Author(s): Thien Ho, Liang Wang, Linfeng Huang, Zhigang Li, Denise W. Pallett, Tamas Dalmay, Kazusato Ohshima, John A. Walsh and Hui Wang

Article Title: Nucleotide bias of DCL and AGO in plant anti-virus gene silencing
Year of publication: 2010
Link to published article:
http://dx.doi.org/10.1007/s13238-010-0100-4
Publisher statement: The original publication is available at www.springerlink.com

Nucleotide bias of DCL and AGO in plant anti-virus gene silencing

Thien Ho[1,2§], Liang Wang[3], Linfeng Huang[1§], Zhigang Li[1], Denise W. Pallett[1], Tamas Dalmay[4], Kazusato Ohshima[5], John A. Walsh[6], Hui Wang[1*]

[1] NERC/Centre for Ecology and Hydrology (CEH) Wallingford, Maclean Building, Benson Lane, Wallingford, Oxfordshire OX10 8BB, UK

[2] Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

[3] Beijing Institute of Genomics, Chinese Academy of Sciences, Beitucheng West Road, Beijing 100029, China

[4] School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

[5] Laboratory of Plant Virology, Faculty of Agriculture, Saga University, 1-banchi, Honjo-machi, Saga 840-8502, Japan

[6] Plant-Virus Interactions Group, Warwick HRI, Warwick University, Wellesbourne, Warwick CV35 9EF, UK

*Corresponding author: Hui Wang. Mailing address: NERC/Centre for Ecology and Hydrology (CEH) Wallingford, Maclean Building, Benson Lane, Wallingford, Oxfordshire OX10 8BB, UK. Phone: +44-(0)1491-838800. Fax: 44-(0)1491-692424. Email: huw@ceh.ac.uk

[§] Current address: LH, Immune Disease Institute, 200 Longwood Avenue, Boston MA 02115, USA; TH, AF Building, University of Dundee at SCRI, Invergowrie, Dundee DD2 5DA, UK

Running title: Nucleotide bias in plant anti-virus gene silencing

Abstract

Plant Dicer-like (DCL) and Argonaute (AGO) are the key enzymes involved in anti-virus post-transcriptional gene silencing (AV-PTGS). Here we show that AV-PTGS exhibited nucleotide preference by calculating a relative AV-PTGS efficiency on processing viral RNA substrates. In comparison with genome sequences of dicot-infecting *Turnip mosaic virus* (TuMV), and monocot-infecting *Cocksfoot streak virus* (CSV), viral-derived small interfering (vsi)RNAs displayed positive correlations between AV-PTGS efficiency and G+C content (GC%). Further investigations on nucleotide contents revealed that the vsiRNA populations had G-biases. This finding was further supported by our analyses of previously reported vsiRNA populations in diverse plant-virus associations, and AGO associated *Arabidopsis* endogenous siRNA populations, indicating that plant AGOs operated with G-preference. We further propose a hypothesis that AV-PTGS imposes selection pressure(s) on the evolution of plant viruses. This hypothesis was supported when potyvirus genomes were analysed for evidence of GC elimination, suggesting that plant virus evolution to have low GC% genomes would have a unique function which is to reduce the host AV-PTGS attack during infections.

Key words: anti-virus post-transcriptional gene silencing, siRNA, nucleotide bias, Dicer-like, Argonaute, plant virus evolution

Introduction

During plant post-transcriptional gene silencing (PTGS), Dicer-like (DCL) RNase-III enzymes cleave RNA molecules with double-stranded (ds) features, producing ds-small interfering RNAs (ds-siRNA) (Ding and Voinnet, 2007; Jinek and Doudna, 2009; Mlotshwa et al., 2008). One strand of the DCL product (guide strand) is incorporated to the PAZ domain of Argonaute (AGO) protein to form a component of the so-called RNA-induced silencing complex (RISC) that interferes with mRNAs based on complementary homology to the guide strand siRNA (e.g. reviewed in (Hock and Meister, 2008; Hutvagner and Simard, 2008; Jinek and Doudna, 2009; Vaucheret, 2008; Zhang et al., 2008). Meanwhile, the other strand (the passenger strand) of the siRNA is cleaved by the RNase-H-like PIWI domain of AGO (Matranga et al., 2005; Tomari et al., 2004). During plant virus infections, viral RNA triggers plant production of virus-derived (v)siRNAs (e.g., reviewed in (Ding and Voinnet, 2007; Mlotshwa et al., 2008) and thereby becomes a direct target of plant antivirus (AV-)PTGS.

Plant viruses have evolved a variety of silencing suppressor proteins that inhibit host AV-PTGS (reviewed by (Burgyan, 2008), indicating that AV-PTGS imposes significant selection pressure on plant virus evolution; suppression of AV-PTGS is advantageous for virus replication and spread. On the other hand, gene silencing derived selection pressure has also been highlighted by escaping viral mutants that emerge during RNA interfering (RNAi) treatments against human viruses (e.g. reviewed in (Grimm and Kay, 2007; Watanabe et al., 2007; Yamamoto and Tsunetsugu-Yokota, 2008) and in mosquito vector (Brackney et al., 2009). Although a high degree of genome polymorphism is a hallmark of plant virus

populations (Elena et al., 2008), mutagenesis has not been considered as a viable viral strategy against plant AV-PTGS. One of the notions is that plant AV-PTGS produces vsiRNAs that target viral sequences at multiple hotspots throughout a virus genome, making the virus impossible to accumulate effective site mutations to escape host AV-PTGS. Indeed, any nucleotide position in a virus genome may be subject to plant AV-PTGS attack (Donaire et al., 2009).

Populations of vsiRNAs can either consist of equal proportions of both plus (sense) and minus (anti-sense) species or be dominated by species originating from the viral plus strand. The former represented a pathway in which ds-RNAs were processed (reviewed in (Aliyari and Ding, 2009), whereas the latter suggested an alternative pathway in which the plus single-stranded (+ss) viral RNAs were predominately targeted (e.g. (Donaire et al., 2009; Ho et al., 2006; Molnar et al., 2005; Qi et al., 2009). However, the +ssRNA targeting scenario appeared increasingly contradictory to observations that viral RNAs at many vsiRNA hotspots lacked detectable fold-back structures, suggesting a possibility of unknown vsiRNA production mechanism (e.g.(Donaire et al., 2008; Donaire et al., 2009; Du et al., 2007; Qi et al., 2009).

Both eudicot and monocot AV-PTGS displayed G+C (GC) preference when producing vsiRNAs against potyvirus (genus *Potyvirus*, family *Potyviridae*) infections (Ho et al., 2008; Ho et al., 2007). A recent report on vsiRNA populations of nine plant viruses further confirmed the GC enrichment feature (Donaire et al., 2009). Here, analysing vsiRNA populations generated by deep sequencing, we further characterised the nucleotide bias during AV-PTGS, including the catalysing steps by DCL (ds-RNA processing) and AGO (guiding strand selection). Furthermore, as plant

AV-PTGS operates with nucleotide bias, we propose a hypothesis that plant viruses may have evolved genome composition biases opposite to the AV-PTGS preference, so as to reduce the severity of AV-PTGS attack during infections. According to the equilibrium theory on genome compositional bias, existing biases are the results of balances made among mutation, selection, and drift (e.g. (Vetsigian and Goldenfeld, 2009; Yang and Nielsen, 2008). Because AV-PTGS determines the survival of viral RNAs in the cytoplasm, this ancient mechanism may have played an important role in genome compositions during virus evolution.

Results

The efficiency of vsiRNA production positively correlated to GC content of potyviral RNA substrate

Potyviruses comprise about 20% of the known plant viruses. A potyvirus has a +ssRNA genome in which a single open reading frame encoding the viral polyprotein is flanked by the 5'- and 3'-untranslated regions (UTR). Small RNA populations of *Brassica juncea* (family Brassicaceae) leaves infected by *Turnip mosaic virus* (TuMV, genus *Potyvirus*, family *Potyviridae*), and *Dactylis glomerata* (family Poaceae) leaves infected by *Cocksfoot streak virus* (CSV, genus *Potyvirus*, family *Potyviridae*) were obtained by high-throughput pyrophosphate sequencing (known as 454 sequencing) (Fahlgren et al., 2007; Rajagopalan et al., 2006). In total 41,647 TuMV vsiRNAs (GEO accession number GSE12053) and 5,631 CSV vsiRNAs (GEO accession number GSE12052) of 15-29 nt long, with 100% match to the TuMV (GBR98, GenBank accession number, EU861593) and CSV (GenBank accession

number, EU119422) genome sequences, were used for further analyses. Both populations were dominated by 21-nt species (>50%, Supporting Figure 1) and had hotspots for both polarities (Supporting Figure 2) similar to that obtained by conventional small RNA cloning and sequencing (Ho et al., 2008; Ho et al., 2007). The potyvirus vsiRNAs originated from both plus and minus strands (TuMV: 50.3% plus polarity, n=20,944, 49.7% minus polarity, n=20,703; CSV: 56.1% plus polarity, n=3,159, 43.9% minus polarity, n=2,472), also similarly as reported previously (Donaire et al., 2009; Ho et al., 2008; Ho et al., 2007). Screening for reverse and complementary vsiRNAs with 2-nt 3'-overhangs revealed only 1,272 possible pairs (TuMV: n=1,238, 2.97% of total; CSV: n=34, 0.60% of total), indicating that the vast majority of the vsiRNA sequences were isolated as ss-vsiRNAs, most likely being the guiding strands recruited and stabilized by the AGO complexes (Hutvagner and Simard, 2008).

To compare the compositional profiles of vsiRNA with those of the viral genomes, complete sets of viral substrate (vsub)RNAs (to PTGS enzymes) were generated *in silico* by using sliding windows of 21, 22, and 24-nt in length, and from both plus and minus polarities, respectively. These vsubRNAs represented the theoretical vsiRNA population based on random vsiRNA production. Obtained vsiRNA populations by sequencing shifted to higher GC% distributions when compared to vsubRNA populations (Figure 1A&B). When a Relative Substrate Efficiency (RSE) of the PTGS machinery was calculated (RSE = vsiRNA% / vsubRNA%, Y-axis, Figure 1A&B) for each GC% category (X-axis, Figure 1A&B), the GC bias was further evident as positive correlations between RSE and GC% (Figure 1). It appeared that both eudicot and monocot plants could not effectively produce or accumulate vsiRNA when GC%<30, shown as RSE close to zero (Figure

1C&D). With 30<GC%<60, two phases of linear correlations were observed. Firstly, when 30<GC%<45, RSE was positively correlated to GC% with factors of 1.9 (RSE=1.92GC-0.62, R^2=0.90, P<0.001) and 0.7 (RSE=0.73GC-0.24, R^2=0.84, P<0.005), for TuMV and CSV, respectively (Figure 1E). Secondly, when 40<GC%<60, RSE increased with GC% by factors of 15.2 (RSE=15.24GC-6.22, R^2=0.92, P<0.001) and 33.3 (RSE=33.27GC-15.28, R^2=0.96, P<0.001), for TuMV and CSV, respectively (Figure 1F). Such divergence may suggest a difference between eudicot and monocot PTGS, and it indicated that the observed correlations were not artefacts introduced during experimentation because a systemic error should be consistent. It appeared that GC%>50 would be needed for achieving RSE≥1 (Figure 1F), a sufficient silencing response against potyviral RNAs in plants (i.e. one proportion of a viral RNA fragment triggered plant production of at least one proportion of vsiRNA). On the other hand, reduction of GC% to <45 would help the viral RNAs to escape from plant PTGS. When GC%>60, greater variations were observed (Figure 1 C&D). It has been implied that too high GC% may affect RISC loading, cleavage, and product release (Pei and Tuschl, 2006).

Plant AGO exhibited G-preference

Based on the Watson-Crick base-pairing rule, the data shown in Figure 1 might suggest that DCLs prefer stable dsRNA substrates (Pei and Tuschl, 2006). However, when compositions of vsiRNAs and the viral genomes were compared, G-enrichments were clearly evident as positive correlations between RSE (logarithm) and G% (Figure 2 A-D). The other nucleotides did not have such a relationship to RSE (Figure 2 A-D), suggesting that G% played a unique role in AV-PTGS, most likely in the AGO selection of guiding strand siRNA from the DCL products.

When the nucleotide content at each vsiRNA position was calculated, G-preference was displayed throughout the entire middle positions (Figure 3C, F, I, L). The nucleotide ratios (vsiRNA% / vsubRNA% = product / substrate) for G were >1 (Figure 3C, F, I, L), indicating that the G over C bias in vsiRNA was not a result of compositional bias in the virus genomes. When the detected vsiRNA strand had a G residue in a certain position, its theoretical complementary strand (the passenger strand) had a C. Thus the G over C biases (Figure 3C, F, I, L) also represented G-enrichments in the detected vsiRNAs compared to their passenger strands. All analysed vsiRNA populations displayed GC bias at terminal positions (Figure 3A, D, G, and M). At the 5'-end, A was discriminated against (Figure 3B, E, H and N), whereas C was preferred (Figure 3C, F, I and L). At the 3'-end, U was discriminated against (Figure 3B, E, H, K), whereas G (Figure 3C, F and I) or GC (Figure 3L) was preferred. The 5'-G avoidance was consistent with previous reports (Donaire et al., 2009; Navarro et al., 2009; Qi et al., 2009) and a 5'-C preference had also been reported for siRNAs against grapevine viroids (Navarro et al., 2009). To determine whether or not the vsiRNA profile may be relevant to the 5'-nucleotide mediated AGO sorting mechanism (Mi et al., 2008; Montgomery et al., 2008; Takeda et al., 2008), populations of vsiRNAs were further divided to 5'-end A-, U-, G-, C-leading sub-populations. Sequence logos (http://weblogo.berkeley.edu/logo.cgi) made for these sub-populations showed that features of G-preference in the middle range positions appeared independent to the 5'-end leading nucleotide (Supporting Figure 3).

A recent report on nine plant viruses (Donaire et al., 2009) confirmed earlier work showing that GC-enrichment in vsiRNA populations is a common feature among different virus/plant associations (Ho et al., 2008; Ho et al., 2007). To

determine whether or not the G-bias (Figures 2&3) is also a common feature of vsiRNA populations, we analysed two independent vsiRNA datasets [NCBI/GEO/GSE16996 (Donaire et al., 2009), and NCBI/GEO/GSE12918 (Qi et al., 2009)]. Significant G-enrichment was evident in repeatedly sequenced vsiRNAs species (read count numbers, n>1) compared to the singletons (read count number, n=1) in *Tobacco rattle virus* (TRV, genus *Tobravirus*) and *Cucumber mosaic virus* (CMV, genus *Cucumovirus*) infections in *Arabidopsis thaliana*; *Cymbidium ringspot* (CymRSV, genus *Tobmbusvirus*), *Potato virus X virus* (PVX, genus *Potexvirus*), and *Pepper mild mottle virus* (PMMoV, genus *Tobamovirus*) infections in *Nicotiana benthamiana*; *Melon necrotic spot virus* (MNSV, genus *Carmovirus*) and *Watermelon mosaic virus* (WMV, genus *Potyvirus*) infections in *Cucumis melo*; and *Tomato yellow leaf curl virus* (TYLCV, genus *Begomovirus*) infection in *Solanum lycopersicum* (Donaire et al., 2009) (Table 1). In *Tobacco mosaic virus* (TMV-Cg strain) infections in *A. thaliana* wild type (Col-0) and RDR deficient plants (*rdr1-1*, *rdr6-15*) (Qi et al., 2009), G-enrichments were also apparent in the repeatedly sequenced vsiRNAs (Table 1). Although G-enrichment was observed in TuMV infection in *B. juncea* (Figures 2 and 3), it was not evident in TuMV infection in *Arabidopsis* (Table 1). This could be due to the small numbers of repeatedly detected vsiRNA species in the TuMV/*Arabidopsis* dataset (only 56 vsiRNAs with count numbers larger than 1, Table 1), although it might suggest that *Arabidopsis* PTGS against TuMV could be different to that of *Brassica*. Overwhelmingly, Table 1 showed that *Arabidopis* and the other plant species produced vsiRNA populations with G-enrichments, showing that the G-preference is a common and possibly ancient feature of plant AV-PTGS.

To further confirm that plant AGO operated with G-bias, we analysed previously reported *A. thaliana* endogenous siRNA populations directly isolated from AGO complexes [NCBI/GEO/GSE10036 (Mi et al., 2008), and NCBI/GEO/GSE16545 (Havecker et al., 2010), Supporting Table 1]. The siRNAs were categorised for their popularity frequencies as singleton species (read count number, n=1), species repeatedly reported 2-10 times (read count numbers, $1<n\leq10$) and species repeatedly reported more than 10 times (read count numbers, $n>10$). Nucleotide contents were compared among these 3 categories by ANOVA (Minitab15) and *t*-Test (Excel, Microsoft Office 2007). For all tested AGOs, significant GC-enrichment (ANOVA, $P<0.001$) appeared in the repeatedly sequenced populations compared to singleton species, showing increased PTGS affinity to GC-rich targets in *Arabidopsis* endogenous siRNA production/accumulation (Figure 4A, and Supporting Fig. 4A). Populations of siRNA associated with different *Arabidopsis* AGOs displayed significant differences (ANOVA, $P<0.001$) in their GC% (Figure 4A), showing that GC% of the sequenced siRNAs was sample dependent rather than fixed by the sequencing methodology. G-enrichment (ANOVA, $P<0.001$) was also evident in the repeatedly sequenced siRNAs (Figure 4B, and Supporting Fig. 4B). Populations of AGO1, AGO2, AGO4, AGO6, and AGO9 associated siRNAs displayed G (Figure 4B, and Supporting Fig. 4B) over C (Figure 4C, and Supporting Fig. 4C) biases, whereas AGO5 associated siRNAs had C (Figure 4C) over G (Figure 4B) biases in all the three frequency categories. However, significant G-enrichment was also displayed in the repeatedly sequenced AGO5-siRNAs compared to singletons (Figure 4B), indicating that AGO5 also operated with G-preference during the guiding strand selection.

Finally, to confirm that the plant AGO associated nucleotide bias is not due to possible systemic biases that may be generated by high throughput sequencing (Linsen et al., 2009), *Drosophila melanogaster* AGO associated siRNA populations (NCBI/GEO/GSE11086, GSM280087 and GSM280088, Supporting Table 1) (Czech et al., 2008) were analysed. No trend in nucleotide preference could be established in the insect system (Supporting Figure 5), showing that the G-bias (Figures 2-4) was unique to the plant system and was unlikely to have been generated by the high throughput sequencing. All independent datasets generated by conventional small RNA cloning and sequencing (Ho et al., 2008; Ho et al., 2007), 454 pyrosequencing (Figures 1-3, Supporting Fig. 4, and Table 1) and Illumina Solexa sequencing (Figure 4, Supporting Fig. 4, and Table 1 for TMV-Cg) showed conforming results. The data on *Arabidopsis* endogenous siRNAs provided direct evidence strongly supporting the observation on G-bias of plant AGOs during AV-PTGS (Figures 2&3).

Evidence of selective pressure on low GC content in the *Potyviridae* genomes

If the nucleotide biases of plant PTGS enzymes are significant in anti-virus function, they would have impacted on virus evolution. According to Figure 1, it would be logical to suggest that plant viruses had evolved to contain low GC contents in their genomes to reduce the PTGS attack during infection. Based on 953 NCBI reference genomes (Supporting Table 2, virus genome sequences with NCBI accession numbers starting with NC_) obtained from the GenBank, plant virus genomes indeed had GC% of 43.5±0.2 (Mean ± S.E., n=953), significantly lower (*P*=0.000, paired *t*-Test) than AU(T)% of 56.5±0.2. *Potyviridae* (the largest family of plant viruses and the most represented in GenBank) genomes had GC% of 42.4±0.2 (n=69, Supporting Table 3), significantly lower (*P*=0.000, paired *t*-Test) than AU% of

57.6±0.2. We then used the *Potyviridae* genome sequences to test if the low GC% may be due to selection pressure or mutational bias (e.g., reviewed by (Hershberg and Petrov, 2008).

A *Potyviridae* +ssRNA genome is composed of a 5'-UTR (untranslated region), a single open reading frame encoding the viral polyprotein, and a 3'-UTR. In the *Potyviridae* genomes, the GC% at the 5'-UTR (33.9±0.8), and 3'-UTR (41.0±0.6) were both significantly lower than that of the coding sequences (CDS, 42.6±0.2, $P<0.005$, *t*-Test) (Figure 5A). GC% at the 1st, 2nd, and 3rd codon positions (GC1, GC2, and GC3) was calculated and plotted against the total GC% of the CDS for each virus genome (Figure 5B). Positive correlations ($P<0.001$, Regression Analysis, MiniTab15) were detected for all codon positions, indicating that all codon positions contributed to the GC elimination (from the assumed theoretical null point, GC%=50) in CDS. However, the correlation efficiency of GC3 (K=1.91) was more than two times greater than those of GC1 (K=0.68) and GC2 (K=0.42) (Figure 5B). The acceleration of GC elimination in GC3 indicated a selection pressure on low GC%. This is because mutagenesis occurs in all codon positions at the same rate, but the 3rd codon position is the most likely place for mutants to survive due to synonymous nucleotide substitutions. Low GC% at UTRs (Figure 5A) and the dynamics of GC elimination in CDS (Figure 5B) supported the hypothesis that *Potyviridae* genomes are subject to a selection pressure for low GC%.

Discussion

Since plant PTGS determines the survival of RNAs in cytoplasma, and translation efficiency of mRNAs (Hock and Meister, 2008; Hutvagner and Simard, 2008; Mi et al., 2008; Montgomery et al., 2008; Takeda et al., 2008; Vaucheret, 2008), it is logical to propose that PTGS mediates a selection pressure (additional to transcriptional and translational factors) during evolution. According to the estimated RSE of plant AV-PTGS (Figure 1E and F), a GC% reduction from 50% to 42% reduces the AV-PTGS efficacy to about 10-fold. Under the framework of the equilibrium hypothesis, current genome compositional bias is the result of a balance made among different driving forces (e.g. (Vetsigian and Goldenfeld, 2009; Yang and Nielsen, 2008). Although our data indicates that the majority of plant viruses have evolved low GC% genomes, it is necessary to note that exceptions are evident. For example, Poaceae-infecting sobemovirus genomes display GC%>50 because the viruses adapt to host codon usages (Zhou et al., 2005). For the *Potyviridae*, molecular evolution may also be affected by other factors, e.g. transmission bottlenecks (Wang et al., 2006), and positive selection during host adaptation (Ohshima et al., 2010; Tan et al., 2005). Therefore, similarly to the strategy of PTGS suppressor proteins (Burgyan, 2008), low GC content would be one of the virus adaptations to the host PTGS system, contributing to the virus survival and evolution. However, as vsiRNAs can be detected, it appears that PTGS suppressors, low GC% genomes, or a combination of the two can not prevent vsiRNA production completely, showing the robustness of plant AV-PTGS as an ancient and fundamental anti-virus mechanism. This concept may also apply to host genome biology because small RNAs are employed to target selfish DNAs such as transposons (e.g. Havecker et al., 2010; Kasschau et al., 2007; Wei et al., 2009).

How could AGO generate nucleotide bias? It has been questioned whether or not plant PTGS machinery produces siRNAs following the same asymmetry rules as detected in animal species (Rajagopalan et al., 2006). Figures 2 and 3 showed that G was preferred over C at most middle positions in vsiRNA asymmetry. This seems to support a hypothesis that flexibility of the RNA-binding PAZ domain may contribute to recognition of siRNA in "initial loading" to the AGO complex (Rashid et al., 2007) and further suggests that the PAZ domain of plant AGO may operate a generic preference of G over C. Alternatively, the PIWI domain (Ma et al., 2005; Parker et al., 2005) may mediate unknown catalytic bias that leads to survival of the G-rich strand. Figure 4 showed variations in nucleotide biases among *Arabidopsis* endogenous siRNAs associated with different AGOs. Preference would be given to the influence derived from protein-protein interactions because RISC complexes are composed of a wide range of protein consortiums (Hock and Meister, 2008; Hutvagner and Simard, 2008; Jinek and Doudna, 2009; Vaucheret, 2008; Zhang et al., 2008).

It remains obscure why a G-rich AGO complex might have advantages for RISC function. AGO1 is believed to be the main AGO involved in AV-PTGS (Brodersen et al., 2008). The AGO4 family incorporates 24-nt DCL3 products and mediates siRNA-induced *de novo* DNA methylation (Chan et al., 2005; Havecker et al., 2010; Matzke et al., 2007; Ruiz-Ferrer and Voinnet, 2009; Zhai et al., 2008). Because all plant AGOs analysed in Figure 4 showed G-preference, it is reasonable to assume that the G-bias may be an ancient feature in plant PTGS rather than limited in AV-PTGS. It is tempting to speculate that the G-bias may be relevant in promoting siRNA induced *de novo* DNA methylation. Indeed, G-rich endogenous siRNA populations correlated to methylcytosine enrichments at precise locations in the

Arabidopsis genome (Lister et al., 2008). Further investigation of nucleotide biases may provide information on the plant gene silencing systems outside of AV-PTSG.

Materials and methods

TuMV and CSV

TuMV (isolate GBR-98) infections in *B. juncea* (Mustard cv. Tendergreen) were established under glasshouse conditions and the infected leaves were processed as described previously (Ho et al., 2007). Leaves of perennial *D. glomerata* (cocksfoot grass) were collected from 32 wild individuals originally from the Yellow Ant Reserve in Wytham Woods, Oxfordshire, UK. Six of them were naturally infected by CSV and these individuals had been maintained in a glasshouse for more than 3 years (Ho et al., 2008). Small RNAs were isolated, ligated to 5'- and 3'-adaptors (Ho et al., 2006), and amplified by RT-PCR using primers containing the adaptor sequences and the 454 sequencing Primer-A (forward) and Primer-B (reverse). The RT-PCR products were pooled before being sequenced by 454 Life Sciences (Branford, USA).

All adaptor sequences were removed from the 454 reads before analysis. The resulting sequences were screened against TuMV (GBR98, GenBank accession number, EU861593) and CSV (CSV630wytham, GenBank accession number, EU119422) genome sequences as described before (Ho et al., 2008; Ho et al., 2007). All non-hit sequences were treated as a single background population. Among the sequences that had affinities to the viral genomes, only hits of 15-29 nt with 100% homology to the reference genomes were used for further analyses. Only 21-nt (DCL-

4 products), 22-nt (DCL-2 products), and 24-nt (DCL-3 products) (Deleris et al., 2006; Liu et al., 2007) were used for detailed nucleotide analyses.

To determine the nucleotide bias in vsiRNA populations, complete theoretical sets of the TuMV and CSV substrate (vsub)RNAs were generated *in silico* for 21, 22, & 24-nt in length with plus and minus polarities, by using sliding window size of a desired length and step size of 1-nt (BioEditor, http://bioeditor.sdsc.edu/). Each 1,000-nt fragment of the ~10,000-nt, +ssRNA viral genomes (except for the last fragments of 798-nt for TuMV and 622-nt for CSV) was analysed independently (Ho et al., 2008; Ho et al., 2007). Resulting data was represented as the Mean ± Standard Error (SE, n=10 fragments). The vsiRNA populations were also sorted into each of the 10 artificial genome fragments according to their 5'-positions. Relative substrate efficiency (RSE) of the AV-PTGS machinery was calculated as the ratio of proportion in vsiRNA population against that of vsubRNA population (RSE = vsiRNA% / vsubRNA%) for each nucleotide content category (i.e. A%, U%, G%, C%, AU%, and GC%). It was assumed that the nucleotide popularities of detected vsiRNA should match those of vsubRNA if the AV-PTGS enzymes process RNA substrates randomly without any bias. Therefore the expected RSE (the null model) is 1. Any RSE>1 indicates a bias for the AV-PTGS process, and any RSE<1 indicates a bias against AV-PTGS. Only populations of the 21-, 22-, 24-nt vsiRNAs were used.

To investigate positional bias, nucleotide contents at each position (1-24 nt, from the 5'-end) of the vsiRNAs and vsubRNAs were calculated for each of the 1000-nt TuMV and CSV genome fragment. Plus and minus polarities were calculated separately. The nucleotide ratio (NR) at each vsiRNA position was calculated as vsiRNA% / vsubRNA% for each genome fragments and represented by the Mean ±

Standard Error (n=10 fragment). Again, the expected NR is 1 under the null model, and NR>1 or NR<1 indicates preference for or against a particular nucleotide at a particular vsiRNA position. WebLogo program (http://weblogo.berkeley.edu/) (Crooks et al., 2004) was used to represent positional bias of vsiRNA populations. The CSV 22- and 24-nt vsiRNA populations had limited numbers of sequences to support robust positional analyses and therefore were not used.

vsiRNAs of other plant viruses

To investigate nucleotide bias in plant AV-PTGS against other plant viruses, two independent vsiRNA datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) site. Non-redundant profiles of *Tobacco mosaic virus* (TMV-Cg) vsiRNA profiles were downloaded from NCBI/GEO/GSE12918 (Qi et al., 2009); *Tobacco rattle virus* (TRV), *Turnip mosaic virus* (TuMV) and *Cucumber mosaic virus* (CMV) infections in *Arabidopsis thaliana*; *Cymbidium ringspot virus* (CymRSV), *Potato virus X virus* (PVX), and *Pepper mild mottle virus* (PMMoV) infections in *Nicotiana benthamiana*; *Melon necrotic spot virus* (MNSV), and *Watermelon mosaic virus* (WMV) infections in *Cucumis melo*; and *Tomato yellow leaf curl virus* infection in *Solanum lycopersicum* (NCBI/GEO/GSE16996) (Donaire et al., 2009) were analysed for G%. Each unique sequence was categorised as either singleton or multiple (read count number, n>1) vsiRNA species. G% was calculated for each unique sequence and represented as the Mean ± SE for singleton and multiples of each virus. Two-tailed homoscedastic *t*-Test (Excel) was performed to compare G% between the singleton and multiple vsiRNA populations.

To obtain nucleotide content profiles of plant viruses, 953 plant virus reference genome segments were downloaded from GenBank (NCBI accession

numbers starting with NC_, Supporting Table 2), including 69 *Potyviridae* genomes (Supporting Table 3). Nucleotide contents of each genome were calculated individually by using the BioEditor program (Bioeditor, http://bioeditor.sdsc.edu/), and Mean ± SE were represented. A paired *t*-Test was performed to compare GC% against AU(T)% among all plant viruses, and among the *Potyviridae* members, respectively. To determine selective pressure on nucleotide usages, GC% at the 5'-, 3'-UTR and protein coding sequences were calculated for each *Potyviridae* genome. GC% of the 1st, 2nd, and 3rd codon positions was also calculated for the encoding sequences and plotted against the total GC%.

AGO-associated endogenous siRNA

To determine if AGO operated with G-bias, we analysed *Arabidopsis* endogenous siRNA populations directly isolated from AGO complexes [NCBI/GEO/GSE10036 (Mi et al., 2008), and NCBI/GEO/GSE16545 (Havecker et al., 2010), Supporting Table 1]. These datasets contain siRNAs associated with AGO1, AGO2, AGO4, AGO5, AGO6, and AGO9. Non-redundant profiles were used, and the siRNAs were categorised as singleton species (read count number, n=1), species repeatedly sequenced 2-10 times (read count numbers, $1<n\leq10$), and species repeatedly sequenced more than 10 times (read count numbers, $n>10$). Nucleotide contents were compared among/between these 3 categories by ANOVA (Minitab) and *t*-Test (Excel). To determine if the nucleotide biases were unique to the plants, *Drosophila* small RNA populations isolated from AGO1 and AGO2 complexes (GEO: GSE11086, (Czech et al., 2008) were also analysed.

Acknowledgements

We are grateful to Charles Godfray and Michael Morecroft for encouragement in the early stages of this work, Nigel Fisher and M. Morecroft for help with field work on CSV infected grasses. This work was supported by the Vietnamese Studentship to TH (Ministry of Education and Training, Decision No 322/QD-TTg), NERC (UK) grants to TD (NER/A/S/2003/00547) and HW (NER/A/S/2003/00548, NE/E008933/1), and CEH Biodiversity research fund to HW (C02875).

References

Aliyari, R., and Ding, S.W. (2009). RNA-based viral immunity initiated by the Dicer family of host immune receptors. Immunol Rev 227, 176-188.

Brackney, D.E., Beane, J.E., and Ebel, G.D. (2009). RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. PLoS pathogens 5, e1000502.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science 320, 1185-1190.

Burgyan, J. (2008). Role of silencing suppressor proteins. Methods Mol Biol 451, 69-79.

Chan, S.W., Henderson, I.R., and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. Nat Rev Genet 6, 351-360.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome research 14, 1188-1190.

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J.A., Sachidanandam, R., *et al.* (2008). An endogenous small interfering RNA pathway in Drosophila. Nature 453, 798-802.

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K.D., Carrington, J.C., and Voinnet, O. (2006). Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science 313, 68-71.

Ding, S.W., and Voinnet, O. (2007). Antiviral immunity directed by small RNAs. Cell 130, 413-426.

Donaire, L., Barajas, D., Martinez-Garcia, B., Martinez-Priego, L., Pagan, I., and Llave, C. (2008). Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. J Virol 82, 5167-5177.

Donaire, L., Wang, Y., Gonzalez-Ibeas, D., Mayer, K.F., Aranda, M.A., and Llave, C. (2009). Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392, 203-214.

Du, Q.S., Duan, C.G., Zhang, Z.H., Fang, Y.Y., Fang, R.X., Xie, Q., and Guo, H.S. (2007). DCL4 targets Cucumber mosaic virus satellite RNA at novel secondary structures. J Virol 81, 9142-9151.

Elena, S.F., Agudelo-Romero, P., Carrasco, P., Codoner, F.M., Martin, S., Torres-Barcelo, C., and Sanjuan, R. (2008). Experimental evolution of plant RNA viruses. Heredity 100, 478-483.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., *et al.* (2007). High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. PLoS ONE 2, e219.

Grimm, D., and Kay, M.A. (2007). Combinatorial RNAi: a winning strategy for the race against evolving targets? Mol Ther 15, 878-888.

Havecker, E.R., Wallbridge, L.M., Hardcastle, T.J., Bush, M.S., Kelly, K.A., Dunn, R.M., Schwach, F., Doonan, J.H., and Baulcombe, D.C. (2010). The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. The Plant cell 22, 321-334.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annu Rev Genet 42, 287-299.

Ho, T., Pallett, D., Rusholme, R., Dalmay, T., and Wang, H. (2006). A simplified method for cloning of short interfering RNAs from Brassica juncea infected with Turnip mosaic potyvirus and Turnip crinkle carmovirus. Journal of virological methods 136, 217-223.

Ho, T., Rusholme Pilcher, R.L., Edwards, M.L., Cooper, I., Dalmay, T., and Wang, H. (2008). Evidence for GC preference by monocot Dicer-like proteins. Biochem Biophys Res Commun 368, 433-437.

Ho, T., Wang, H., Pallett, D., and Dalmay, T. (2007). Evidence for targeting common siRNA hotspots and GC preference by plant Dicer-like proteins. FEBS Lett 581, 3267-3272.

Hock, J., and Meister, G. (2008). The Argonaute protein family. Genome biology 9, 210.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol 9, 22-32.

Jinek, M., and Doudna, J.A. (2009). A three-dimensional view of the molecular machinery of RNA interference. Nature 457, 405-412.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol 5, e57.

Linsen, S.E., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R.K., Fritz, B., Wyman, S.K., de Bruijn, E., Voest, E.E., *et al.* (2009). Limitations and possibilities of small RNA digital gene expression profiling. Nature methods 6, 474-476.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133, 523-536.

Liu, B., Chen, Z., Song, X., Liu, C., Cui, X., Zhao, X., Fang, J., Xu, W., Zhang, H., Wang, X., *et al.* (2007). Oryza sativa dicer-like4 reveals a key role for small interfering RNA silencing in plant development. The Plant cell 19, 2705-2718.

Ma, J.B., Yuan, Y.R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. (2005). Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature 434, 666-670.

Matranga, C., Tomari, Y., Shin, C., Bartel, D.P., and Zamore, P.D. (2005). Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. Cell 123, 607-620.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A.J. (2007). Targets of RNA-directed DNA methylation. Curr Opin Plant Biol 10, 512-519.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., *et al.* (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell 133, 116-127.

Mlotshwa, S., Pruss, G.J., and Vance, V. (2008). Small RNAs in viral infection and host defense. Trends Plant Sci 13, 375-382.

Molnar, A., Csorba, T., Lakatos, L., Varallyay, E., Lacomme, C., and Burgyan, J. (2005). Plant virus-derived small interfering RNAs originate predominantly from highly structured single-stranded viral RNAs. Journal of Virology 79, 7812-7818.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008).

Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell 133, 128-141.

Navarro, B., Pantaleo, V., Gisel, A., Moxon, S., Dalmay, T., Bisztray, G., Di Serio, F., and Burgyan, J. (2009). Deep sequencing of viroid-derived small RNAs from grapevine provides new insights on the role of RNA silencing in plant-viroid interaction. PLoS One 4, e7686.

Ohshima, K., Akaishi, S., Kajiyama, H., Koga, R., and Gibbs, A.J. (2010). Evolutionary trajectory of turnip mosaic virus populations adapting to a new host. The Journal of general virology 91, 788-801.

Parker, J.S., Roe, S.M., and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature 434, 663-666.

Pei, Y., and Tuschl, T. (2006). On the art of identifying effective and specific siRNAs. Nature methods 3, 670-676.

Qi, X., Bao, F.S., and Xie, Z. (2009). Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. PLoS ONE 4, e4971.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev 20, 3407-3425.

Rashid, U.J., Paterok, D., Koglin, A., Gohlke, H., Piehler, J., and Chen, J.C. (2007). Structure of Aquifex aeolicus argonaute highlights conformational flexibility of the PAZ domain as a potential regulator of RNA-induced silencing complex function. J Biol Chem 282, 13824-13832.

Ruiz-Ferrer, V., and Voinnet, O. (2009). Roles of plant small RNAs in biotic stress responses. Annual review of plant biology 60, 485-510.

Takeda, A., Iwasaki, S., Watanabe, T., Utsumi, M., and Watanabe, Y. (2008). The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins. Plant Cell Physiol 49, 493-500.

Tan, Z., Gibbs, A.J., Tomitaka, Y., Sanchez, F., Ponz, F., and Ohshima, K. (2005). Mutations in Turnip mosaic virus genomes that have adapted to Raphanus sativus. The Journal of general virology 86, 501-510.

Tomari, Y., Du, T., Haley, B., Schwarz, D.S., Bennett, R., Cook, H.A., Koppetsch, B.S., Theurkauf, W.E., and Zamore, P.D. (2004). RISC Assembly Defects in the Drosophila RNAi Mutant armitage. Cell 116, 831-841.

Vaucheret, H. (2008). Plant ARGONAUTES. Trends Plant Sci 13, 350-358.

Vetsigian, K., and Goldenfeld, N. (2009). Genome rhetoric and the emergence of compositional bias. Proceedings of the National Academy of Sciences of the United States of America 106, 215-220.

Wang, H., Huang, L.F., and Cooper, J.I. (2006). Analyses on mutation patterns, detection of population bottlenecks, and suggestion of deleterious-compensatory evolution among members of the genus Potyvirus. Arch Virol 151, 1625-1633.

Watanabe, T., Umehara, T., and Kohara, M. (2007). Therapeutic application of RNA interference for hepatitis C virus. Adv Drug Deliv Rev 59, 1263-1276.

Wei, Y., Chen, S., Yang, P., Ma, Z., and Kang, L. (2009). Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. Genome biology 10, R6.

Yamamoto, T., and Tsunetsugu-Yokota, Y. (2008). Prospects for the therapeutic application of lentivirus-based gene therapy to HIV-1 infection. Curr Gene Ther 8, 1-8.

Yang, Z., and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25, 568-579.

Zhai, J., Liu, J., Liu, B., Li, P., Meyers, B.C., Chen, X., and Cao, X. (2008). Small RNA-directed epigenetic natural variation in Arabidopsis thaliana. PLoS Genet 4, e1000056.

Zhang, X., Segers, G.C., Sun, Q., Deng, F., and Nuss, D.L. (2008). Characterization of hypovirus-derived small RNAs generated in the chestnut blight fungus by an inducible DCL-2-dependent pathway. J Virol 82, 2613-2619.

Zhou, H., Wang, H., Huang, L.F., Naylor, M., and Clifford, P. (2005). Heterogeneity in codon usages of sobemovirus genes. Arch Virol 150, 1591-1605.

Table

Table 1. Guanine enrichment in repeatedly detected species in vsiRNA populations

	vsiRNA				*t*Test
	Singleton		Multiple		*P* value
Virus	Unique sequence No.	G% (mean ± se)	Unique sequence No.	G% (mean ± se)	Single vs Multiple
MNSV*	6596	26.1 ± 0.1	5190	29.0 ± 0.1	<0.001
TRV*	2526	28.0 ± 0.2	887	32.0 ± 0.3	<0.001
CymRSV*	2159	31.9 ± 0.2	1799	35.1 ± 0.2	<0.001
PMMoV*	1793	25.3 ± 0.2	782	27.6 ± 0.3	<0.001
CMV*	1292	30.0 ± 0.3	372	33.8 ± 0.5	<0.001
WMV*	973	23.8 ± 0.3	180	25.5 ± 0.6	<0.05
TYLCV*	514	27.5 ± 0.4	207	29.6 ± 0.5	<0.005
TuMV*	359	28.5 ± 0.6	56	26.4 ± 1.6	=0.176
PVX*	126	32.1 ± 0.7	57	36.9 ± 0.8	<0.001
TMV/wt**	1751	22.5 ± 0.2	4279	26.1 ± 0.1	<0.001
TMV/*rdr*1**	1351	24.5 ± 0.2	1390	26.6 ± 0.2	<0.001
TMV/*rdr*6**	1673	23.3 ± 0.2	3578	25.4 ± 0.1	<0.001

* Data downloaded from NCBI/GEO/GSE16996. *Tobacco rattle Tobravirus* (TRV), *Turnip mosaic Potyvirus* (TuMV) and *Cucumber mosaic Cucumovirus* (CMV) infections were in *Arabidopsis thaliana*. *Cymbidium ringspot Tombusvirus* (CymRSV), *Potato virus X Potexvirus* (PVX), and *Pepper mild mottle Tobamovirus* (PMMoV) infections were in *Nicotiana benthamiana*. *Melon necrotic spot Carmovirus* (MNSV), and *Watermelon mosaic Potyvirus* (WMV) infections were in *Cucumis melo*. And *Tomato yellow leaf curl Begomovirus* infection was in *Solanum lycopersicum* (Donaire et al., 2009).

** Data downloaded from the NCBI/GEO/GSE12918. Tobacco Mosaic Tobamovirus Cg strain (TMV-Cg) infections were made in *Arabidopsis thaliana* Col-0 (wild type, wt), *rdr1-1* (SAIL_672F11, RDR1 deficiency), and *rdr6-15* (SAIL_617H07, RDR6 deficiency) lines (Qi et al., 2009).

Figure legends

Figure 1. Correlations of relative substrate efficiency (RSE) to GC%

The mean of proportions (%, Y-axis) of each GC-content category (X-axis) were plotted for vsubRNAs generated *in silico* (grey labelled) and sequenced vsiRNAs (black labelled) for TuMV (Panel-A) and CSV (Panel-B). Marks of diamond, triangle, and square represent 21, 22, and 24-nt species, respectively. RSEs were plotted against GC-content for TuMV (Panel-C) and CSV (Panel-D). Dashed lines represent regressions of RSE against GC% of TuMV 21, 22, and 24-nt species (open marks, Panels E, F), and solid lines show those of CSV 21, 22-nt species (filled marks, Panels E, F). Error bars show the standard error (SE, n=10, each of the ~10,000-nt viral genomes were divided to 10 fragments of ~1000-nt).

Figure 2. Relationships of nucleotide contents to RSE

RSE (Y-axis, log scale) of A (green), U (red), C (blue), and G (yellow) were plotted along the nucleotide content (X-axis) for vsiRNAs of TuMV 21-nt (Panel-A), TuMV 22-nt (Panel-B), TuMV 24-nt (Panel-C), and CSV 21-nt (Panel-D). Solid lines with filled marks, and dashed lines with open marks represent vsiRNAs with plus and minus polarity, respectively. Error bars represent SE (n=10 viral genome fragments of 1000-nt).

Figure 3. Position bias of nucleotides among TuMV and CSV vsiRNAs

Mean and SE (n=10 viral genome fragments of 1000-nt) of nucleotide ratio (vsiRNA%/vsubRNA%, Y-axis) of G+C (black), A+U (grey), A (green), U (red), C (blue), and G (yellow) were plotted for each nucleotide position (X-axis) for vsiRNAs of TuMV 21-nt (Panels A-C), TuMV 22-nt (Panels D-F), TuMV 24-nt (Panels G-I),

and CSV 21-nt (Panels J-L). Filled and open labels represent vsiRNAs with plus and minus polarity, respectively.

Figure 4. G-bias in *Arabidopsis* AGO associated siRNAs

Arabidopsis endogenous siRNA populations directly isolated from AGO complexes were downloaded from GenBank (NCBI/GEO/GSE10036 and GSE16545, Supporting Table 1). GC% (Panel A), G% (Panel B), C% (Panel C), A% (Panel D), and U% (Panel E) were calculated and represented (mean±S.E.) for singleton species (open bar, count number, n=1), siRNAs repeatedly sequenced less than 10 times (gray bar, count number, $1<n\leq10$), and species repeatedly sequenced for more than 10 times (black bar, count number, $n>10$). Asterisks (*) indicate statistically significant difference (*t*Test, $P<0.05$) to the next column in left. Replicate datasets of AGO4, AGO6, and AGO9 were shown in Supporting Fig. 4.

Figure 5. GC elimination in the *Potyviridae* genomes

Bar chart (Panel A) shows the Mean±SE of GC% calculated for the 5'- UTR, 3'-UTR, and CDS of the *Potyviridae* genomes (n=69). Scatter plot (Panel B) shows the GC% (Y-axis) at the 1st (diamond labelled), 2nd (square labelled), and 3rd (triangle labelled) codon positions against the overall genome GC% (X-axis) with the regression trendlines ($P<0.001$, n=69). The theoretical null point is genome GC% (X-axis) = 50%.

Supporting information

Supporting Table 1: AGO associated small RNA datasets used in this study

Supporting Table 2: GC contents of 953 plant virus reference genomes

Supporting Table 3: GC contents of 69 *Potyviridae* genomes

Supporting Figure 1: Length distributions of TuMV and CSV vsiRNAs

Supporting Figure 2: Distributions of TuMV and CSV vsiRNA hotspots

Supporting Figure 3: Compositional profiles of vsiRNAs sorted by the 5'-end leading nucleotide

Supporting Figure 4: Nucleotide contents in *Arabidopsis* AGO associated siRNAs

Supporting Figure 5: Nucleotide contents in *Drosophila* AGO associated siRNAs