Role of unpaired cysteines and individual oligosaccharides,” *J. Gen. Virol.*, vol. 78, no. 12, pp. 3233–3245, 1997.

[57] J. N. Varghese and P. M. Colman, “Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2 armstrong resolution,” *J. Mol. Biol.*, vol. 221, pp. 473–486, 1991.

[58] R. Wagner, T. Wolff, A. Herwig, S. Pleschka, and H. D. Klenk, “Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics.,” *J. Virol.*, vol. 74, no. 14, pp. 6316–23, 2000.

[59] C. Liu, M. C. Eichelberger, R. W. Compans, and G. M. Air, “Influenza type A virus neuraminidase

does not play a role in viral entry, replication, assembly, or budding," *J. Virol.*, vol. 69, no. 2, pp. 1099–1106, 1995.

[60] R. W. H. Ruigrok, A. Barge, P. Durrer, J. Brunner, K. Ma, and G. R. Whittaker, "Membrane interaction of influenza virus M1 protein," *Virology*, vol. 267, no. 2, pp. 289–298, 2000.

[61] L. Wakefield and G. G. Brownlee, "RNA-binding properties of influenza a virus matrix protein M1," *Nucleic Acids Res.*, vol. 17, no. 21, pp. 8569–8580, 1989.

[62] C. Elster, K. Larsen, J. Gagnon, R. W. H. Ruigrok, and F. Baudin, "Influenza virus M1 protein binds to RNA through its nuclear localization signal," *J. Gen. Virol.*, vol. 78, no. 7, pp. 1589–1596, 1997.

[63] C. Chaimayo, T. Hayashi, A. Underwood, E. Hodges, and T. Takimoto, "Selective incorporation of vRNP into influenza A virions determined by its specific interaction with M1 protein," *Virology*, vol. 505, no. January, pp. 23–32, 2017.

[64] M. Bui, G. Whittaker, and a Helenius, "Effect of M1 protein and low pH on nuclear transport of influenza virus ribonucleoproteins.," *J. Virol.*, vol. 70, no. 12, pp. 8391–401, 1996.

[65] Z. Ye, D. Robinson, and R. R. Wagner, "Nucleus-targeting domain of the matrix protein (M1) of influenza virus," *J Virol*, vol. 69, no. 3, pp. 1964–1970, 1995.

[66] X. Huang, T. Liu, J. Muller, R. A. Levandowski, and Z. Ye, "Effect of influenza virus matrix protein and viral RNA on ribonucleoprotein formation and nuclear export," *Virology*, vol. 287, no. 2, pp. 405–416, 2001.

[67] T. Shimizu, N. Takizawa, K. Watanabe, K. Nagata, and N. Kobayashi, "Crucial role of the influenza virus NS2 (NEP) C-terminal domain in M1 binding and nuclear export of vRNP," *FEBS Lett.*, vol. 585, no. 1, pp. 41–46, 2011.

[68] L. J. Holsinger and R. Alams, "Influenza virus M2 integral membrane protein is a homotetramer stabilized by formation of disulfide bonds," *Virology*, vol. 183, no. 1, pp. 32–43, 1991.

[69] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A

virus," *Nature*, vol. 451, no. 7178, pp. 591–595, 2008.

[70] Y. Tang, F. Zaitseva, R. A. Lamb, and L. H. Pinto, "The gate of the influenza virus M2 proton channel is formed by a single tryptophan residue," *J. Biol. Chem.*, vol. 277, no. 42, pp. 39880–39886, 2002.

[71] C. Wang, R. A. Lamb, and L. H. Pinto, "Activation of the M2 ion channel of influenza virus: a role for the transmembrane domain histidine residue," *Biophys. J.*, vol. 69, no. 4, pp. 1363–1371, 1995.

[72] K. Takeuchi and R. A. Lamb, "Influenza virus M2 protein ion channel activity stabilizes the native form of fowl plague virus hemagglutinin during intracellular transport.," *J. Virol.*, vol. 68, no. 2, pp. 911–9, 1994.

[73] A. García-Sastre *et al.*, "Influenza A virus lacking the NS1 gene replicates in interferon- deficient systems," *Virology*, vol. 252, no. 2, pp. 324–330, 1998.

[74] S. de la Luna, P. Fortes, A. Beloso, and J. Ortin, "Influenza virus NS1 protein enhances the rate of translation initiation of viral mRNAs," *J Virol*, vol. 69, no. 4, pp. 2427–2433, 1995.

[75] T. Aragón, S. De Luna, I. Novoa, J. Ortín, A. Nieto, and C. N. De Biotecnologı, "Eukaryotic Translation Initiation Factor 4GI Is a Cellular Target for NS1 Protein , a Translational Activator of Influenza Virus Eukaryotic Translation Initiation Factor 4GI Is a Cellular Target for NS1 Protein , a Translational Activator of Influenza Vir," *Society*, vol. 20, no. 17, pp. 6259–6268, 2000.

[76] I. Burgui, T. Aragón, J. Ortín, and A. Nieto, "PABP1 and eIF4GI associate with influenza virus NS1 protein in viral mRNA translation initiation complexes," *J. Gen. Virol.*, vol. 84, no. 12, pp. 3263–3274, 2003.

[77] a M. Falcón, P. Fortes, R. M. Marión, a Beloso, and J. Ortín, "Interaction of influenza virus NS1 protein and the human homologue of Staufen in vivo and in vitro.," *Nucleic Acids Res.*, vol. 27, no. 11, pp. 2241–2247, 1999.

[78] Y. Qiu and R. M. Krug, “The influenza virus NS1 protein is a poly(A)-binding protein that inhibits nuclear export of mRNAs containing poly(A).,” *J. Virol.*, vol. 68, no. 4, pp. 2425–2432, 1994.

[79] M. E. Nemeroff, S. M. L. Barabino, Y. Li, W. Keller, and R. M. Krug, “Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3’ end formation of cellular pre-mRNAs,” *Mol. Cell*, vol. 1, no. 7, pp. 991–1000, 1998.

[80] D. L. Noah, K. Y. Twu, and R. M. Krug, “Cellular antiviral responses against influenza A virus are countered at the posttranscriptional level by the viral NS1A protein via its binding to a cellular protein required for the 3’ end processing of cellular pre-mRNAS,” *Virology*, vol. 307, no. 2, pp. 386–395, 2003.

[81] N. Satterly *et al.*, “Influenza virus targets the mRNA export machinery and the nuclear pore complex,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 6, pp. 1853–1858, 2007.

[82] B. G. Hale, R. E. Randall, J. Ortin, and D. Jackson, “The multifunctional NS1 protein of influenza A viruses,” *J. Gen. Virol.*, vol. 89, no. 10, pp. 2359–2376, 2008.

[83] J. Dubois, O. Terrier, and M. Rosa-Calatrava, “Influenza viruses and mRNA splicing: Doing more with less,” *MBio*, vol. 5, no. 3, pp. 1–13, 2014.

[84] R. A. Lamb, P. R. Etkind, and P. W. Choppin, “Evidence for a ninth influenza viral polypeptide,” *Virology*, vol. 91, no. 1, pp. 60–78, 1978.

[85] G. Neumann, M. T. Hughes, and Y. Kawaoka, “Influenza A virus NS2 protein mediates vRNP nuclear export through NES-independent interaction with hCRM1,” *EMBO J.*, vol. 19, no. 24, pp. 6751–6758, 2000.

[86] K. Iwatsuki-horimoto, T. Horimoto, Y. Fujii, and Y. Kawaoka, “Generation of In uenza A Virus NS2 (NEP) Mutants with an Altered Nuclear Export Signal Sequence,” *Society*, vol. 78, no. 18, pp. 10149–10155, 2004.

[87] R. E. O’Neill, J. Talon, and P. Palese, “The influenza virus NEP (NS2 protein) mediates the nuclear

export of viral ribonucleoproteins," *EMBO J.*, vol. 17, no. 1, pp. 288–296, 1998.

[88] N. C. Robb, M. Smith, F. T. Vreede, and E. Fodor, "NS2/NEP protein regulates transcription and replication of the influenza virus RNA genome," *J. Gen. Virol.*, vol. 90, no. 6, pp. 1398–1407, 2009.

[89] B. Mänz, L. Brunotte, P. Reuther, and M. Schwemmle, "Adaptive mutations in NEP compensate for defective H5N1 RNA replication in cultured human cells," *Nat. Commun.*, vol. 3, no. May, 2012.

[90] M. Cohen *et al.*, "Influenza A penetrates host mucus by cleaving sialic acids with neuraminidase," *Virol. J.*, vol. 10, pp. 1–13, 2013.

[91] K. Martin and A. Helenius, "Transport of incoming influenza virus nucleocapsids into the nucleus," *Trends Cell Biol.*, vol. 2, no. 1, p. 8, 1992.

[92] Y. ying Chou, N. S. Heaton, Q. Gao, P. Palese, R. Singer, and T. Lionnet, "Colocalization of Different Influenza Viral RNA Segments in the Cytoplasm before Viral Budding as Shown by Single-molecule Sensitivity FISH Analysis," *PLoS Pathog.*, vol. 9, no. 5, 2013.

[93] A. J. Hay, B. Lomniczi, A. R. Bellamy, and J. J. Skehel, "Transcription of the influenza virus genome," *Virology*, vol. 83, no. 2, pp. 337–355, 1977.

[94] I. Widjaja, E. de Vries, P. J. M. Rottier, and C. A. M. de Haan, "Competition between Influenza A Virus Genome Segments," *PLoS One*, vol. 7, no. 10, 2012.

[95] D. Vester *et al.*, "Real-time RT-qPCR assay for the analysis of human influenza A virus transcription and replication dynamics," *J. Virol. Methods*, vol. 168, no. 1–2, pp. 63–71, 2010.

[96] G. I. Shapiro, T. Gurney Jr., and R. M. Krug, "Influenza virus gene expression: control mechanisms at early and late times of infection and nuclear-cytoplasmic transport of virus-specific RNAs," *J. Virol.*, vol. 61, no. 3, pp. 764–773, 1987.

[97] E. Hatada and M. Hasegawa, "Control of influenza virus gene expression: quantitative analysis of each viral RNA species in infected cells," *J. Biochem.*, vol. 105, no. 4, pp. 537–546, 1989.

[98] K. Watanabe, H. Handa, K. Mizumoto, and K. Nagata, “Mechanism for inhibition of influenza virus RNA polymerase activity by matrix protein,” *J Virol*, vol. 70, no. 1, pp. 241–247, 1996.

[99] L. Zhao *et al.*, “New Insights into the Nonconserved Noncoding Region of the Subtype-Determinant Hemagglutinin and Neuraminidase Segments of Influenza A Viruses,” *J. Virol.*, vol. 88, no. 19, pp. 11493–11503, 2014.

[100] B. Crescenzo-Chaigne, C. V. S. Barbezange, S. Léandri, C. Roquin, C. Berthault, and S. Van Der Werf, “Incorporation of the influenza A virus NA segment into virions does not require cognate non-coding sequences,” *Sci. Rep.*, vol. 7, no. August 2016, pp. 1–13, 2017.

[101] Y. Kobayashi, B. Dadonaite, N. van Doremalen, Y. Suzuki, W. S. Barclay, and O. G. Pybus, “Computational and molecular analysis of conserved influenza A virus RNA secondary structures involved in infectious virion production,” *RNA Biol.*, vol. 13, no. 9, pp. 883–894, 2016.

[102] R. Flick and G. Hobom, “Transient bicistronic vRNA segments for indirect selection of recombinant influenza viruses,” *Virology*, vol. 262, no. 1, pp. 93–103, 1999.

[103] A. J. W. te Velthuis and E. Fodor, “Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis,” *Nat. Rev. Microbiol.*, vol. 14, no. 8, pp. 479–493, 2016.

[104] A. Rodriguez-Frandsen, R. Alfonso, and A. Nieto, “Influenza virus polymerase: Functions on host range, inhibition of cellular response to infection and pathogenicity,” *Virus Res.*, vol. 209, pp. 23–38, 2015.

[105] J. Braam, I. Ulmanen, and R. M. Krug, “Molecular model of a eucaryotic transcription complex: Functions and movements of influenza P proteins during capped RNA-primed transcription,” *Cell*, vol. 34, no. 2, pp. 611–618, 1983.

[106] S. K. Biswas and D. P. Nayak, “Mutational analysis of the conserved motifs of influenza A virus polymerase basic protein 1.,” *J. Virol.*, vol. 68, no. 3, pp. 1819–26, 1994.

[107] D. Guilligay *et al.*, “The structural basis for cap binding by influenza virus polymerase subunit

PB2," *Nat. Struct. Mol. Biol.*, vol. 15, no. 5, pp. 500–506, 2008.

[108] A. Dias *et al.*, "The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit," *Nature*, vol. 458, no. 7240, pp. 914–918, 2009.

[109] A. Pflug *et al.*, "Capped RNA primer binding to influenza polymerase and implications for the mechanism of cap-binding inhibitors," *Nucleic Acids Res.*, vol. 46, no. 2, pp. 956–971, 2018.

[110] S. Reich, D. Guilligay, and S. Cusack, "An in vitro fluorescence based study of initiation of RNA synthesis by influenza B polymerase," *Nucleic Acids Res.*, vol. 45, no. 6, pp. 3353–3368, 2016.

[111] E. Thierry *et al.*, "Influenza Polymerase Can Adopt an Alternative Configuration Involving a Radical Repacking of PB2 Domains," *Mol. Cell*, vol. 61, no. 1, pp. 125–137, 2016.

[112] J. Mukaigawa and D. P. Nayak, "Two signals mediate nuclear localization of influenza virus (A/WSN/33) polymerase basic protein 2.," *J. Virol.*, vol. 65, no. 1, pp. 245–53, 1991.

[113] F. Tarendeau *et al.*, "Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit," *Nat. Struct. Mol. Biol.*, vol. 14, no. 3, pp. 229–233, 2007.

[114] S. Boivin and D. J. Hart, "Interaction of the influenza A virus polymerase PB2 C-terminal region with importin ?? isoforms provides insights into host adaptation and polymerase assembly," *J. Biol. Chem.*, vol. 286, no. 12, pp. 10439–10448, 2011.

[115] T. Naito, F. Momose, A. Kawaguchi, and K. Nagata, "Involvement of Hsp90 in Assembly and Nuclear Import of Influenza Virus RNA Polymerase Subunits," *J. Virol.*, vol. 81, no. 3, pp. 1339–1349, 2007.

[116] E. Fodor and M. Smith, "The PA Subunit Is Required for Efficient Nuclear Accumulation of the PB1 Subunit of the Influenza A Virus RNA Polymerase Complex The PA Subunit Is Required for Efficient Nuclear Accumulation of the PB1 Subunit of the Influenza A Virus RNA Polymerase Compl," *Society*, vol. 78, no. 17, pp. 9144–9153, 2004.

[117] A. Nieto, S. de la Luna, J. Barcena, A. Portela, and J. Ortin, "Complex structure of the nuclear

translocation signal of influenza virus polymerase PA subunit.," *J. Gen. Virol.*, vol. 75 (Pt 1), no. 1, pp. 29–36, 1994.

[118] S. Huet, S. V. Avilov, L. Ferbitz, N. Daigle, S. Cusack, and J. Ellenberg, "Nuclear Import and Assembly of Influenza A Virus RNA Polymerase Studied in Live Cells by Fluorescence Cross-Correlation Spectroscopy," *J. Virol.*, vol. 84, no. 3, pp. 1254–1264, 2010.

[119] S. T. Nath and D. P. Nayak, "Function of two discrete regions is required for nuclear localization of polymerase basic protein 1 of A/WSN/33 influenza virus (H1 N1)," *Mol Cell Biol*, vol. 10, no. 8, pp. 4139–4145, 1990.

[120] E. C. Hutchinson, O. E. Orr, S. M. Liu, O. G. Engelhardt, and E. Fodor, "Characterization of the interaction between the influenza A virus polymerase subunit PB1 and the host nuclear import factor ran-binding protein 5," *J. Gen. Virol.*, vol. 92, no. 8, pp. 1859–1869, 2011.

[121] S. Jakel, "Importins fulfil a dual function as nuclear import receptors and cytoplasmic chaperones for exposed basic domains," *EMBO J.*, vol. 21, no. 3, pp. 377–386, 2002.

[122] C. Swale *et al.*, "Structural characterization of recombinant IAV polymerase reveals a stable complex between viral PA-PB1 heterodimer and host RanBP5," *Sci. Rep.*, vol. 6, no. April, pp. 1–14, 2016.

[123] U. Desselberger, V. R. Racaniello, J. J. Zazra, and P. Palese, "The 3' and 5'-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity," *Gene*, vol. 8, no. 3, pp. 315–328, 1980.

[124] D. H. Turner and D. H. Mathews, "NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp. 2009–2011, 2009.

[125] A. I. Tomescu, N. C. Robb, N. Hengrung, E. Fodor, and A. N. Kapanidis, "Single-molecule FRET reveals a corkscrew RNA structure for the polymerase-bound influenza virus promoter," *Proc.*

Natl. Acad. Sci., vol. 111, no. 32, pp. E3335–E3342, 2014.

[126] L. L. Poon, D. C. Pritlove, E. Fodor, and G. G. Brownlee, "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.*, vol. 73, no. 4, pp. 3473–6, 1999.

[127] L. L. Poon, D. C. Pritlove, J. Sharps, and G. G. Brownlee, "The RNA polymerase of influenza virus, bound to the 5' end of virion RNA, acts in cis to polyadenylate mRNA.," *J. Virol.*, vol. 72, no. 10, pp. 8214–9, 1998.

[128] H. Zheng, H. A. Lee, P. Palese, and A. García-Sastre, "Influenza A virus RNA polymerase has the ability to stutter at the polyadenylation site of a viral RNA template during RNA replication.," *J. Virol.*, vol. 73, no. 6, pp. 5240–3, 1999.

[129] X. Li and P. Palese, "Characterization of the polyadenylation signal of influenza virus RNA.," *J. Virol.*, vol. 68, no. 2, pp. 1245–9, 1994.

[130] S. Reich *et al.*, "Structural insight into cap-snatching and RNA synthesis by influenza polymerase," *Nature*, vol. 516, no. 7531, pp. 361–366, 2014.

[131] N. C. Robb *et al.*, "Single-molecule FRET reveals the pre-initiation and initiation conformations of influenza virus promoter RNA," *Nucleic Acids Res.*, vol. 44, no. 21, pp. 10304–10315, 2016.

[132] P. Gerlach, H. Malet, S. Cusack, and J. Reguera, "Structural insights into bunyavirus replication and its regulation by the vRNA promoter," *Cell*, vol. 161, no. 6, pp. 1267–1279, 2015.

[133] F. Ferron, F. Weber, J. C. de la Torre, and J. Reguera, "Transcription and replication mechanisms of Bunyaviridae and Arenaviridae L proteins," *Virus Res.*, vol. 234, pp. 118–134, 2017.

[134] A. Amroun, S. Priet, X. de Lamballerie, and G. Quérat, "Bunyaviridae RdRps: structure, motifs, and RNA synthesis machinery," *Crit. Rev. Microbiol.*, vol. 43, no. 6, pp. 753–778, 2017.

[135] K. Hara, T. Kashiwagi, N. Hamada, and H. Watanabe, "Basic amino acids in the N-terminal half of the PB2 subunit of influenza virus RNA polymerase are involved in both transcription and

replication," *J. Gen. Virol.*, vol. 98, no. 5, pp. 900–905, 2017.

[136] C. Chu *et al.*, "Functional analysis of conserved motifs in influenza virus PB1 protein.," *PLoS One*, vol. 7, no. 5, pp. 3–10, 2012.

[137] A. J. W. Te Velthuis, N. C. Robb, A. N. Kapanidis, and E. Fodor, "The role of the priming loop in influenza A virus RNA synthesis," *Nat. Microbiol.*, vol. 1, no. 5, pp. 1–16, 2016.

[138] J. Oymans and A. J. W. Te Velthuis, "Initiation , Elongation , and Realignment during Influenza Virus," *J. Virol.*, vol. 92, no. 3, pp. e01775-17, 2018.

[139] J. Oymans and A. J. W. Te Velthuis, "A Mechanism for Priming and Realignment during Influenza A Virus Replication.," *J. Virol.*, vol. 92, no. 3, pp. e01773-17, 2018.

[140] A. Pflug, M. Lukarska, P. Resa-Infante, S. Reich, and S. Cusack, "Structural insights into RNA synthesis by the influenza virus transcription-replication machine," *Virus Res.*, vol. 234, pp. 103–117, 2017.

[141] J. T. Perez *et al.*, "Influenza A virus-generated small RNAs regulate the switch from transcription to replication," *Proc. Natl. Acad. Sci.*, vol. 107, no. 25, pp. 11525–11530, 2010.

[142] S. Zhang, J. Wang, Q. Wang, and T. Toyoda, "Internal initiation of influenza virus replication of viral RNA and complementary RNA in vitro," *J. Biol. Chem.*, vol. 285, no. 52, pp. 41194–41201, 2010.

[143] A. J. W. Velthuis, N. C. Robb, A. N. Kapanidis, and E. Fodor, "The role of the priming loop in in fl uenza A virus RNA synthesis," *Nat. Microbiol.*, vol. 651, no. March, pp. 1–16, 2016.

[144] D. Garcin and D. Kolakofsky, "Tacaribe arenavirus RNA synthesis in vitro is primer dependent and suggests an unusual model for the initiation of genome replication.," *J. Virol.*, vol. 66, no. 3, pp. 1370–1376, 1992.

[145] A. York, N. Hengrung, F. T. Vreede, J. T. Huiskonen, and E. Fodor, "Isolation and characterization of the positive-sense replicative intermediate of a negative strand RNA virus," *Proc Natl Acad Sci*

USA, vol. 110, pp. E4238–E4245, 2013.

[146] N. Jorba, R. Coloma, and J. Ortín, "Genetic trans-complementation establishes a new model for influenza virus RNA transcription and replication," *PLoS Pathog.*, vol. 5, no. 5, 2009.

[147] E. Fodor, D. C. Pritlove, and G. G. Brownlee, "The Influenza-Virus Panhandle Is Involved in the Initiation of Transcription," *J Virol*, vol. 68, no. 6, pp. 4092–4096, 1994.

[148] R. M. Presti *et al.*, "Quaranfil, Johnston Atoll, and Lake Chad Viruses Are Novel Members of the Family Orthomyxoviridae," *J. Virol.*, vol. 83, no. 22, pp. 11599–11606, 2009.

[149] T. Sandvik, E. Rimstad, and S. Mjaaland, "The viral RNA 3'- and 5'-end structure and mRNA transcription of infectious salmon anaemia virus resemble those of influenza viruses," *Arch. Virol.*, vol. 145, no. 8, pp. 1659–1669, 2000.

[150] D. Guilligay *et al.*, "Comparative structural and functional analysis of Orthomyxovirus polymerase cap-snatching domains," *PLoS One*, vol. 9, no. 1, 2014.

[151] M. B. Leahy, J. T. Dessens, and P. a Nuttall, "In vitro polymerase activity of Thogoto virus: evidence for a unique cap-snatching mechanism in a tick-borne orthomyxovirus.," *J. Virol.*, vol. 71, no. 11, pp. 8347–8351, 1997.

[152] A. V. Bussetti *et al.*, "Genomic and antigenic characterization of Jos virus," *J. Gen. Virol.*, vol. 93, no. 2, pp. 293–298, 2012.

[153] B. C. Ramirez, D. Garcin, L. a Calvert, D. Kolakofsky, and a L. Haenni, "Capped nonviral sequences at the 5' end of the mRNAs of rice hoja blanca virus RNA4.," *J. Virol.*, vol. 69, no. 3, pp. 1951–1954, 1995.

[154] J. J. Walia and B. W. Falk, "Fig mosaic virus mRNAs show generation by cap-snatching," *Virology*, vol. 426, no. 2, pp. 162–166, 2012.

[155] P. Vialat and M. Bouloy, "Germiston virus transcriptase requires active 40S ribosomal subunits and utilizes capped cellular RNAs," *J. Virol.*, vol. 66, no. 2, pp. 685–693, 1992.

[156] F. van Poelwijk, J. Kolkman, and R. Goldbach, “Sequence analysis of the 5’ ends of tomato spotted wilt virus N mRNAs,” *Arch Virol*, vol. 141, no. 1, pp. 177–184, 1996.

[157] D. Duijsings, R. Kormelink, and R. Goldbach, “In vivo analysis of the TSWV cap-snatching mechanism: Single base complementarity and primer length requirements,” *EMBO J.*, vol. 20, no. 10, pp. 2545–2552, 2001.

[158] I. Van Knippenberg, M. Lamine, R. Goldbach, and R. Kormelink, “Tomato spotted wilt virus transcriptase in vitro displays a preference for cap donors with multiple base complementarity to the viral template,” *Virology*, vol. 335, no. 1, pp. 122–130, 2005.

[159] L. Huiet, P. A. Feldstein, J. H. Tsai, and B. W. Falk, “The maize stripe virus major noncapsid protein messenger RNA transcripts contain heterogeneous leader sequences at their 5’ termini,” *Virology*, vol. 197, no. 2. pp. 808–812, 1993.

[160] E. M. Estabrook, J. Tsai, and B. W. Falk, “In vivo transfer of barley stripe mosaic hordeivirus ribonucleotides to the 5’ terminus of maize stripe tenuivirus RNAs.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 14, pp. 8304–8309, 1998.

[161] M. Yao, T. Zhang, T. Zhou, Y. Zhou, X. Zhou, and X. Tao, “Repetitive prime-and-realign mechanism converts short capped RNA leaders into longer ones that may be more suitable for elongation during rice stripe virus transcription initiation,” *J. Gen. Virol.*, vol. 93, no. 1, pp. 194–202, 2012.

[162] X. Liu *et al.*, “Inherent properties not conserved in other tenuiviruses increase priming and realignment cycles during transcription of Rice stripe virus,” *Virology*, vol. 496, pp. 287–298, 2016.

[163] X. Liu *et al.*, “Rice Stripe Tenuivirus Has a Greater Tendency To Use the Prime-and-Realign Mechanism in Transcription of Genomic than in Transcription of Antigenomic Template RNAs,” *J. Virol.*, vol. 92, no. 1, pp. e01414-17, 2018.

[164] D. K. Dobie *et al.*, “Analysis of LaCrosse virus S mRNA 5’ termini in infected mosquito cells and

Aedes triseriatus mosquitoes.," *J. Virol.*, vol. 71, no. 6, pp. 4395–4399, 1997.

[165] H. Jin and R. M. Elliott, "Characterization of Bunyamwera virus S RNA that is transcribed and replicated by the L protein expressed from recombinant vaccinia virus," *J. Virol.*, vol. 67, no. 3, pp. 1396–404, 1993.

[166] D. Coupeau, F. Claine, L. Wiggers, B. Martin, N. Kirschvink, and B. Muylkens, "Characterization of messenger RNA termini in Schmallenberg virus and related Simbuviruses," *J. Gen. Virol.*, vol. 94, no. PART 11, pp. 2399–2405, 2013.

[167] A. Amroun, S. Priet, and G. Querat, "Toscana virus cap-snatching and initiation of transcription," *J. Gen. Virol.*, vol. 98, no. 11, pp. 2676–2688, 2017.

[168] H. Jin and R. M. Elliott, "Non-viral sequences at the 5' ends of Dugbe nairovirus S mRNAs," *J. Gen. Virol.*, vol. 74, no. 10, pp. 2293–2297, 1993.

[169] Y. Zheng *et al.*, "Actinidia chlorotic ringspot-associated virus: a novel emaravirus infecting kiwifruit plants," *Mol. Plant Pathol.*, vol. 18, no. 4, pp. 569–581, 2017.

[170] D. Garcin *et al.*, "The 5' Ends of Hantaan virus (Bunyaviridae) RNAs Suggest a Prime-and-Realign Mechanism for the Initiation of RNA Synthesis.," *J. Virol.*, vol. 69, no. 9, pp. 5754–5762, 1995.

[171] J. F. Simons and R. F. Pettersson, "Host-derived 5' ends and overlapping complementary 3' ends of the two mRNAs transcribed from the ambisense S segment of Uukuniemi virus.," *J. Virol.*, vol. 65, no. 9, pp. 4741–4748, 1991.

[172] E. Cheng and M. A. Mir, "Signatures of Host mRNA 5' Terminus for Efficient Hantavirus Cap Snatching," *J. Virol.*, vol. 86, no. 18, pp. 10173–10185, 2012.

[173] M. A. Mir, S. Sheema, A. Haseeb, and A. Haque, "Hantavirus nucleocapsid protein has distinct m7G cap- and RNA-binding sites," *J. Biol. Chem.*, vol. 285, no. 15, pp. 11357–11368, 2010.

[174] M. A. Mir, W. A. Duran, B. L. Hjelle, C. Ye, and A. T. Panganiban, "Storage of cellular 5' mRNA caps in P bodies for viral cap-snatching," *Proc. Natl. Acad. Sci.*, vol. 105, no. 49, pp. 19294–19299,

2008.

[175] R. Raju, L. Raju, D. Hacker, D. Garcin, R. Compans, and D. Kolakofsky, “Nontemplated bases at the 5′ ends of tacaribe virus mRNAs,” *Virology*, vol. 174, no. 1, pp. 53–59, 1990.

[176] B. J. Meyer and P. J. Southern, “Concurrent sequence analysis of 5′ and 3′ RNA termini by intramolecular circularization reveals 5′ nontemplated bases and 3′ terminal heterogeneity for lymphocytic choriomeningitis virus mRNAs.,” *J. Virol.*, vol. 67, no. 5, pp. 2621–7, 1993.

[177] S. J. Polyak, S. Zheng, and D. G. Harnish, “5′ termini of Pichinde arenavirus S RNAs and mRNAs contain nontemplated nucleotides,” *J. Virol.*, vol. 69, no. 5, pp. 3211–3215, 1995.

[178] B. Morin *et al.*, “The N-terminal domain of the arenavirus L protein is an RNA endonuclease essential in mRNA transcription,” *PLoS Pathog.*, vol. 6, no. 9, 2010.

[179] M. Lelke, L. Brunotte, C. Busch, and S. Gunther, “An N-Terminal Region of Lassa Virus L Protein Plays a Critical Role in Transcription but Not Replication of the Virus Genome,” *J. Virol.*, vol. 84, no. 4, pp. 1934–1944, 2010.

[180] X. Qi *et al.*, “Cap binding and immune evasion revealed by Lassa nucleoprotein structure,” *Nature*, vol. 468, no. 7325, pp. 779–785, 2010.

[181] R. a Lamb and P. W. Choppin, “Synthesis of influenza virus polypeptides in cells resistant to alpha-amanitin: evidence for the involvement of cellular RNA polymerase II in virus replication.,” *J. Virol.*, vol. 23, no. 3, pp. 816–819, 1977.

[182] O. Engelhardt, M. Smith, and E. Fodor, “Association of the influenza A virus RNA-dependent RNA polymerase with cellular RNA polymerase II,” *J. Virol.*, vol. 79, no. 9, pp. 5812–5818, 2005.

[183] A. Y. Chan, F. T. Vreede, M. Smith, O. G. Engelhardt, and E. Fodor, “Influenza virus inhibits RNA polymerase II elongation,” *Virology*, vol. 351, no. 1, pp. 210–217, 2006.

[184] E. B. Rasmussen and J. T. Lis, “In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes.,” *Proc. Natl. Acad. Sci.*, vol. 90, no. 17, pp. 7923–7927, 1993.

[185] H.-M. Lv, X.-L. Guo, R.-X. Gu, and D.-Q. Wei, “Free Energy Calculations and Binding Analysis of Two Potential Anti- Influenza Drugs with Polymerase Basic Protein-2 (PB2),” *Protein Pept. Lett.*, vol. 18, no. 10, pp. 1002–1009, 2011.

[186] Y. J. Wang *et al.*, “Computational Studies on the Substrate Interactions of Influenza A Virus PB2 Subunit,” *PLoS One*, vol. 7, no. 9, pp. 1–7, 2012.

[187] L. Hooker, R. R. Sully, B. Handa, N. Ono, H. Koyano, and K. Klumpp, “Quantitative analysis of influenza virus RNP interaction with RNA cap structures and comparison to human cap binding protein eIF4E,” *Biochemistry*, vol. 42, no. 20, pp. 6234–6240, 2003.

[188] M. L. Li, P. Rao, and R. M. Krug, “The active sites of the influenza cap-dependent endonuclease are on different polymerase subunits,” *EMBO J.*, vol. 20, no. 8, pp. 2078–2086, 2001.

[189] M. B. Leahy, D. C. Pritlove, L. L. M. Poon, and G. G. Brownlee, “Mutagenic Analysis of the 5’ Arm of the Influenza A Virus Virion RNA Promoter Defines the Sequence Requirements for Endonuclease Activity,” *J. Virol.*, vol. 75, no. 1, pp. 134–142, 2001.

[190] M. B. Leahy, H. C. Dobbyn, and G. G. Brownlee, “Hairpin Loop Structure in the 3’ Arm of the Influenza A Virus Virion RNA Promoter Is Required for Endonuclease Activity,” *J. Virol.*, vol. 75, no. 15, pp. 7042–7049, 2001.

[191] M. K. Lee, S. H. Bae, C. J. Park, H. K. Cheong, C. Cheong, and B. S. Choi, “A single-nucleotide natural variation (U4 to C4) in an influenza A virus promoter exhibits a large structural change: Implications for differential viral RNA synthesis by RNA-dependent RNA polymerase,” *Nucleic Acids Res.*, vol. 31, no. 4, pp. 1216–1223, 2003.

[192] P. Yuan *et al.*, “Crystal structure of an avian influenza polymerase PA N reveals an endonuclease active site,” *Nature*, vol. 458, no. 7240, pp. 909–913, 2009.

[193] J. Reguera *et al.*, “Comparative Structural and Functional Analysis of Bunyavirus and Arenavirus Cap-Snatching Endonucleases,” *PLoS Pathog.*, vol. 12, no. 6, pp. 1–24, 2016.

[194] T. Crepin, A. Dias, A. Palencia, C. Swale, S. Cusack, and R. W. H. Ruigrok, “Mutational and Metal Binding Analysis of the Endonuclease Domain of the Influenza Virus Polymerase PA Subunit,” *J. Virol.*, vol. 84, no. 18, pp. 9096–9104, 2010.

[195] M. W. Shaw and R. A. Lamb, “A specific sub-set of host-cell mRNAs prime influenza virus mRNA synthesis,” *Virus Res.*, vol. 1, no. 6, pp. 455–467, 1984.

[196] A. R. Beaton and R. M. Krug, “Selected host cell capped RNA fragments prime influenza viral RNA transcription in vivo,” *Nucleic Acids Res.*, vol. 9, no. 17, pp. 4423–4436, 1981.

[197] O. Terrier *et al.*, “Cellular transcriptional profiling in human lung epithelial cells infected by different subtypes of influenza A viruses reveals an overall down-regulation of the host p53 pathway,” *Virol. J.*, vol. 8, pp. 1–11, 2011.

[198] Y. He *et al.*, “Influenza A Virus Replication Induces Cell Cycle Arrest in G0/G1 Phase,” *J. Virol.*, vol. 84, no. 24, pp. 12832–12840, 2010.

[199] A. Bercovich-Kinori *et al.*, “A systematic view on influenza induced host shutoff,” *Elife*, vol. 5, no. AUGUST, 2016.

[200] P. and S. Langmead, Trapnell, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biol.*, 2009.

[201] R Core Team, “R: A Language and Environment for Statistical Computing.” Vienna, Austria, 2018.

[202] RStudio Team, “RStudio: Integrated Development Environment for R.” Boston, MA, 2016.

[203] M. Corporation and S. Weston, “doParallel: Foreach Parallel Adaptor for the ‘parallel’ Package.” 2017.

[204] Y. Xie, “knitr: A General-Purpose Package for Dynamic Report Generation in R.” 2018.

[205] H. Wickham *et al.*, “ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.” 2018.

[206] C. O. Wilke, “cowplot: Streamlined Plot Theme and Plot Annotations for ‘ggplot2.’” 2018.

[207] O. Wagih, “ggseqlogo: A ‘ggplot2’ Extension for Drawing Publication-Ready Sequence Logos.” 2017.

[208] H. Wickham, “reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package.” 2017.

[209] A. Kassambara, “ggpubr: ‘ggplot2’ Based Publication Ready Plots.” 2018.

[210] A. R. Davis and B. M. Znosko, “Thermodynamic characterization of single mismatches found in naturally occurring RNA,” *Biochemistry*, vol. 46, no. 46, pp. 13425–13436, 2007.

[211] A. R. Davis and B. M. Znosko, “Positional and neighboring base pair effects on the thermodynamic stability of RNA single mismatches,” *Biochemistry*, vol. 49, no. 40, pp. 8669–8679, 2010.

[212] L. Y. Geer *et al.*, “The NCBI BioSystems database,” *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, 2009.

[213] Y. Bao *et al.*, “The influenza virus resource at the National Center for Biotechnology Information.,” *J. Virol.*, vol. 82, no. 2, pp. 596–601, 2008.

[214] F. Mignone, C. Gissi, S. Liuni, and G. Pesole, “Untranslated regions of mRNAs.,” *Genome Biol.*, vol. 3, no. 3, p. REVIEWS0004, 2002.

[215] M. Sémon, D. Mouchiroud, and L. Duret, “Relationship between gene expression and GC-content in mammals: Statistical significance and biological relevance,” *Hum. Mol. Genet.*, vol. 14, no. 3, pp. 421–427, 2005.

[216] N. B. Leontis, “The non-Watson-Crick base pairs and their associated isostericity matrices,” *Nucleic Acids Res.*, vol. 30, no. 16, pp. 3497–3531, 2002.

[217] M. W. Shaw, R. a Lamb, B. W. Erickson, D. J. Briedis, and P. W. Choppin, “Complete nucleotide sequence of the neuraminidase gene of influenza B virus.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 79, no. 22, pp. 6817–6821, 1982.

[218] T. A. Kunkel and P. S. Alexander, “The base substitution fidelity of eucaryotic DNA polymerases: Mispairing frequencies, site preferences, insertion preferences, and base substitution by

dislocation," *J. Biol. Chem.*, vol. 261, no. 1, pp. 160–166, 1986.

[219] K. Bebenek and T. A. Kunkel, "Frameshift errors initiated by nucleotide misincorporation.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 13, pp. 4946–50, 1990.

[220] K. H. Lee and B. L. Seong, "The position 4 nucleotide at the 3' end of the influenza virus neuraminidase vRNA is involved in temporal regulation of transcription and replication of neuraminidase RNAs and affects the repertoire of influenza virus surface antigens," *J. Gen. Virol.*, vol. 79, no. 8, pp. 1923–1934, 1998.

[221] P. J. Kranzusch *et al.*, "Assembly of a functional Machupo virus polymerase complex," *Proc. Natl. Acad. Sci.*, vol. 107, no. 46, pp. 20069–20074, 2010.

[222] A. Kohl, E. F. Dunn, A. C. Lowen, and R. M. Elliott, "Complementarity, sequence and structural elements within the 3' and 5' non-coding regions of the Bunyamwera orthobunyavirus S segment determine promoter strength," *J. Gen. Virol.*, vol. 85, no. 11, pp. 3269–3278, 2004.

[223] J. N. Barr and G. W. Wertz, "Bunyamwera Bunyavirus RNA Synthesis Requires Cooperation of 3′- and 5′-Terminal Sequences," *J. Virol.*, vol. 78, no. 3, pp. 1129–1138, 2004.

[224] M. Hass, M. Westerkofsky, S. Muller, B. Becker-Ziaja, C. Busch, and S. Gunther, "Mutational Analysis of the Lassa Virus Promoter," *J. Virol.*, vol. 80, no. 24, pp. 12414–12419, 2006.