Co-evolving pairs of complementary nucleotide sequence regions as candidates for bundling signals in viruses with segmented genomes

Yoshiyuki Suzuki

Graduate School of Natural Sciences, Nagoya City University, 1 Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan

ARTICLE INFO

ABSTRACT

Some of RNA viruses with segmented genomes are believed to undergo selective packaging, in which one copy of each genomic segment is incorporated into a virion. The selective packaging is thought to be mediated by supramolecular complex formation of genomic RNA segments through inter-segmental base pairing of complementary nucleotide sequence regions termed the bundling signal. Here the nucleotide sequences of 10 genomic segments (large: L1, L2, and L3; medium: M1, M2, and M3; and small: S1, S2, S3, and S4) for 29 avian (ARV) and 8 pteropine (PRV) orthoreovirus isolates were analyzed to identify co-evolving pairs of complementary nucleotide sequence regions within and between genomic segments. The co-evolving pairs within and between genomic segments were predicted to be involved in intra-segmental and inter-segmental interactions, respectively. One co-evolving pair each was identified within M1 and S3, and both pairs indeed tended to constitute stems in the intra-segmental RNA secondary structures. In addition, one co-evolving pair each was identified between S3 and S4, L2 and S4, and L3 and M1, and the co-evolving pair between S3 and S4 tended to constitute loops in the intra-segmental RNA secondary structures. The co-evolving pair identified between S3 and S4 was considered as a candidate for the bundling signal in ARV and PRV.

1. Introduction

Genomes of RNA viruses belonging to the families *Arenaviridae*, *Birnaviridae*, *Bromoviridae*, *Bunyaviridae*, *Chrosoviridae*, *Closteroviridae*, *Cystoviridae*, *Orthomyxoviridae*, *Partitiviridae*, *Picobirnaviridae*, and *Reoviridae* are divided into segments (McDonald et al. 2016). By virtue of the segmented nature of the genome, these viruses can generate mutants by means of exchanging genomic segments upon co-infection of a cell, called the reassortment (McDonald et al. 2016). In fact, the species is defined as a group of isolates that can reassort with each other in these viruses (Attoui et al. 2011; McCauley et al. 2011).

Among the RNA viruses with segmented genomes, those belonging to the families *Orthomyxoviridae* and *Reoviridae* possess highly divided genomes comprising 6–8 and 9–12 segments, respectively (Carstens 2011; Condit 2013). These viruses are believed to undergo selective packaging, in which one copy of each genomic segment is incorporated into a virion (Hutchinson et al. 2010; McDonald and Patton 2011). The mechanism of selective packaging is not fully understood. However, it has been postulated that genomic segments of RNA are bundled as a supra-molecular complex through inter-segmental interaction of complementary nucleotide sequence regions, termed the bundling signal (Goto et al. 2013), and that intra-segmental RNA secondary structures facilitate the interaction (Suzuki and Kobayashi 2013; Kobayashi et al. 2016).

Generally, functions of molecules such as RNAs and proteins are exerted through formation of secondary and tertiary structures. Nucleotide or amino acid sequences may evolve neutrally or nearly neutrally as long as the structures of molecules are conserved (Kimura, 1983, 1985). Consequently, structures are sometimes more useful than sequences for clarifying distant evolutionary relationships among molecules (Nasir and Caetano-Anolles 2015; Seligmann and Raoult 2016). Formation of structures in RNAs and proteins is mediated by interactions of nucleotides and amino acids, respectively. Interacting nucleotides and amino acids may co-evolve under the constraint for maintaining structures. Therefore, interacting nucleotides and amino acids may be identified from the analysis of co-evolution (Lin et al., 2012; Rashidipour et al. 2016; Marouzi et al. 2017; Skwark et al. 2017).

The pair of complementary nucleotide sequence regions constituting the bundling signal may co-evolve through compensatory mutations under the constraint for maintaining complementarity (Kimura 1983, 1985). Co-evolution in the bundling signal may suppress reassortment of wild-type and mutant genomic segments and may eventually promote speciation (Hutchinson et al. 2010; McDonald and Patton 2011). Therefore, a bundling signal may be represented as a pair of complementary nucleotide sequence regions between genomic segments with a signature of co-evolution during speciation (Boyce and

McCrae 2015; Boyce et al. 2016).

Avian (ARV) and pteropine (PRV) orthoreoviruses are closely related in the family *Reoviridae* but do not appear to reassort with each other (Attoui et al. 2011; Farkas et al. 2016). ARV is an etiological agent of various diseases including arthritis and synovitis in birds worldwide (Day 2009). PRV is an emerging and zoonotic agent causing respiratory diseases in humans transmitted from the reservoir of fruit bats in Southeast Asia (Tan et al. 2017). These viruses possess 10 genomic segments of RNA, which are classified according to the size; i.e., large (L1, L2, and L3), medium (M1, M2, and M3), and small (S1, S2, S3, and S4). It has been reported that supra-molecular complex formation of genomic segments in reoviruses is initiated with bundling of small genomic segments (Sung and Roy 2014; Fajardo et al. 2015, 2017). The purpose of the present study was to search for candidates of bundling signals in ARV and PRV from the analysis of genomic sequences.

2. Materials and methods

2.1. Sequence data

The complete genomic sequences of 29 ARV and 8 PRV isolates were retrieved from the International Nucleotide Sequence Database (INSD) on July 22, 2017 (Supplementary Table S1). Multiple alignment for the total of 37 nucleotide sequences was made for each of 10 genomic segments using the computer program MAFFT (version 7.305b) (Katoh et al. 2002). The 10 multiple alignments obtained were analyzed for identifying the co-evolving pairs of complementary nucleotide sequence regions as described below.

2.2. Data analysis

From the comparison of 10 multiple alignments, pairs of mono-nucleotide positions that were complementary in each of ARV and PRV isolates were identified within and between genomic segments. Here the nucleotide positions containing a gap in any of ARV and PRV isolates were eliminated from the analysis, and only the Watson-Crick base pairs of U and A, and C and G were regarded as complementary, though U and G may form a wobble base pair, to reduce false-positives in prediction of co-evolving pairs of complementary nucleotide sequence regions that were functionally relevant. Among the pairs of complementary mono-nucleotide positions identified above, co-varying pairs were defined as the pairs in which the bases in ARV isolates were different from those in PRV isolates. The proportion of co-varying pairs among all pairs of complementary mono-nucleotide positions was designated p_{cov}.

In the above analysis, pairs of complementary mono-nucleotide positions were identified regardless of whether they were involved in the direct interaction or not. Therefore, the pairs identified above should consist of not only the interacting pairs but also the non-interacting pairs that were complementary by chance and may interact with different mono-nucleotide positions from each other. Compensatory mutations that do not disrupt complementary are more likely to be neutral or nearly neutral and cause co-evolution in the interacting pairs than in the non-interacting pairs. Consequently, the proportion of co-varying pairs among the interacting pairs is expected to be greater than that among the non-interacting pairs, and p_{cov} may represent an intermediate value of these proportions by the law of large numbers. Therefore, a pair of complementary nucleotide sequence regions with the length of 2 or more bases including a greater number of co-varying pairs of complementary mono-nucleotide positions than expected from p_{cov} may be co-evolving and thus interacting (Lin et al., 2012; Rashidipour et al. 2016; Marouzi et al. 2017; Skwark et al. 2017).

Based on this rationale, 10 multiple alignments obtained above were further analyzed for identifying pairs of complementary nucleotide sequence regions with the length of 2 or more bases within and between genomic segments. For each pair with the length of X bases, a

statistical test was conducted to examine whether the number of co-varying pairs of complementary mono-nucleotide positions included (Y) was significantly greater than that expected from p_{cov}. Specifically, the binomial probability for the occurrence of Y or more events in X trials was computed with the parameter value of p_{cov}. Correction for multiple testing was conducted with the total number of the pairs with the length of X bases by the methods of Bonferroni (Sokal and Rohlf 1995) and Benjamini and Hochberg (1995) for obtaining 5% family-wise false-positive rate (FW-FPR) and 5% false-discovery rate (FDR), respectively.

3. Results

3.1. Co-evolving pairs of complementary nucleotide sequence regions

Lack of reassortment between ARV and PRV isolates was confirmed by the phylogenetic analysis for each of 10 genomic segments; ARV and PRV isolates always formed distinct clusters in the phylogenetic tree (data not shown). The length of the pairs of complementary nucleotide sequence regions identified within and between genomic segments (Supplementary Table S1) ranged from 1 to 6 bases (Table 1). It should be noted that the pairs listed in Table 1 overlap; i.e., a pair with the length of n bases includes ($n - i + 1$) pairs with the length of i bases ($1 \leq i \leq n$). The p_{cov} value was obtained to be 0.0337 (= 332,720/9,883,373). For each pair of complementary nucleotide sequence regions with a particular length, a statistical test of co-evolution was conducted with the correction for multiple testing. For example, for the pairs with the length of 5 bases, 38 statistical tests were conducted with the correction (Table 1). Overall, 4 pairs of complementary nucleotide sequence regions with the length of 5 or 6 bases were judged as co-evolving with the 5% FW-FPR, and additional 9 pairs with the length of 3 to 6 bases were judged as co-evolving with the 5% FDR (Table 1). The nucleotide sequences of ARV and PRV isolates at the 13 co-evolving pairs are presented in Table 2, in which the positions are numbered according to the isolate S1133 for ARV (Teng et al. 2014) and the isolate Talikud-74 for PRV (Taniguchi et al. 2017). After eliminating overlaps, 5 co-evolving pairs remained; one co-evolving pair each with 6 bases long within M1, 5 bases long within S3, 4 bases long between S3 and S4, 3 bases long between L2 and S4, and 3 bases long between L3 and M1 (Table 2).

It has been reported that the supra-molecular complex formation of genomic segments in reoviruses is initiated with bundling of small genomic segments (Sung and Roy 2014; Fajardo et al. 2015, 2017). Therefore, only the small genomic segments (S1, S2, S3, and S4) were

Table 1
Numbers of pairs of complementary nucleotide sequence regions in ARV and PRV isolates.

		Y^d					
	X^c	4	3	2	1	0	Total
All[a]	6	0[e]	*1*	0	0	0	1
	5	**1**	*2*	0	1	34	38
	4	0	*5*	1	6	467	479
	3	N.A.[f]	*4*	14	86	3455	3559
	2	N.A.	N.A.	2213	6018	446,829	455,060
	1	N.A.	N.A.	N.A.	332,720	9,530,653	9,883,373
Small[b]	5	**1**	0	0	0	5	6
	4	0	*3*	0	2	35	40
	3	N.A.	1	6	11	274	292
	2	N.A.	N.A.	108	317	12,173	12,598
	1	N.A.	N.A.	N.A.	15,574	262,064	277,638

[a] All of 10 genomic segments were used for the analysis.
[b] Only 4 small genomic segments (S1, S2, S3, and S4) were used for the analysis.
[c] Length of co-evolving pairs of complementary nucleotide sequence regions.
[d] Number of co-varying pairs of complementary mono-nucleotide positions.
[e] Numbers of pairs of complementary nucleotide sequence regions judged as co-evolving with 5% FW-FPR and 5% FDR are bold-faced and italicized, respectively.
[f] Not applicable.