Adenine·cytosine substitutions are an alternative pathway of compensatory mutation in angiosperm ITS2

XINWAN ZHANG,^{1,3} YONG CAO,^{1,3} WEI ZHANG,^{1,2} and MARK P. SIMMONS²

¹Marine College, Shandong University, Weihai 264209, China

²Department of Biology, Colorado State University, Fort Collins, Colorado 80523, USA

ABSTRACT

Compensatory mutations are crucial for functional RNA because they maintain RNA configuration and thus function. Compensatory mutation has traditionally been considered to be a two-step substitution through the GU-base-pair intermediate. We tested for an alternative AC-mediated compensatory mutation (ACCM). We investigated ACCMs by using a comprehensive sampling of ribosomal internal transcribed spacer 2 (ITS2) from 3934 angiosperm species in 80 genera and 55 families. We predicted ITS2 consensus secondary structures by using LocARNA for structure-based alignment and partitioning paired and unpaired regions. We examined and compared the substitution rates and frequencies among base pairs by using RNA-specific models. Base-pair states of ACCMs were mapped onto the inferred phylogenetic trees to infer their evolution. All types of compensatory mutations involving the AC intermediate were observed, but the most frequent substitutions were with AU or GC pairs, which are part of the AU–AC–GC pathway. Compared with the GU intermediate, AC had a lower frequency and higher mutability. Within the AU–AC–GC pathway, the AU–AC substitution rate was much slower than the AC–GC substitution rate. No consistently higher overall rate was identified for either pathway among all 80 sampled lineages, though compensatory mutations through the AC intermediate averaged about half that through the GU intermediate. These results demonstrate an alternative compensatory mutation between AU and GC that helps address the controversial inference of inferred simultaneous double substitutions.

Keywords: base-pair substitution; compensatory mutations; ribosomal ITS2; secondary structure

INTRODUCTION

The functions of RNA molecules are based on their secondary and tertiary structures, which are formed by intramolecular base pairs (Côté et al. 2002; Pendrak and Roberts 2011; Woolford and Baserga 2013). Any mutation occurring on one side of a base pair could cause RNA conformational changes and thus is expected to be deleterious. Compensatory mutation or compensatory base change (CBC), wherein a substitution on one side of a base pair is compensated by a substitution on the other side of that base pair, is expected to restore base-pairing (Kimura 1985; Chen et al. 1999; Ivankov et al. 2014). In the context of a fitness landscape, the intermediate variant is inferred to represent a deep valley between adaptive peaks (Kimura 1985; Kirby et al. 1995; Meer et al. 2010). Identifying the intermediate base pair prior to the compensatory mutation is necessary to understand RNA evolution and adaptation (Meer et al. 2010; Kusumi et al. 2016).

The GU/UG base pair has a special status among the 12 possible noncanonical combinations (Masquida and Westhof 2000; Varani and McClain 2000; Mokdad et al. 2006; Ananth et al. 2013). It is generally recognized that the frequency and stability of the GU/UG pair are less than Watson–Crick (WC) base pairs, but still higher than all other noncanonical base pairs (Masquida and Westhof 2000; Savill et al. 2001; Ananth et al. 2013). Therefore RNA-substitution models always include the GU/UG pair to account for this compensatory mutation process (e.g., AU → GU → GC; Rousset et al. 1991; Tillier and Collins 1995; Savill et al. 2001). In contrast, the AC/CA pair, which is the other purine–pyrimidine intermediate, has rarely been assessed (Meer et al. 2010; Marin Rodrigues et al. 2017).

Crystal-structure analyses indicate that the two characteristics which make UG pairs relatively stable, hydrogen bonds and translation of the pyrimidine and purine toward

³These authors contributed equally to this work.

Corresponding author: wzhang@sdu.edu.cn

Article is online at <http://www.rnajournal.org/cgi/doi/10.1261/rna.072660.119>.

© 2020 Zhang et al. This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

major groove and minor grooves, respectively, are both shared with the C·A⁺ base pair (Jang et al. 1998; Pan et al. 1998; Brovarets and Hovorun 2016). In some RNA functional sites, C·A⁺ pairs can be used in place of U·G pairs because of their structural similarity (Gabriel et al. 1996; Pan and Sundaralingam 1999; Masquida and Westhof 2000). This type of C·A⁺ pair (the other being tautomeric C·A pair; Wang et al. 2011), which is formed by protonation and is stable in a weak acidic environment, has been repeatedly observed by X-ray crystallography (Hunter et al. 1986; Jang et al. 1998; Garg and Heinemann 2018) and nuclear-magnetic resonance (NMR; Boulard et al. 1992; Durant and Davis 1999; Huppler et al. 2002) in both chemically synthesized oligoribonucleotides as well as cleavage sites of an *in vivo* lead-zyme, the *Bacillus stearothermophilus* DNA polymerase I large fragment, and human interleukin (IL)-6 mRNA. The C·A pair can also form a cognate-base-pair conformation in the polymerase active site (Wang et al. 2011; Brovarets and Hovorun 2016). These detailed biochemical findings for AC/CA structures contrast with our relative ignorance of their evolutionary relevance.

Detection of unstable base pairs remains problematic in practice. X-ray crystallography or NMR spectroscopy remain indispensable for determining the base-pair composition of crystal structures (e.g., Wang et al. 2011; Brovarets and Hovorun 2016; Garg and Heinemann 2018). But these approaches have limitations when applied to the CA/AC unstable pair because it is both infrequent and transient, and hence seldom observed. Furthermore, the formation of hydrogen bonds between noncanonical base pairs, either through protonation or tautomerization, is dependent upon the crystal environment, which makes it difficult to visualize in the crystal structure (Wang et al. 2011; Kimsey et al. 2015). Finally, determining structures larger than tRNA remains difficult. Recently, a more sophisticated NMR-relaxation-dispersion method has been developed to facilitate the visualization of transient Watson–Crick-like mismatches in DNA and RNA, but it is restricted to GT and GU base pairs (Kimsey et al. 2015).

As an alternative to biochemical-detection methods, nucleotide-substitution models infer the substitution process and are not limited to canonical-base-pair states. By assuming that the substitution process is constant within a given lineage (Yang 1994; Jow et al. 2002), equilibrium frequency and rate parameters can be defined, including substitutions involving transient AC/CA base pairs. Here, we applied the 16-state RNA model, which accounts for all base-pair combinations (Allen and Whelan 2014), to the ribosomal RNA region internal transcribed spacer 2 (ITS2), which is a widely used phylogenetic marker in plants (Álvarez and Wendel 2003; Chen et al. 2010; Qin et al. 2017; Li et al. 2019), in order to determine if AC/CA has a similar role as GU/UG in RNA compensatory mutations, as shown in Figure 1.

FIGURE 1. A schematic representation of substitution processes involving compensatory mutations. The base-pair substitutions in black are existing pathways proposed by Tillier and Collins (1998); the base-pair substitutions in red are proposed in this study. The thickness of the solid arrows and their shading indicate the substitution rates inferred from our data sets. The dashed arrows indicate the simultaneous double-substitution pathway that was not supported by our best-fit 16D/E/F/J model for 70 of the 80 lineages. Letter(s) on the arrows represent substitution rates between base pairs.

RESULTS

Characteristics of AC-base-pair substitutions from the best-fit double-substitution rate matrices

The best-fit single-substitution models for 70 of our 80 study lineages were RNA16D (67 lineages), RNA16E (1 lineage), RNA16F (1 lineage), and RNA16J (1 lineage); none of which allows for simultaneous substitutions of both nucleotides in a base pair (Supplemental Table S1). However, the 16D/E/F single-substitution models are subject to an artifact wherein the total probability of change between AU (UA) and GC (CG) is identical for both the GU (UG) and AC (CA) intermediates (i.e., $\alpha_n \cdot \beta_n$ is always equal to $\mu_n \cdot \gamma_n$; Supplemental Fig. S1). Given that this artifact of the 16D/E/F models is negatively determinate to the rates that we are interested in, we instead used double (simultaneous) substitution models to estimate rates (Supplemental Table S2).

The best-fit double-substitution models for these 80 lineages were 16C (61 lineages) and 16A (19 lineages). These models assign each base pair an equilibrium frequency and substitution rate with the other base pairs. By summing across the substitution rates for each base pair we obtained the net rate of change from one base pair to the others (“mutability”).

There are eight possible CBC pathways, which can be divided into four groups based on the same WC base-pair change, albeit through different intermediates (e.g., CG can change into UA through CA or UG; Table 1; Fig. 1). Our averaged results for the 80 lineages indicate that substitution rates from these intermediates to WC pairs differ greatly but are generally proportional within each

TABLE 1. Comparison of average substitution elements between intermediate base pairs across the 80 best-fit double-substitution rate matrices among all eight possible compensatory substitutions

Pathway ^a	Frequency of intermediate base pair	Mutability of intermediate base pair	Rate from WC base pair to intermediate base pair	Rate from intermediate base pair to another WC pair	Overall rate	Rate ratio
AU → AC → GC	0.0111	3.2122	0.0701 (μ_1)	1.8800 (γ_1)	0.1367 ($\mu_1 * \gamma_1$)	56.6390 (γ_1 / μ_1)
AU → GU → GC	0.0786	1.6389	0.2828 (α_1)	0.9064 (β_1)	0.3066 ($\alpha_1 * \beta_1$)	3.6924 (β_1 / α_1)
UA → CA → CG	0.0107	3.8632	0.0688 (μ_2)	2.5713 (γ_2)	0.1896 ($\mu_2 * \gamma_2$)	49.9194 (γ_2 / μ_2)
UA → UG → CG	0.0755	1.8846	0.2496 (α_2)	1.2627 (β_2)	0.3332 ($\alpha_2 * \beta_2$)	5.8868 (β_2 / α_2)
GC → AC → AU	0.0111	3.2122	0.0721 (μ_3)	1.0103 (γ_3)	0.0747 ($\mu_3 * \gamma_3$)	28.5807 (γ_3 / μ_3)
GC → GU → AU	0.0786	1.6389	0.2642 (α_3)	0.4273 (β_3)	0.1393 ($\alpha_3 * \beta_3$)	1.8215 (β_3 / α_3)
CG → CA → UA	0.0107	3.8632	0.0668 (μ_4)	0.9760 (γ_4)	0.0684 ($\mu_4 * \gamma_4$)	22.5078 (γ_4 / μ_4)
CG → UG → UA	0.0755	1.8846	0.2569 (α_4)	0.4426 (β_4)	0.1264 ($\alpha_4 * \beta_4$)	2.009 (β_4 / α_4)
Average	0.0109 (AC/CA) 0.0771 (GU/UG)	3.5377 (AC/CA) 1.7618 (GU/UG)	0.0695 (μ) 0.2634 (α)	1.6094 (γ) 0.7598 (β)	0.1174 ($\mu * \gamma$) 0.2264 ($\alpha * \beta$)	39.4117 (γ / μ) 3.3524 (β / α)

^aThe eight possible pathways can be divided into four groups based on the same Watson–Crick (WC) pair changes through different intermediate base pairs. The substitution elements of each group are marked by the different subscript numbers used in Figure 1.

pathway. For example, the rate of AC → GC (γ_1 , 1.8800; Table 1) is very different from its counterpart of CA → CG (γ_2 , 2.5713). But they are both approximately twice the rates of GU → GC (β_1 , 0.9064) and UG → CG (β_2 , 1.2627), respectively, in each of these two CBC pathways. Given that other rates and frequencies were similar among the four groups (Table 1; Fig. 2; Supplemental Fig. S2), we chose to focus on one group (AU → GC) to demonstrate these common characteristics.

Taken across all 80 lineages with the best-fit double-substitution models, the average AC frequency was only 14% that of the average GU frequency (0.0111 AC vs. 0.0786 GU; Table 1; Supplemental Table S2). However, the AC mutability was higher than the GU mutability for 63/80 lineages, averaging 196% that of GU (3.2122 AC vs. 1.6389 GU; Table 1; Supplemental Table S2). The lower frequency but higher mutability of AC is largely distinct from the higher frequency and lower mutability of GU (Fig. 2A). Taken together, AC → GC and AC → AU account for 90.0% of the six possible one-site base pair changes from AC (1.8800 GC, 1.0103 AU, 0.0771 CC, 0.0817 AG, 0.0860 UC, 0.0770 AA; Fig. 2B). The high mutability of AC (Fig. 2A) and this substitution bias (Fig. 2B) indicate

an additional pathway of compensatory mutation involving an AC intermediate despite the low frequency of AC base pairs (Figs. 1, 2A).

FIGURE 2. Comparisons of substitution rates and base-pair frequencies between AC (CA) and GU (UG) intermediate base pairs in CBCs. All results are shown based on the best-fit double-substitution rate matrix. (A) Frequency-mutability scatter plot of AC and GU base pairs and their chemical structures. (B) Box-plots for single-substitution rates from AC to six other base pairs. (C) Comparisons of rate heterogeneity between AC- and GU-mediated CBCs. The two substitution-rate ratios for AC-mediated CBCs are generally higher than those for GU-mediated CBCs. (D) Comparisons of the total probabilities of change from AU (UA) to GC (CG) based on multiplying the pairs of single-substitution rates. $\alpha_1 * \beta_1$ and $\mu_1 * \gamma_1$ represent the probability of GU- and AC-mediated CBCs, respectively (Fig. 1).

We compared the compensatory substitution rates of AU→GC between the two alternative compensatory-mutation pathways (Fig. 1) and found that the average mutation rate from AU to GU (α_1 ; 0.2828, Table 1; Supplemental Table S2) was 403% higher than that for AU to AC (μ_1 ; 0.0701). But the average mutation rate from GU to GC (β_1 ; 0.9064) was 52% lower than that for AC to GC (γ_1 ; 1.8800, Table 1; Supplemental Table S2). For each compensatory mutation, the substitution rate between AU and the intermediate base pair (AC or GU) was always lower than that between the intermediate base pair and GC (AC or GU→GC; $\mu_1 < \gamma_1$, $\alpha_1 < \beta_1$; Table 1; Fig. 1; Supplemental Table S2). We found that in AC-mediated CBCs, γ was always higher while μ was always lower than their GU-mediated counterparts β and α , respectively. Thus $\gamma_1/\mu_1 > \beta_1/\alpha_1$ in 75/80 lineages (15.3-fold on average; Fig. 2C, the four exceptions [*Asarum*, *Fraxinus*, *Rhodiola* and *Viburnum*], for which $\mu_1 = 0$, are not shown in Fig. 2C; Supplemental Table S2). Again, these substitution rates are consistent with our inference that AC-mediated CBCs are a distinct process from GU-mediated CBCs.

The total probabilities of change from AU to GC based on the double-substitution models are shown in Table 1, Supplemental Table S2, and Figure 2D. Only three lineages have equal probabilities for $\alpha^*\beta$ and $\mu^*\gamma$. Of the remaining 77 lineages, 25 had higher probabilities of AC-mediated CBCs and 52 had higher probabilities of GU-mediated CBCs, indicating no consistently higher rate for either pathway. Taken across all eight CBC pathways, we observed that the total probabilities of CBC change via AC and CA intermediate base pairs was about half of those through GU and UG intermediates (0.1174 $\mu^*\gamma$ vs. 0.2264 $\alpha^*\beta$; Table 1).

Examination of CBCs among lineages

We recorded the number of base pairs involved in compensatory mutations to help assess the power of our model-based rate estimates. Compensatory base changes (CBCs) were inferred at 353 base pairs among 65 of the 80 sampled lineages (Supplemental Table S3). This count is based on those base-pair sites that have both AU (UA) and GC (CG) as well as either UA (AU) or CG (GC) present. From one to 30 CBCs were inferred for each of these 65 lineages. For example, 30 base-pair positions with CBCs were inferred among the 405 sampled species of *Allium*, whereas a single base-pair position with a CBC was inferred among the nine species of *Circaea* sampled (Supplemental Table S3). Among these 353 base-pair positions, 26 also included the AC (for positions with AU and GC) or CA (for positions with UA and CG) intermediate state but not GU or UG, whereas 189 positions included the GU or UG intermediate state but not AC or CA. These counts correspond with our model-parameter

estimates, both of which indicate that the AC/CA intermediate is less frequent than the GU/UG intermediate. Of the remaining 138 positions, 101 included both AC or CA and GU or UG.

We counted the occurrence of hemi-CBCs (change at only one of the two base-pair positions) for the remaining 15 non-CBC lineages (Supplemental Table S4). We inferred that 60 AC/CA-mediated hemi-CBCs occurred based on those base-pair sites that have at least one of the following four base-pair combinations: AC–AU, AC–GC, CA–UA, and CA–CG. Likewise, we inferred 138 GU/UG-mediated hemi-CBCs. These counts of hemi-CBCs are consistent with our CBC results and indicate that these 15 non-CBC lineages can also provide useful information in estimating CBC rates. On average fewer species were sampled in non-CBC lineages, and taxon undersampling may account for the lack of inferred CBCs in these 15 lineages.

Comparison of CBCs and model-rate estimates between LocARNA and 4SALE alignments

The consensus secondary structure generated separately by LocARNA and 4SALE delimited similar stem and CBC positions for our three test genera: a total 40 CBC positions of 206 stem positions by LocARNA versus 34 of 187 positions by 4SALE (Supplemental Table S5). Among these CBC positions, 39 included an intermediate base pair in the LocARNA alignments and 31 in the 4SALE alignments. When attention is restricted to those positions for which only one intermediate base pair is included, a total 16 GU/UG-mediated CBCs are inferred in both pairs of alignments, whereas the sum of AC/CA-mediated CBCs dropped from six to four, including zeroes for *Nepenthes* and *Potentilla*. We further compared AC frequency and substitution rates from AC to the six other single-substitution base pairs between the LocARNA and 4SALE alignments (Supplemental Table S6). The AC frequency increased in *Hypericum* but decreased in *Nepenthes* and *Potentilla*. In five of six cases, the AC→GC and AC→AU substitution rates were lower for the 4SALE alignments, but in every case these were the two highest of the six substitution rates from AC, both of which are part of the GC–AC–AU pathway.

Taken across all eight possible compensatory-substitution pathways in these three genera, the average AC/CA frequency is identical in both alignments (0.0110; Supplemental Table S7), whereas there is a decrease in mutability for 4SALE alignments (1.6012 vs. 1.3130). The average overall rates of the AC/CA intermediate also decreased from the LocARNA alignment to the 4SALE alignment (0.0434 vs. 0.0398; Supplemental Table S7). But the average overall rates relative to the GU/UG intermediate in each of these two alignments are similar (0.18; 0.0434/0.2467 vs. 0.0398/0.2267). Taken together, our results

based on the LocARNA and 4SALE alignments are similar, and both support the alternative AC pathway.

Optimization of AC/CA-mediated CBCs

Both model-parameter estimates and our counts of base-pair positions for which CBCs are inferred indicate the occurrence of AC-mediated CBCs. To complement these results, we also optimized potential AC/CA-mediated CBCs on the inferred gene trees. Despite the instability of the AC/CA base pairs, our dense taxon and specimen sampling enabled us to directly identify four unambiguous

AC/CA-mediated CBCs among the 26 base-pair sites including AU–AC–GC or UA–CA–CG (Supplemental Table S3). Each of these four CBC sites was observed in a different lineage. In three of the four cases (Fig. 3A–C), specimens with the AC/CA intermediates were resolved as a paraphyletic group separating taxa with the stable base pairs, and Fitch optimization of the AC/CA intermediate was unambiguous. In the fourth case (Fig. 3D), the clade of specimens with base pair CA is sister to the clade of specimens with the base pair CG, and Fitch optimization is ambiguous but consistent with the UA–CA–CG pathway. These four cases are unlikely to be sequencing artifacts

FIGURE 3. Optimization of AC/CA-mediated CBCs on ITS2 gene trees for four lineages. AC-mediated CBCs are inferred in (A) lineage *Aralia* and (B) lineage *Astilbe*; CA-mediated CBCs are inferred in (C) lineage *Meconopsis* and (D) lineage *Celastrus*. Base-pair states involved in CBCs are indicated using different colors for the applicable branches and species. Species names in bold font indicate the outgroup, which was used to root each tree. Branches with ≥ 0.5 posterior probabilities for Bayesian inference are highlighted in bold. Numbers following a species name indicate GenBank accession numbers.

because the AC/CA intermediate was observed in numerous specimens and at least three species. We identified all four possible AC/CA-mediated CBCs (Fig. 1). In *Aralia* we inferred a GC → AC → AU CBC at position 145 (stem IV; Fig. 3A). In *Astilbe* we inferred a reverse AU → AC → GC CBC at position 17 (stem I; Fig. 3B). In *Meconopsis* we inferred a CG → CA → UA CBC at position 80 (stem II, Fig. 3C). In *Celastrus*, we inferred a reverse UA → CA → CG CBC at position 98 (stem III; Fig. 3D). All four of these cases of AC/CA-mediated CBCs support the CBC species concept (Wolf et al. 2013), which was proposed based on GU-mediated CBCs, because none of the species were inferred to retain both stable base pairs of a given CBC.

DISCUSSION

An accurate description of RNA evolution is important for both systematic and functional biology. In this study, we inferred the transient noncanonical base pairs by using model-parameter estimates and investigated the base-pair substitution process in RNA stems. The GU base pair has traditionally been regarded as the intermediate of the two-step compensatory mutation from AU to GC (Rousset et al. 1991). Our results demonstrate that AC is another intermediate, thereby expanding the model of compensatory mutations (Tillier and Collins 1995, 1998).

It has traditionally been assumed that since the noncanonical base pairs are rare, they probably provide little phylogenetic signal (Tillier and Collins 1998; Higgs 2000; Savill et al. 2001). Our results demonstrate that compensatory mutations occur on RNA stems mainly through GU or AC intermediates, thereby indicating their importance in RNA evolution (Fig. 1). Rousset et al. (1991) found that substitutions between AU and GC always occurred through a GU intermediate in the rapidly evolving D1 and D2 domains of *Drosophila* large subunit (LSU) rRNA, which was the basis for most of the GU-mediated-compensatory-substitution models (Schöniger and Von Haeseler 1994; Muse 1995; Rzhetsky 1995; Tillier and Collins 1995, 1998; Higgs 1998). However, other authors using slowly evolving SSU rRNA sampled from the rRNA database (<http://bioinformatics.psb.ugent.be/webtools/rRNA/>; Van de Peer et al. 1998) found that the rate of simultaneous substitutions at both positions is greater than the rate of single substitutions through the GU intermediate (Tillier and Collins 1998; Savill et al. 2001). These authors inferred that some compensatory substitutions between WC base pairs have not necessarily gone through the GU intermediate. This inference is consistent with our findings of an alternative AC-mediated CBC, for which the overall rate is roughly half that of GU-mediated CBCs ($0.1174 \mu^*\gamma$, $0.2264 \alpha^*\beta$; Table 1). Simultaneous substitutions were also inferred, albeit at a still lower rate (Fig. 1; $0.088 r_{ds}$; data not shown). Given the high mutability of AC, AC (CA) intermediates are transient and frequently unob-

served. Hence we propose that AC-mediated CBCs are frequently misinterpreted as simultaneous double substitutions, artificially increasing the inferred rate of double substitutions and therefore the fit of double-substitution models (Tillier and Collins 1998). This proposal helps reconcile the controversial issue of interpreting double substitutions in rate matrices (Savill et al. 2001).

Since base change on one side of a pair often decreases or increases the stability of the stem, which is deleterious or advantageous to RNA function (Kimura 1985; Chen and Stephan 2003; Meer et al. 2010), a concern is how selective constraint acts on different base changes of a compensatory mutation. We found that the relatively stable GU intermediate had a higher frequency and that its substitute rates differed only slightly during the CBC process (Fig. 1). In contrast, the unstable AC intermediate is consistently rare and the substitution rates varied substantially during the two-step process, indicating that AC is under different selective pressures than GU. These observations are consistent with Kimura's (1985) hypothesis that deleterious intermediates can be maintained at low frequency but cannot reach fixation until another complementary mutation occurs. In the context of GU and AC, we infer that the GU intermediate can persist longer than the AC intermediate until a compensatory mutation occurs, and that both of these alternatives frequently occur in angiosperm ITS2. The generality of our ITS2-based results and inferences should be tested in other RNA regions.

The function of RNA molecules depends on their secondary and tertiary structures, which are determined by their base-pair interactions. A better understanding of the base-pair-change process is thus important for elucidating the mechanisms of RNA functional evolution. Both our model-parameter estimates and base-pair counts among diverse flowering-plant lineages indicate that the AC (CA) base pair is an alternative intermediate to the conventional GU (UG) base pair. These two intermediate pathways have different substitution rates that are consistent with their different levels of stability. Although the occurrence of AC-mediated compensatory mutations is roughly half that of GU, incorporation of the AC intermediate helps address the controversial inference of simultaneous double substitutions.

MATERIALS AND METHODS

Taxon sampling and sequence acquisition

We sampled lineages for which ITS2 sequences are available from closely related species as well as multiple specimens per species in most cases in order to sample low-frequency base pairs. We focused on empirical studies that presented plant DNA barcodes because they generally have sufficient inter- and intraspecific sampling to achieve effective species identification (Hebert

et al. 2004; www.ibol.org). For the study lineages, all sequences with the annotation “internal transcribed spacer 2” or “internal transcribed spacer” were retrieved from GenBank between September 2018 and February 2019. Incomplete ITS2 sequences were excluded. We sampled most of the major flowering-plant lineages (Chase et al. 2016). We assessed species coverage and validity of species names according to the Plant List (<http://www.theplantlist.org/>). A total of 80 lineages, each consisting of a single genus, subgenus, or species group, were sampled. Taken across all 80 lineages, 3945 species and 13,741 sequences were sampled. These lineages are distributed among 55 families and 35 orders (Supplemental Fig. S3; Supplemental Table S8; Chase et al. 2016).

Alignment partition and phylogenetic analyses using RNA-specific models

ITS2 boundaries were identified by using GenBank annotations or hidden Markov models implemented in the ITS2 Ribosomal RNA Database (<http://its2.bioapps.biozentrum.uni-wuerzburg.de/>; Ankenbrand et al. 2015). We aligned the raw sequences using the MAFFT auto-select strategy (Katoh and Standley 2013) and then imported the sequence alignments into LocARNA 1.9.2.1 (Supplemental Fig. S4A,B; Will et al. 2007, 2012). LocARNA was used for consensus-secondary-structure annotation for each of the 80 lineages independently of each other. The consensus secondary structure generated by LocARNA was viewed and refined by using PseudoViewer3 (Byun and Han 2009) according to the conserved “four-helix model” of ITS2 in Eukaryota (Coleman 2003; Schultz et al. 2005). We partitioned the sequence alignment into unpaired and paired regions according to their consensus secondary structure (Supplemental Fig. S5), and then selected among conventional DNA models and RNA-specific models with a Perl script (model_selection.pl) from PHASE package 3.0 (Allen and Whelan 2014). This Perl script selects among two DNA models (HKY85 and REV), seven RNA 7-state models, and nine RNA 16-state models. We used Allen and Whelan’s (2014) likelihood-correction method to facilitate comparison between the 4-, 7-, and 16-state models.

To test whether our LocARNA alignments are robust, we also applied an alternative alignment method that uses a 12×12 scoring matrix to three genera for which the greatest potential number of AC-mediated CBCs (as measured by the sixth and tenth columns of Supplemental Table S3) were inferred based on the LocARNA alignments (*Hypericum*, *Nepenthes*, and *Potentilla*; the 405 sampled species of *Allium* were excluded because we were unable to obtain a consensus secondary structure from 4SALE). With 4SALE, sequence-structure data are coded for each of the four nucleotides with three structural states (paired left, paired right or unpaired), wherein the scoring function for substitutions and gap costs are estimated specifically for ITS2 (Wolf et al. 2014). We performed this alignment according to Schultz and Wolf’s (2009) tutorial as follows. First, the Vienna format of each ITS2 secondary structure was predicted by homology modeling from the online ITS2 Database (Koetschan et al. 2012; Ankenbrand et al. 2015). Second, sequences with homologous structures were synchronously aligned by using 4SALE 1.7 (Seibel et al. 2006, 2008). Third, the graphical form of the consensus secondary structure was generated after manual refinement using the 4SALE editor.

Fourth, the 50% majority consensus secondary structure was transformed manually into Vienna format for subsequent analyses as described above for the LocARNA alignment.

Gene-tree analyses for each of the 80 separate lineages were performed using the PHASE 3.0 package (Jow et al. 2002; Hudelot et al. 2003; Allen and Whelan 2014) with the best-fit mixed model (HKY85 or REV for unpaired regions and a 16-state RNA-base-pair model for paired regions; Allen and Whelan 2014). Bayesian MCMC phylogenetic inference was performed using the mcmcPHASE program (Allen and Whelan 2014). Two independent runs that each consisted of four MCMC chains were run for one million generations each, sampling every 100 generations, with a burn-in of 3000 (30%) trees. The mcmcsummarize program in the PHASE package was then used to calculate the majority-rule-consensus topology and posterior probabilities. These trees were viewed by using FigTree 1.5.4 (<http://tree.bio.ed.ac.uk/software/figtree/>) and edited by using the interactive tree of life (iTOL: <https://itol.embl.de/>). When applicable, the inferred trees were rooted using the same outgroups that the original authors applied in their previous DNA barcoding or phylogenetic studies. For the remaining 15 lineages, we added two or three outgroups from closely related lineages. For each data set, base-pair frequencies and substitution-rate parameter values were generated using the mcmcsummarize program from the PHASE package. Statistical analyses were then performed to summarize these results using SPSS 22.0 (SPSS).

Inferring the compensatory substitution process

To infer CBCs, both positions of a base pair should be considered at the same time. To facilitate this inference, we transformed all paired bases into alternative symbols as follows. Base-pair information was included in the sequence-structure alignment generated from LocARNA (Supplemental Fig. S4B; Will et al. 2007, 2012). Each base pair was treated as a unit and transformed into a 28-symbol coding matrix using RNastat (Supplemental Fig. S4C1; Subbotin et al. 2007). To supplement our model-based results, we optimized potential AC-mediated compensatory mutations on the inferred trees by using Fitch (parsimony; Fitch 1971) optimization. We did so for aligned positions that contained one of the following four base-pair combinations: F–D–M (representing AU–AC–GC), P–H–J, M–D–F, or J–H–P in the 28-symbol matrix (Supplemental Fig. S4C).

Supplementary data, including ITS2 sequence matrices of both secondary-structure guided and base-pair transformed alignments, gene trees from both the best-fit models (16C/D/E/F/J) and the best-fit double-substitution models (16A/C), optimization of AC/CA-mediated CBCs on nine gene trees, Microsoft Excel files containing the raw and summarized data of rate matrices from both the best-fit models and the best-fit double-substitution models, and supplemental data for an alternative alignment method that uses a 12×12 scoring matrix for three genera are posted at: https://figshare.com/articles/Supplemental_data_for_Adenine_cytosine_substitutions_are_an_alternative_pathway_of_compensatory_mutation_in_angiosperm_ITS2_/8285447.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank the editor and two anonymous reviewers for several constructive criticisms with which to improve the manuscript. This work was supported by the National Natural Science Foundation of China (grant 81673551) and Young Scholars Program of Shandong University, Weihai (2017WHWLJH05).

Received July 18, 2019; accepted November 14, 2019.

REFERENCES

- Allen JE, Whelan S. 2014. Assessing the state of substitution models describing noncoding RNA evolution. *Genome Biol Evol* **6**: 65–75. doi:10.1093/gbe/evt206
- Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol Phylogenet Evol* **29**: 417–434. doi:10.1016/S1055-7903(03)00208-2
- Ananth P, Goldsmith G, Yathindra N. 2013. An innate twist between Crick's wobble and Watson–Crick base pairs. *RNA* **19**: 1038–1053. doi:10.1261/ma.036905.112
- Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F. 2015. ITS2 database V: twice as much. *Mol Biol Evol* **32**: 3030–3032. doi:10.1093/molbev/msv174
- Boulard Y, Cognet JA, Gabarro-Arpa J, Le Bret M, Sowers LC, Fazakerley GV. 1992. The pH dependent configurations of the CA mispair in DNA. *Nucleic Acids Res* **20**: 1933–1941. doi:10.1093/nar/20.8.1933
- Brovarets OO, Hovorun DM. 2016. By how many tautomerisation routes the Watson–Crick-like A:C* DNA base mispair is linked with the wobble mismatches? A QM/QTAIM vision from a biological point of view. *Struct Chem* **27**: 119–131. doi:10.1007/s11224-015-0687-4
- Byun Y, Han K. 2009. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics* **25**: 1435–1437. doi:10.1093/bioinformatics/btp252
- Chase MW, Christenhusz MJ, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc* **181**: 1–20. doi:10.1111/boj.12385
- Chen Y, Stephan W. 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster* *Adh* gene. *Proc Natl Acad Sci* **100**: 11499–11504. doi:10.1073/pnas.1932834100
- Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, Stephan W. 1999. RNA secondary structure and compensatory evolution. *Genes Genet Syst* **74**: 271–286. doi:10.1266/ggs.74.271
- Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, et al. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* **5**: e8613. doi:10.1371/journal.pone.0008613
- Coleman AW. 2003. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet* **19**: 370–375. doi:10.1016/S0168-9525(03)00118-5
- Côté CA, Greer CL, Peculis BA. 2002. Dynamic conformational model for the role of ITS2 in pre-rRNA processing in yeast. *RNA* **8**: 786–797. doi:10.1017/S1355838202023063
- Durant PC, Davis DR. 1999. Stabilization of the anticodon stem-loop of tRNA^{Lys,3} by an A⁺-C base-pair and by pseudouridine. *J Mol Biol* **285**: 115–131. doi:10.1006/jmbi.1998.2297
- Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Biol* **20**: 406–416. doi:10.1093/sysbio/20.4.406
- Gabriel K, Schneider J, McClain WH. 1996. Functional evidence for indirect recognition of G-U in tRNA^{Ala} by alanyl-tRNA synthetase. *Science* **271**: 195–197. doi:10.1126/science.271.5246.195
- Garg A, Heinemann U. 2018. A novel form of RNA double helix based on G-U and C-A⁺ wobble base pairing. *RNA* **24**: 209–218. doi:10.1261/ma.064048.117
- Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc Natl Acad Sci* **101**: 14812–14817. doi:10.1073/pnas.0406166101
- Higgs PG. 1998. Compensatory neutral mutations and the evolution of RNA. *Genetica* **102**: 91–101. doi:10.1023/A:1017059530664
- Higgs PG. 2000. RNA secondary structure: physical and computational aspects. *Q Rev Biophys* **33**: 199–253. doi:10.1017/S0033583500003620
- Hudlot C, Gowri-Shankar V, Jow H, Rattray M, Higgs PG. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol Phylogenet Evol* **28**: 241–252. doi:10.1016/S1055-7903(03)00061-7
- Hunter WN, Brown T, Anand NN, Kennard O. 1986. Structure of an adenine cytosine base pair in DNA and its implications for mismatch repair. *Nature* **320**: 552. doi:10.1038/320552a0
- Huppler A, Nikstad LJ, Allmann AM, Brow DA, Butcher SE. 2002. Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. *Nat Struct Mol Biol* **9**: 431–435. doi:10.1038/nsb800
- Ivanov DN, Finkelstein AV, Kondrashov FA. 2014. A structural perspective of compensatory evolution. *Curr Opin Struct Biol* **26**: 104–112. doi:10.1016/j.sbi.2014.05.004
- Jang SB, Hung LW, Chi YI, Holbrook EL, Carter RJ, Holbrook SR. 1998. Structure of an RNA internal loop consisting of tandem C-A⁺ base pairs. *Biochemistry* **37**: 11726–11731. doi:10.1021/bi980758j
- Jow H, Hudlot C, Rattray M, Higgs PG. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol Biol Evol* **19**: 1591–1601. doi:10.1093/oxfordjournals.molbev.a004221
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780. doi:10.1093/molbev/mst010
- Kimsey IJ, Petzold K, Sathyamoorthy B, Stein ZW, Al-Hashimi HM. 2015. Visualizing transient Watson–Crick-like mispairs in DNA and RNA duplexes. *Nature* **519**: 315–320. doi:10.1038/nature14227
- Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. *J Genet* **64**: 7–19. doi:10.1007/BF02923549
- Kirby DA, Muse SV, Stephan W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci* **92**: 9047–9051. doi:10.1073/pnas.92.20.9047
- Koetschan C, Hackl T, Müller T, Wolf M, Förster F, Schultz J. 2012. ITS2 database IV: interactive taxon sampling for internal transcribed spacer 2 based phylogenies. *Mol Phylogenet Evol* **63**: 585–588. doi:10.1016/j.ympev.2012.01.026
- Kusumi J, Ichinose M, Takefu M, Piskol R, Stephan W, Iizuka M. 2016. A model of compensatory molecular evolution involving multiple sites in RNA molecules. *J Theor Biol* **388**: 96–107. doi:10.1016/j.jtbi.2015.10.008
- Li M, Zhao H, Zhao F, Jiang L, Peng H, Zhang W, Simmons M P. 2019. Alternative analyses of compensatory base changes in an ITS2 phylogeny of *Corydalis* (Papaveraceae). *Ann Bot* **124**: 233–243. doi:10.1093/aob/mcz062

- Marin Rodrigues MV, Müller T, Buchheim MA, Canbäck B, Wolf M. 2017. THESES db: the algae 18S rDNA sequence–structure database for inferring phylogenies. *Phycologia* **56**: 186–192. doi:10.2216/16-71.1
- Masquida B, Westhof E. 2000. On the wobble GoU and related pairs. *RNA* **6**: 9–15. doi:10.1017/S1355838200992082
- Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **464**: 279–283. doi:10.1038/nature08691
- Mokdad A, Krasovska MV, Sponer J, Leontis NB. 2006. Structural and evolutionary classification of G/U wobble basepairs in the ribosome. *Nucleic Acids Res* **34**: 1326–1341. doi:10.1093/nar/gkl025
- Muse SV. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics* **139**: 1429–1439.
- Pan B, Sundaralingam M. 1999. Mismatched base pairing in RNA crystal structures. *Int J Quantum Chem* **75**: 275–287. doi:10.1002/(SICI)1097-461X(1999)75:3<275::AID-QUA17>3.0.CO;2-J
- Pan B, Mitra SN, Sundaralingam M. 1998. Structure of a 16-mer RNA duplex r(GCAGACUAAAUCUGC)₂ with wobble C:A⁺ mismatches. *J Mol Biol* **283**: 977–984. doi:10.1006/jmbi.1998.2140
- Pendrak ML, Roberts DD. 2011. Ribosomal RNA processing in *Candida albicans*. *RNA* **17**: 2235–2248. doi:10.1261/ma.028050.111
- Qin Y, Li M, Cao Y, Gao Y, Zhang W. 2017. Molecular thresholds of ITS2 and their implications for molecular evolution and species identification in seed plants. *Sci Rep-UK* **7**: 17316. doi:10.1038/s41598-017-17695-2
- Rousset F, Pelandakis M, Solignac M. 1991. Evolution of compensatory substitutions through GU intermediate state in *Drosophila* rRNA. *P Natl Acad Sci* **88**: 10032–10036. doi:10.1073/pnas.88.22.10032
- Rzhetsky A. 1995. Estimating substitution rates in ribosomal RNA genes. *Genetics* **141**: 771–783.
- Savill NJ, Hoyle DC, Higgs PG. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* **157**: 399–411.
- Schöniger M, Von Haeseler A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol Phylogenet Evol* **3**: 240–247. doi:10.1006/mpev.1994.1026
- Schultz J, Wolf M. 2009. ITS2 sequence–structure analysis in phylogenetics: a how-to manual for molecular systematics. *Mol Phylogenet Evol* **52**: 520–523. doi:10.1016/j.ympev.2009.01.008
- Schultz J, Maisel S, Gerlach D, Müller T, Wolf M. 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA* **11**: 361–364. doi:10.1261/rna.7204505
- Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M. 2006. 4SALE—a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics* **7**: 498. doi:10.1186/1471-2105-7-498
- Seibel PN, Müller T, Dandekar T, Wolf M. 2008. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Res Notes* **1**: 91. doi:10.1186/1756-0500-1-91
- Subbotin SA, Sturhan D, Vovlas N, Castillo P, Tambe JT, Moens M, Baldwin JG. 2007. Application of the secondary structure model of rRNA for phylogeny: D2–D3 expansion segments of the LSU gene of plant-parasitic nematodes from the family Hoplolaimidae Filipjev, 1934. *Mol Phylogenet Evol* **43**: 881–890. doi:10.1016/j.ympev.2006.09.019
- Tillier ER, Collins RA. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Mol Biol Evol* **12**: 7–15. doi:10.1093/oxfordjournals.molbev.a040195
- Tillier ER, Collins RA. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**: 1993–2002.
- Van de Peer Y, Caers A, De Rijk P, De Wachter R. 1998. Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Res* **26**: 179–182. doi:10.1093/nar/26.1.179
- Varani G, McClain WH. 2000. The G-U wobble base pair: a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep* **1**: 18–23. doi:10.1093/embo-reports/kvd001
- Wang W, Hellinga HW, Beese LS. 2011. Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc Natl Acad Sci* **108**: 17644–17648. doi:10.1073/pnas.1114496108
- Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3**: e65. doi:10.1371/journal.pcbi.0030065
- Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. 2012. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**: 900–914. doi:10.1261/ma.029041.111
- Wolf M, Chen S, Song J, Ankenbrand M, Müller T. 2013. Compensatory base changes in ITS2 secondary structures correlate with the biological species concept despite intragenomic variability in ITS2 sequences—a proof of concept. *PLoS One* **8**: e66726. doi:10.1371/journal.pone.0066726
- Wolf M, Koetschan C, Mueller T. 2014. ITS2, 18S, 16S or any other RNA—simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. *Gene* **546**: 145–149. doi:10.1016/j.gene.2014.05.065
- Woolford JL, Baserga SJ. 2013. Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics* **195**: 643–681. doi:10.1534/genetics.113.153197
- Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J Mol Evol* **39**: 105–111. doi:10.1007/BF00178256

RNA

A PUBLICATION OF THE RNA SOCIETY

Adenine-cytosine substitutions are an alternative pathway of compensatory mutation in angiosperm ITS2

Xinwan Zhang, Yong Cao, Wei Zhang, et al.

RNA 2020 26: 209-217 originally published online November 20, 2019

Access the most recent version at doi:[10.1261/rna.072660.119](https://doi.org/10.1261/rna.072660.119)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2019/11/20/rna.072660.119.DC1>

References

This article cites 62 articles, 19 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/26/2/209.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>