



Article (refereed) - postprint

Ma, Jinmin; Pallett, Denise; Jiang, Hui; Hou, Yong; Wang, Hui. 2015. **Mutational bias of Turnip Yellow Mosaic Virus in the context of host anti-viral gene silencing**.

# Mutational Bias of *Turnip Yellow Mosaic Virus* in the Context of Host Anti-viral Gene Silencing

Jinmin Ma [1], Denise Pallett [2], Hui Jiang [1], Yong Hou [1], Hui Wang [1,2,3*]

1. BGI-shenzhen, Beishan Road, Yantian, Shenzhen 518083, China

2. NERC/Centre for Ecology & Hydrology, Benson Lane, Wallingford, Oxfordshire OX10 8BB, UK

3. Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Correspondence: huw@ceh.ac.uk and huiwang789@gmail.com

**Abstract**

Plant Dicer-like (DCL) enzymes exhibit a GC-preference during anti-viral post-transcriptional gene silencing (PTGS), delivering an evolutionary selection pressure resulting in plant viruses with GC-poor genomes. However, some viruses, e.g. *Turnip yellow mosaic virus* (TYMV, genus *Tymovirus*) have GC-rich genomes, raising the question as to whether or not DCL derived selection pressure affects these viruses. In this study we analysed the virus-derived small interfering RNAs from TYMV-infected leaves of *Brassica juncea* and showed that the TYMV population accumulated a mutational bias with AU replacing GC (GC-AU), demonstrating PTGS pressure. Interestingly, at the highly polymorphic sites the GC-AU bias was no longer observed. This suggests the presence of an unknown mechanism preventing mutational drift of the viral population and maintaining viral genome stability, despite the host PTGS pressure.

**Key Words**: Anti-viral post-transcriptional gene silencing, *Turnip Yellow Mosaic Virus*, *Brassica juncea*, virus population, mutational bias

## Introduction

Plant virus infections trigger host anti-viral responses which are mediated by post-transcriptional gene silencing (PTGS, also known as RNA interference in animals) (Ding and Voinnet, 2007; Pumplin and Voinnet, 2013). Plant PTGS is mediated by the Dicer-like (DCL) RNAse III enzymes which cleave double-stranded (ds)RNAs. There are four families of plant DCL enzymes involved in anti-viral PTGS that produces virus-derived small interfering (vsi)RNAs in different lengths (DCL-4: 21 nt, DCL-2: 22 nt, DCL-3: 24 nt) (Deleris et al., 2006). Unlike animal Dicers, plant DCLs operate with a GC bias, *i.e.*, producing vsiRNAs with a bias for GC-rich regions of the viral genomes (Donaire et al., 2009; Ho et al., 2006; Ho

et al., 2007; Ho et al., 2010; Miozzi et al., 2013; Yan et al., 2010; Zhang et al., 2014). It has been hypothesised that a GC-bias in vsiRNA production results in a selection pressure on plant virus genome evolution. Evidence obtained by analysing the vsiRNA populations supports the hypothesis that this pressure results in the evolution of GC-poor (AT/U-rich) genomes and reduces host anti-viral PTGS responses during infection (Ho et al., 2010). However, despite the fact that majority of the plant viruses have GC-poor genomes, there are some plant viruses that display GC-rich genomes. This raises the question as to how the GC-rich viruses deal with the plant anti-viral PTGS attack and how they maintain genomic integrity. In this study we analysed the mutation profiles of a GC-rich, positive single-stranded (+ss)RNA virus, *Turnip Yellow Mosaic Virus* (TYMV, genome size 6318-6320 nt long, genus *Tymovirus*, family *Tymoviridae*) using the vsiRNA population defined by the small RNA sequencing protocol based on the Illumina Hi-Seq platform. Next generation sequencing (NGS) technology has previously been shown to be a powerful tool for studying viral ecology (Stobbe and Roossinck, 2014; Virgin, 2014) and viral populations (Beerenwinkel and Zagordi, 2011; Skums et al., 2014; Skums et al., 2013; Watson et al., 2013; Willerth et al., 2010; Wright et al., 2011). Small (s)RNA sequencing is particularly useful for characterizing host anti-viral immunity mediated by PTGS (Donaire et al., 2009; Pallett et al., 2010; Wu et al., 2010). Infection with TYMV triggers vsiRNA production by DCLs (Jakubiec et al., 2012) and the virus encodes a PTGS suppressor protein, P69 (Chen et al., 2004). In this study, more GC to AU (GC-AU) mutations were observed at the GC-dominant positions than the AU-GC mutations at the AU-dominant positions, demonstrating that PTGS derived selection pressure operated within the TYMV population. However, this GC-AU bias was not detected in the highly polymorphic sites in the TYMV genome, suggesting that the viral genome composition may be maintained by an unknown mechanism.

**Materials and Methods**

Seedlings of *Brassica juncea* were manually inoculated two weeks post germination with a local TYMV strain (TYMV-Ox-S, NCBI Accession No: KP883302) isolated from leaves of wild *Brassica rapa* grown on the banks of the river Thames, Culham, Oxfordshire, England, in 2011. Four weeks after inoculation, non-inoculated young leaves showing yellowing symptoms were collected from 10 *B. juncea* seedlings and sRNA (< 200-nt) was extracted using the mirVana miRNA isolation kit (Cat AM1560, Ambion, Life Technologies Ltd, Paisley, UK) following the manufacturer's protocol. The experiments were repeated twice independently. In the first experiment, leaves with severe yellowing symptoms were selected in order to increase the yield of TYMV vsiRNAs. In the second experiment, young leaves with less severe symptoms were selected and a negative control sRNA sample was produced using leaves from mock (water) inoculated *B. juncea* seedlings. Concentrations of the sRNA extracts were measured by using the Nanodrop 1000 UV spectrophotometer (Nanodrop Products, Wilmington, USA). The sRNA extracts were sequenced by the Beijing Genomics Institute (BGI, Hong Kong, China) using the Illumina small RNA protocol. The deep sequencing dataset was deposited in the NIH Short Read Archive with Accession Numbers SRR2017660 (TYMV-Pre, the first experiment), SRR1867781 (TYMV-NGS, the second experiment) and SRR1867782 (TGM-CK, negative control).

Raw sequence reads were subjected to a standard Illumina Solexa quality control pipeline (with the average single base error rate cut-off value of 0.01). Reads shorter than 18 nt and longer than 44 nt were removed, together with reads having more than 2 bases with Phred quality scores below 10. Non-coding RNAs (rRNA, tRNA, snRNA, snoRNA) were screened and removed using BLAST 2.2.23 against the databases of Rfam (http://www.sanger.ac.uk/software/Rfam), the GenBank noncoding RNA database (http://www.ncbi.nlm.nih.gov/) and miRNAs (miRBase Release-18). The remaining reads

were assembled using SOAPdenovo v1.05 (http://soap.genomics.org.cn/soapdenovo.html) with Kmer values of 15 and 17 (Li et al., 2010). The *De novo* assembly produced 17 TYMV contigs (5,664 nt) from the TYMV-NGS library but nil from the control library. An advanced assembly of the TYMV contigs was performed using PHRAP with the parameters: 50-100 overlap match, vector bound 30, max gap 5 (de la Bastide and McCombie, 2007). The resulting virus consensus sequence (NCBI Accession Number: KP883301, TYMV_Ox_NGS) was 6319 nt long and similar to the sequence determined by the Sanger sequencing (TYMV-Ox-S, NCBI Accession No KP883302) with a difference of only 10 nucleotides. A Neighbour Joint phylogenetic tree (Supplementary Figure S1) was constructed using Clustal X and Mega-6 programs (Chenna et al., 2003; Tamura et al., 2013) to show the phylogenetic relationship between the Oxford isolate and the other reported TYMV isolates (Dreher and Bransom, 1992; Keese et al., 1989; Morch et al., 1988; Skotnicki et al., 1992).

Reads in the two TYMV libraries were mapped against the TYMV_Ox_NGS sequence using BOWTIE (Langmead et al., 2009) with two mismatches allowed. The coverage (Depth) and nucleotide diversity (Pi) was calculated for each position of the virus genome (Supplementary Tables S1 and S2). To test the null hypothesis that the rate of A and U (AU) mutations detected at the G and C (GC) dominant sites (GC-AU) were the same as that of GC mutations detected at the AU dominant sites (AU-GC), a minor nucleotide pair rate (the sum of A%+U% at a GC dominant site, GC-AU; or the sum of G%+C% at an AU dominant site, AU-GC) was calculated for each viral genome position, and used to compare GC-AU to AU-GC using the Analysis of Variance (ANOVA, MiniTab). To investigate whether or not our observations were affected by possible experimental artefacts, e.g., sequencing error and bias, the minor nucleotide pair rates (GC-AU and AU-GC) were analysed for their interactions with sequencing Depth and Pi using the General Linear Model (also known as Generalized Linear Models, GLM, MiniTab). The dependent variable (known

as the Response in MiniTab) was the minor nucleotide pair rate (A%+U% at a GC dominant site or G%+C% at an AU dominant site) (Supplementary Tables S1 & S2). The independent variables (known as Factors in MiniTab) were Site Domination (code-AU&GC), Sequencing Depth (code-depth) and Nucleotide Diversity (code-Pi) (Supplementary Tables S1 & S2). The samples (observations) were the minor nucleotide pair rates observed at each genome positions (Supplementary Tables S1 & S2). To detect the significant interactions between the code-AU&GC and code-depth, GLM model of "code-AU&GC" * "code-depth" was used. To detect significant interactions between code-AU&GC and code-Pi, GLM model of "code-AU&GC" * "code-Pi" was used. A probability lower than 0.05 ($P < 0.05$) was interpreted as statistically significant. An observation of positive correlation between Pi and GC-AU > AU-GC bias is indicative of a virus response to host PTGS immunity whereas observations of sequencing Depth affecting GC-AU > AU-GC suggest possibilities of experimental artefacts.

**Results and Discussion**

In all three samples, approximately 10 million sRNA reads were obtained (SRR2017660, TYMV-Pre, 1,911,164 unique reads; SRR1867781, TYMV-NGS, 4,691,814 unique reads; SRR1867782, Control, 3,022,140 unique reads). All libraries were dominated by 24 nt long sRNA species (Supplementary Figure S2), indicating good sRNA quality.

In both TYMV samples, TYMV vsiRNAs covered the virus genome fully. The GC contents of the TYMV vsiRNAs were 56.32% and 56.09% for TYMV-Pre and TYMV-NGS respectively, comparable with that of the TYMV_Ox_NGS consensus sequence (56.32%). There was no GC enrichment observed for the TYMV vsiRNAs, in contrast to reports of other plant viruses with GC-poor genomes (Donaire et al., 2009; Ho et al., 2006; Ho et al., 2007; Ho et al., 2010; Miozzi et al., 2013; Yan et al., 2010; Zhang et al., 2014). It is likely the high GC content of the TYMV genome masked the GC-preference of the plant DCLs, as

previously observed in the micro(mi)-RNA productions from GC-rich precursor miRNAs (pre-miRNA) in *Gramineae* species (Ho et al., 2006; Ho et al., 2007). However, there were significant GC increments in vsiRNA species that were detected repeatedly (Figure 1A, TYMV-Pre, df=3, F=8.25, $P<0.001$; TYMV-NGS, df=3, F=10.33, $P<0.001$, ANOVA, MiniTab), demonstrating evidence of plant DCL mediated GC-bias in vsiRNA production (Donaire et al., 2009; Ho et al., 2006; Ho et al., 2007; Ho et al., 2010; Miozzi et al., 2013; Yan et al., 2010; Zhang et al., 2014). The minor nucleotide pair rates at the TYMV genome positions were also significantly different (Figure 1B). The GC-AU were higher than the AU-GC in both TYMV-Pre (df=1, F=146.16, $P<0.001$, ANOVA, MiniTab) and TYMV-NGS (df=1, F=39.17, $P<0.001$) (Figure 1B), indicating that there were higher rates of AU mutations detected at GC dominant sites than the rates of GC mutations detected at AU dominant sites. This suggests that the TYMV populations in both samples were under a selection pressure mediated by the GC-preference of plant PTGS. However, due to the possibility that experimental factors may affect the NGS performance (Sims et al., 2014), we further investigated the viral mutation profiles (Supplementary Tables S1 and S2) of the TYMV samples.

If the accumulation of GC-AU mutations (Figure 1B) was due to genuine biological processes it would positively correlate to Pi, *i.e.*, the higher Pi the greater GC-AU > AU-GC. The GLM analysis showed such evident in both TYMV samples (Supplementary Figure S3). However, we also detected experimental factors that influenced the GC-AU and AU-GC. Due to the sequence quality of the two TYMV datasets (described in following paragraphs), we used the TYMV-NGS library to further clarify the mutational bias of TYMV.

From the TYMV-NGS library, 1,091,704 TYMV derived vsiRNAs (74,716 unique) were mapped against the TYMV_Ox_NGS consensus sequence (Figure 2). The TYMV vsiRNAs were dominated by the 21 nt species (Figure 2A) compared to the 24 nt domination

of the total sRNA (Supplementary Figure S2B) indicating that the host DCL-4 pathway was dominant in TYMV vsiRNA production (Deleris et al., 2006). The TYMV vsiRNAs were generated from both plus (32.3%) and minus (67.7%) strands of the virus genome (Figure 2B), suggesting that the double-stranded viral intermediates severed as the DCL targets. In addition to the result shown in Figure 1B, the Kruskal-Wallis test also showed that the AU mutation rate at the GC-dominant sites (GC-AU, n=3554, Mid=0.23%) was significantly higher (H=704.54, df=1, *P*=0.000) than the GC mutation rate at the AU-dominant sites (AU-GC, n=2765, Mid=0.12%), resulting in the rejection of the null hypothesis that GC-AU is equal to AU-GC. Although the plants were sampled at a single time point, the observed single nucleotide polymorphism profile (Supplementary Table S2) represented a snapshot of the TYMV population. The significantly elevated GC-AU rate (compared to AU-GC) suggested that the TYMV population was biased towards accumulating GC-AU mutations.

If there were significant experimental artefacts generated due to procedure, e.g., sequencing errors (Sims et al., 2014), the GC-AU mutational bias would be observed in relation to the sequencing depth of each site, *i.e*., the more time a site was subjected to analysis, the more (or less) experimental artefacts would be displayed. Therefore, we divided the TYMV sites into 10 groups according to the sequence coverage (Depth) in descending order, *i.e.*, Depth-Group-A represented the top 10% of the mostly sequenced sites and Depth-Group-B represented the next 10% of sites etc. (Supplementary Table S2). Analysis with GLM detected statistically significant interactions (df=9, F=2.34, *P*=0.012) between the site domination (Code-AU-GC) and the sequencing depth (Code-depth) (Supplementary Figure S4) showing that the sequencing depth (ranging from 10 – 41254, n=6319 sites) was a factor contributing to the difference between GC-AU and AU-GC (Figure 3A). When sequencing depth > 2886 (Supplementary Table S2) in the Depth-Groups A-D (Figure 3B), the difference between GC-AU and AU-GC became stable indicating that sequencing Depth was no longer

a significant factor (df=3, F=0.02, *P*=0.995, GLM, Supplementary Figure S4). Therefore there was no detectable experimental artefact on the observed GC-AU bias (GC-AU > AU-GC, Figure 3B) in TYMV-NGS Depth-Groups A-D. Using these sites in TYMV-NGS, we went further to clarify the relationship of the GC-AU bias and nucleotide diversity. The TYMV-Pre dataset was not used for such analysis as the experimental factor (Depth) continued to affect the difference between GC-AU and AU-GC regardless of sequencing depth (Supplementary Figure S5).

Using only the TYMV-NGS library, the nucleotide diversity (Pi) was calculated for each site, and the whole genome was divided into another 10 groups in descending order, each containing 10% of the sites (i.e. Pi-Group-A contained the top 10% most polymorphic sites etc., Supplementary Table S2). Analysing the sites that belonged to the Depth-Groups A-D (2528 sites, Figure 3B) with GLM detected a significant interaction (df=9, F=1.88, *P*=0.050) between the Site Domination (Code-AU-GC) and Pi (Code-Pi) (Supplementary Figure S6). The low statistical confidence was mainly due to the sites in Pi-Group-A in which GC-AU was similar to AU-GC (Figure 4A). Removal of the anomalous Pi-Group-A from the analysis showed that the difference between GC-AU and AU-GC positively correlated to Pi, *i.e.*, the higher the Pi, the greater in GC-AU > AU-GC (Figure 4B, regression line y = 0.4927x – 0.041, $R^2$ =0.9748). This was a clear demonstration that the observed GC-AU bias was due to genuine biological function (virus response to host PTGS immunity) rather than experimental process.

By only using sites of deeply sequenced sites (Depth-Groups A-D, Figure 3) in TYMV-NGS, we removed the experimental bias and subsequently obtain evidence that the TYMV genome were more likely to accumulate GC-AU mutations than AU-GC mutations (Figure 4), thus demonstrating a selection pressure for the TYMV population to replace GC with AU. This new evidence supports previous reports that the plant anti-viral PTGS system

operates with a GC-bias (Donaire et al., 2009; Ho et al., 2006; Ho et al., 2007; Ho et al., 2010; Miozzi et al., 2013; Yan et al., 2010; Zhang et al., 2014) and that such a GC-bias delivers a selection pressure on plant virus evolution to develop GC-poor genomes (Ho et al., 2010). However, it remains as a mystery how some plant viruses like TYMV maintain a GC-rich genome. Indeed, we observed that the GC-AU bias was no longer in evidence in the top 10% polymorphic sites, *i.e.* Pi-Group-A (Pi >= 0.93%, Figure 4A) in TYMV-NGS. As there is ample evidence that the plant viruses have evolved a wide range of PTGS suppressors to combat the host anti-viral PTGS (Pumplin and Voinnet, 2013; Voinnet et al., 1999), it is tempting to speculate that plant viruses may have evolved diverse strategies against the PTGS mediated selection pressure of GC-AU replacement. It is plausible that there is an unknown mechanism which maintains the TYMV genome stability.

**Acknowledgements**

The work was supported by the NERC (UK) grants NE/I000593/1 and NE/L012863/1 to HW and the Ministry of Science and Technology of China (International Collaboration Research: 2011DFA33220).

**References**

Beerenwinkel, N., Zagordi, O., 2011. Ultra-deep sequencing for the analysis of viral populations. Current opinion in virology 1, 413-418.

Chen, J., Li, W.X., Xie, D., Peng, J.R., Ding, S.W., 2004. Viral virulence protein suppresses RNA silencing-mediated defense but upregulates the role of microrna in host gene expression. The Plant cell 16, 1302-1313.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic acids research 31, 3497-3500.

de la Bastide, M., McCombie, W.R., 2007. Assembling genomic DNA sequences with PHRAP. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] Chapter 11, Unit11-14.

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

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

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

Dreher, T.W., Bransom, K.L., 1992. Genomic RNA sequence of turnip yellow mosaic virus isolate TYMC, a cDNA-based clone with verified infectivity. Plant molecular biology 18, 403-406.

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

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

Ho, T., Wang, L., Huang, L., Li, Z., Pallett, D.W., Dalmay, T., Ohshima, K., Walsh, J.A., Wang, H., 2010. Nucleotide bias of DCL and AGO in plant anti-virus gene silencing. Protein Cell 1, 847-858.

Jakubiec, A., Yang, S.W., Chua, N.H., 2012. Arabidopsis DRB4 protein in antiviral defense against Turnip yellow mosaic virus infection. The Plant journal : for cell and molecular biology 69, 14-25.

Keese, P., Mackenzie, A., Gibbs, A., 1989. Nucleotide sequence of the genome of an Australian isolate of turnip yellow mosaic tymovirus. Virology 172, 536-546.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10, R25.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., Wang, J., 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome research 20, 265-272.

Miozzi, L., Pantaleo, V., Burgyan, J., Accotto, G.P., Noris, E., 2013. Analysis of small RNAs derived from tomato yellow leaf curl Sardinia virus reveals a cross reaction between the major viral hotspot and the plant host genome. Virus research 178, 287-296.

Morch, M.D., Boyer, J.C., Haenni, A.L., 1988. Overlapping open reading frames revealed by complete nucleotide sequencing of turnip yellow mosaic virus genomic RNA. Nucleic acids research 16, 6157-6173.

Pallett, D.W., Ho, T., Cooper, I., Wang, H., 2010. Detection of Cereal yellow dwarf virus using small interfering RNAs and enhanced infection rate with Cocksfoot streak virus in wild cocksfoot grass (Dactylis glomerata). J Virol Methods 168, 223-227.

Pumplin, N., Voinnet, O., 2013. RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence. Nature reviews. Microbiology 11, 745-760.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P., 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nature reviews. Genetics 15, 121-132.

Skotnicki, M.L., Mackenzie, A.M., Gibbs, A.J., 1992. Turnip yellow mosaic virus variants produced from DNA clones encoding their genomes. Arch Virol 127, 25-35.

Skums, P., Artyomenko, A., Glebova, O., Ramachandran, S., Mandoiu, I., Campo, D.S., Dimitrova, Z., Zelikovsky, A., Khudyakov, Y., 2014. Computational framework for next-generation sequencing of heterogeneous viral populations using combinatorial pooling. Bioinformatics 31, 682-690.

Skums, P., Mancuso, N., Artyomenko, A., Tork, B., Mandoiu, I., Khudyakov, Y., Zelikovsky, A., 2013. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. BMC bioinformatics 14 Suppl 9, S2.

Stobbe, A.H., Roossinck, M.J., 2014. Plant virus metagenomics: what we know and why we need to know more. Frontiers in plant science 5, 150.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution 30, 2725-2729.

Virgin, H.W., 2014. The virome in mammalian physiology and disease. Cell 157, 142-150.

Voinnet, O., Pinto, Y.M., Baulcombe, D.C., 1999. Suppression of gene silencing: a general strategy used by diverse DNA and RNA viruses of plants. Proceedings of the National Academy of Sciences of the United States of America 96, 14147-14152.

Watson, S.J., Welkers, M.R., Depledge, D.P., Coulter, E., Breuer, J.M., de Jong, M.D., Kellam, P., 2013. Viral population analysis and minority-variant detection using short read next-generation sequencing. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368, 20120205.

Willerth, S.M., Pedro, H.A., Pachter, L., Humeau, L.M., Arkin, A.P., Schaffer, D.V., 2010. Development of a low bias method for characterizing viral populations using next generation sequencing technology. PLoS One 5, e13564.

Wright, C.F., Morelli, M.J., Thebaud, G., Knowles, N.J., Herzyk, P., Paton, D.J., Haydon, D.T., King, D.P., 2011. Beyond the consensus: dissecting within-host viral population

diversity of foot-and-mouth disease virus by using next-generation genome sequencing. Journal of Virology 85, 2266-2275.

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E.C., Li, W.X., Ding, S.W., 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proceedings of the National Academy of Sciences of the United States of America 107, 1606-1611.

Yan, F., Zhang, H., Adams, M.J., Yang, J., Peng, J., Antoniw, J.F., Zhou, Y., Chen, J., 2010. Characterization of siRNAs derived from rice stripe virus in infected rice plants by deep sequencing. Arch Virol 155, 935-940.

Zhang, Y., Yan, C., Kuang, H., 2014. GC content fluctuation around plant small RNA-generating sites. FEBS Lett 588, 764-769.

**Figure legends**

Figure 1: Plant GC biases in TYMV vsiRNA production and GC-AU bias in TYMV mutagenesis.

Data was based on both of the TYMV infected samples (TYMV-Pre and TYMV-NGS). Panel A shows the increments of GC contents in repeatedly detected TYMV reads (redundant reads that have identical sequences, $P<0.001$ for both TYMV-Pre and TYMV-NGS series), indicating that the TYMV vsiRNA hotspots (Figure 2B) had higher GC contents than the non-hotspot regions. The X-axis numbers indicate read redundancy categories of n=1, $1<n\leq10$, $10<n\leq100$ and $n>100$. Panel B shows that the rates of AU mutations detected at the GC dominant sites (GC-AU) were higher ($P<0.001$ for both the samples) than those of GC mutations detected at the AU dominant sites (AU-GC), suggesting that the TYMV populations accumulated mutations with a GC-AU bias.

Figure 2: Length and mapping distributions of TYMV vsiRNAs.

Panel A shows the length distribution of the TYMV vsiRNAs from the infected sample TYMV-NGS. X-axis represents the length range (nt) and the Y-axis represents the proportion (%) of each length species. Panel B shows the vsiRNA distribution along the TYMV genome positions (X-axis) and the vsiRNA coverage (read number, Y-axis) on either plus or minus strand polarity. The outer lines represent data of total vsiRNA reads and the inner lines represent data of unique vsiRNA reads.

Figure 3: The minor nucleotide pair rate were affected by sequencing depth of TYMV sites.

Each TYMV_OX_NGS position was deemed as either AU or GC dominant site according to the consensus sequence. All sites were divided into 10 groups (A-J) according to their sequencing depths in descending order. Panel A showed the minor nucleotide pair rate (%, Y-axis, mean ± SE) of AU to GC substitutions (AU-GC, open column) at AU dominant sites or that of GC to AU substitutions (GC-AU, filled column) at GC dominant sites. GLM was performed using the MiniTab program and the result showed that the difference of GC-AU and AU-GC was significantly affected by sequencing depth ($P$=0.012, Supplementary Figure S4), suggesting experimental biases. Panel B showed that the difference of minor nucleotide pair rates (GC-AU > AU-GC, Y-axis) stabilized in the Depth-Groups A-D (labelled by letters, X-axis, sequencing depth, mean ± SE). Among the A-D groups, the experimental factor did not affect the GC-AU > AU-GC ($P$=0.995, GLM, Supplementary Figure S4). Therefore, sites of Depth-Group A-D were used for further analysis.

Figure 4: The minor nucleotide pair rate were affected by nucleotide diversity of individual TYMV sites.

All TYMV_OX_NGS positions were deemed as either AU or GC dominant site according to the consensus sequence. All sites were divided into 10 groups (A-J) according to their nucleotide diversities (Pi) in descending order. Only sites belonging to the Depth-Groups A-D (Figure 3B, Depth $\geqslant$ 2886, Supplementary Table S2) were used for this analysis. Panel A shows the minor nucleotide pair rates (Y-axis, mean $\pm$ SE) of AU to GC substitutions (AU-GC, open column) at AU dominant sites or that of GC to AU substitutions (GC-AU, filled column) at GC dominant sites. GLM analysis showed that the difference of GC-AU and AU-GC was affected by Pi ($P$=0.05, Supplementary Figure S6) and the relatively low confidence was mainly due to a dramatically elevated AU-GC in Pi-Group A (Pi > 0.93%). Panel B shows that the difference of GC-AU and AU-GC (GC-AU > AU-GC, Y-axis) positively correlates to Pi (X-axis, mean ± SE) in sites of Pi-Groups B-J (labelled by letters), supporting the hypothesis that the TYMV population has a GC-AU mutational bias. The regression line was y = 0.4927x – 0.041 ($R^2$ =0.9748).

**Supplementary Materials**

Supplementary Table S1: Sequencing depth and mutation profile of the TYMV population in sample TYMV-Pre.

Supplementary Table S2: Sequencing depth and mutation profile of the TYMV population in sample TYMV-NGS.

Supplementary Figure S1: Phylogenetic tree of TYMV.

The sequence alignment was made using the Clustal X program (http://www.clustal.org/) (Chenna et al., 2003) and an unrooted Neighbour Joint tree was constructed with 1000

bootstraps by Mega 6 program (http://www.megasoftware.net/) (Tamura et al., 2013). The black dot labels the TYMV_OX_NGS consensus sequence used for mapping the NGS reads. The Oxford strain was more closely related to a Netherlands isolate (KJ690173) than to TYMV isolates from other parts of the world: Australian isolates [J04373, (Keese et al., 1989); AF035403, (Skotnicki et al., 1992)], an American isolate [X16378, (Dreher and Bransom, 1992)] and another European isolate [X07441, (Morch et al., 1988)].

Supplementary Figure S2: Length distribution of the small RNA libraries.

Length distribution of the small RNAs were obtained from the TYMV-infected samples TYMV-Pre (Panel A), TYMV-NGS (Panel B) and Mock-inoculated sample (Panel C). X-axis shows the length range (nt) and the Y-axis shows the proportion (%) of each length species.

Supplementary Figure S3: GLM analysis for the minor nucleotide pair rate affected by nucleotide diversity (Pi) of individual TYMV sites.

The TYMV_OX_NGS genome positions were deemed as either AU or GC dominant (code-AU-GC) according to the consensus sequence. All sites were divided into 10 groups according to their nucleotide diversities (Code-Pi) in descending order. GLM was performed using the MiniTab program. The X-axis shows the code groups and the Y-axis represents minor nucleotide pair rate (%). Panels A (Sample TYMV-Pre) & B (Sample TYMV-NGS) showed that (1) the GC-AU rates were significantly higher than that of the AU-GC rates (TYMV-Pre, df=1, F=45.78, *P*=0.000; TYMV-NGS, df=1, F=34.82, *P*=0.000), (2) Pi significantly affected the minor nucleotide pair rates (TYMV-Pre, df=9, F=234.12, *P*=0.000; TYMV-NGS, df=9, F=125.52, *P*=0.000) and (3) the difference between GC-AU and AU-GC

was significantly affected by Pi (interaction between the two factors, TYMV-Pre, df=9, F=20.78, *P*=0.000; TYMV-NGS, df=9, F=6.28, *P*=0.000).

Supplementary Figure S4: GLM analysis of the minor nucleotide pair rate affected by sequencing coverage of TYMV sites.

Each TYMV_OX_NGS position was deemed as either AU or GC dominant (code-AU-GC) according to the consensus sequence. All sites were divided into 10 groups according to their sequencing depths (Code-depth) in descending order. GLM was performed using the MiniTab program. The code groups also label the X-axis (Code-depth for Panels A and C, and Code-AU-GC for Panels B and D). The Y-axis represents minor nucleotide pair rate (%). Asterisks (Panels A & B) show a significant interaction between Code-AU-GC and Code-depth (*P*=0.012) but no longer in Code-depth Groups A-D (Panels C & D, *P*=0.995).

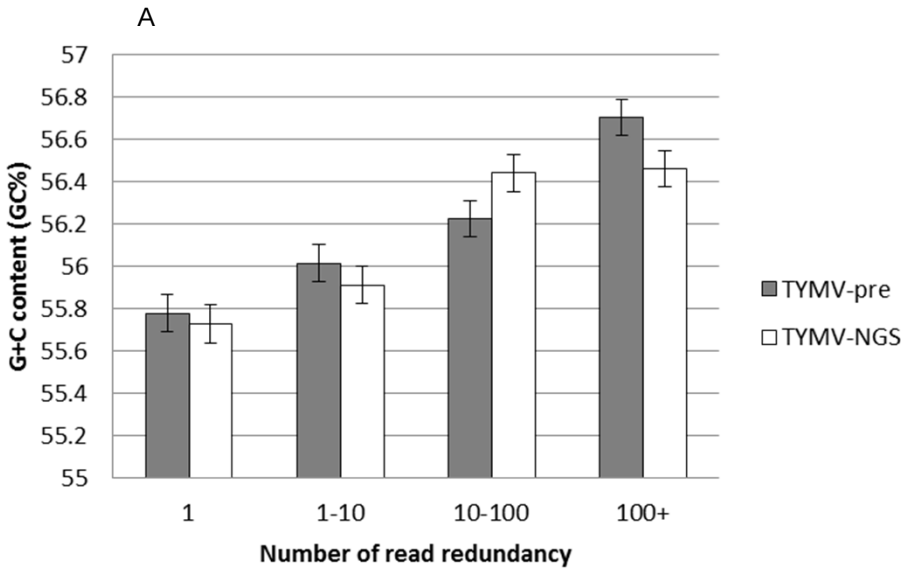
Supplementary Figure S5: GLM analysis for the minor nucleotide pair rate affected by sequencing coverage of TYMV sites in the TYMV-Pre sample.

Each nucleotide position of the TYMV_OX_NGS consensus sequence was deemed as either AU or GC dominant (Code-AU-GC) according to consensus sequence. All sites were divided into 10 groups according to their sequencing depths (Code-depth) in descending order. GLM was performed using the MiniTab program. The X-axis shows the code groups and the Y-axis represents minor nucleotide pair rate (%). The result showed that (1) GC-AU rates were significantly higher than that of AU-GC rates (df=1, F=32.49, *P*=0.000), (2) the minor nucleotide pair rates were significantly affected by the sequencing coverage (Code-depth) (df=9, F=3.95, *P*=0.000) but (3) the interaction between the two factors (Code-AU-GC and Code-depth) was not consistent (df=9, F=1.31, *P*=0.224), indicating that there were unsettled experimental factors in the sequencing run.

Supplementary Figure S6: GLM analysis of the minor nucleotide pair rate affected by nucleotide diversity of individual TYMV sites.

All TYMV_OX_NGS positions were deemed as either AU or GC dominant (Code-AU-GC) according to the consensus sequence. All sites were divided into 10 groups according to their nucleotide diversities (Code-Pi) in descending order. Only sites belonging to the Code-depth Groups A-D (Figure 3C&D, Code-depth, depth > 2886) were used. GLM was performed using the MiniTab program. The code groups also label the X-axis (Code-Pi for Panels A and C, and Code-AU-GC for Panels B and D). The Y-axis represents the minor nucleotide pair rate (%). Panels A & B show that the interaction between Site Domination (Code-AU-GC) and Pi (Code-Pi) was not consistent ($P$ = 0.05) due to the anomalous situation of Code-Pi Group-A sites (Pi > 0.93%). Asterisks (Panels C & D) show a highly significant interaction between the two factors was detected without the Code-Pi Group-A sites ($P$ = 0.000).

Figure 1


A
57
56.8
56.6
56.4
56.2
56
55.8
55.6
55.4
55.2
55
G+C content (GC%)
1
1-10
10-100
100+
Number of read redundancy
TYMV-pre
TYMV-NGS

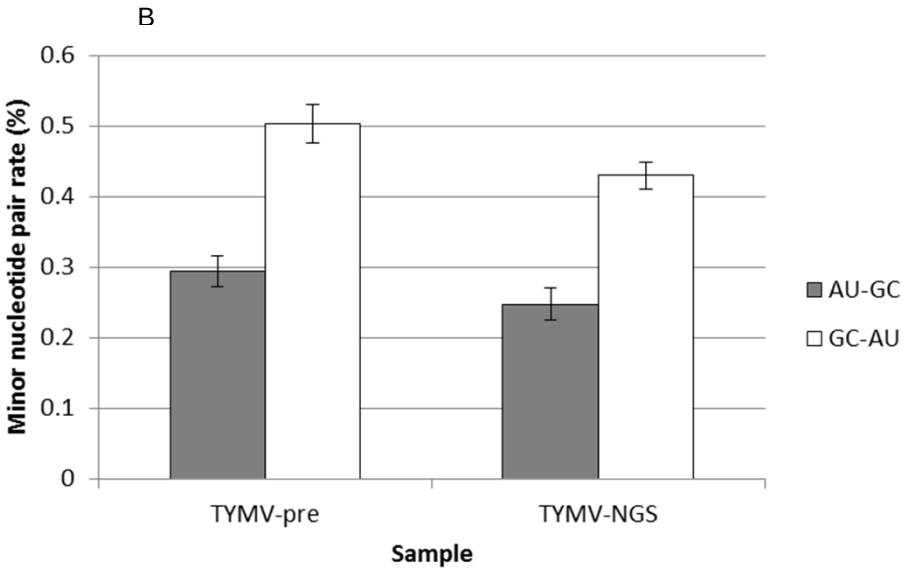
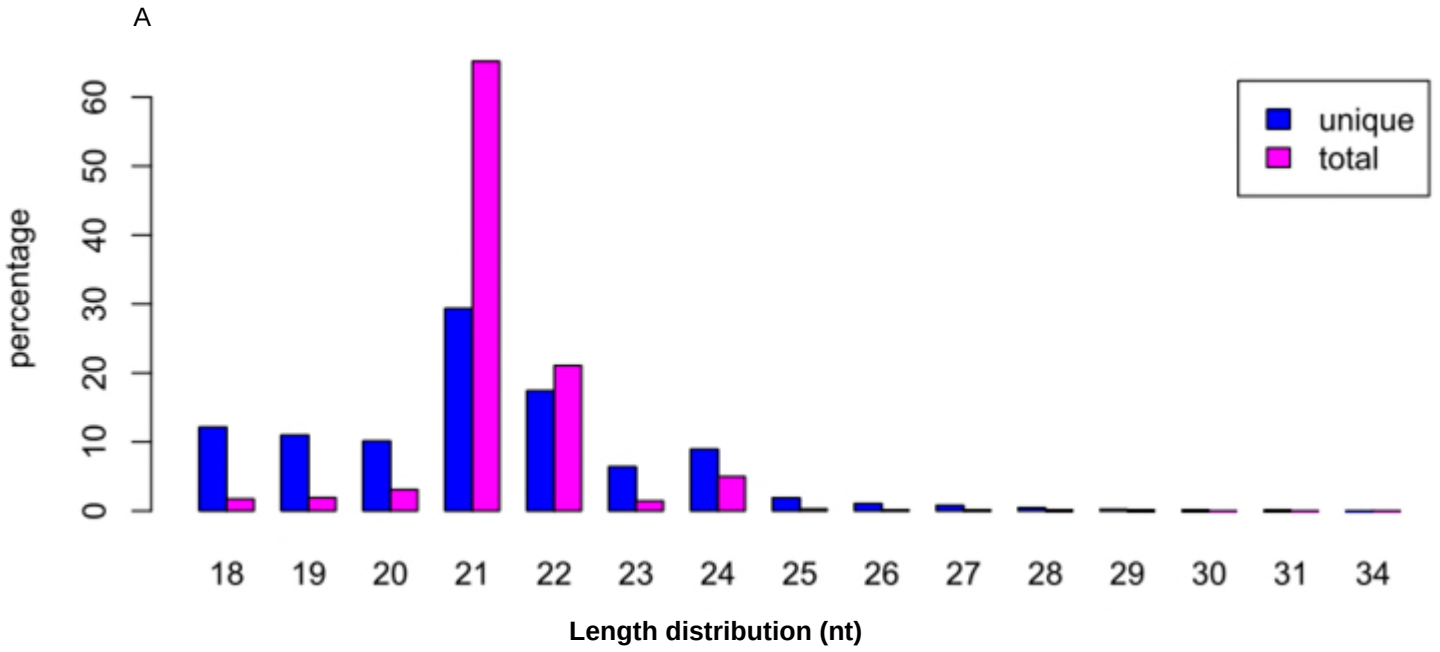

B
0.6
0.5
0.4
0.3
0.2
0.1
0
Minor nucleotide pair rate (%)
TYMV-pre
TYMV-NGS
Sample
AU-GC
GC-AU

Figure 2


A
percentage
0
10
20
30
40
50
60
unique
total
18
19
20
21
22
23
24
25
26
27
28
29
30
31
34
Length distribution (nt)

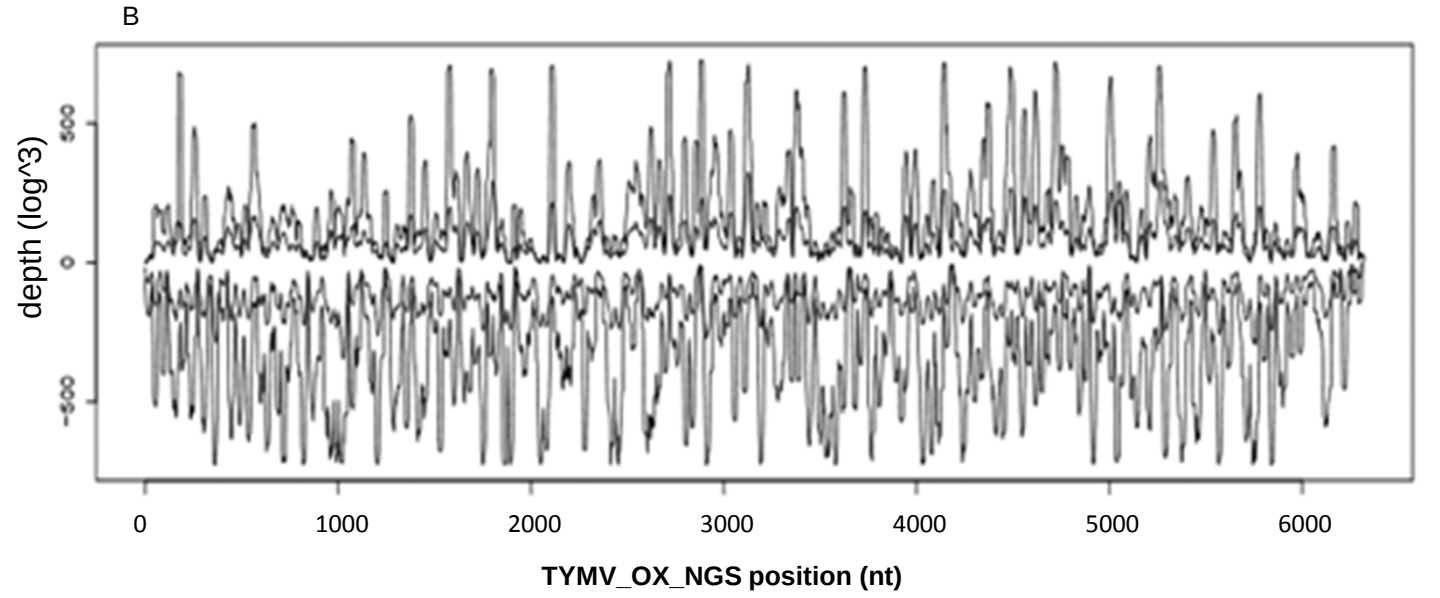
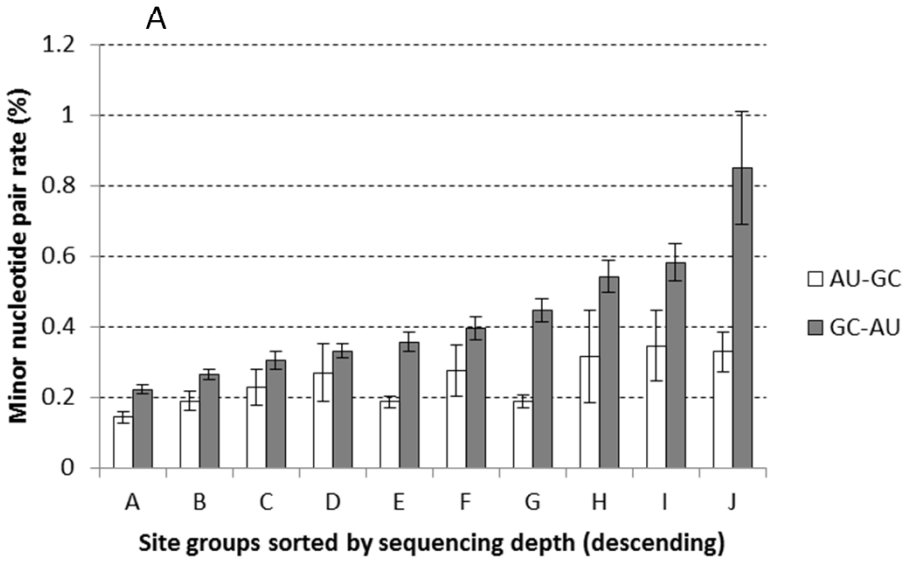

B
depth (log^3)
500
0
-500
0
1000
2000
3000
4000
5000
6000
TYMV_OX_NGS position (nt)

Figure 3


A
1.2
1
0.8
0.6
0.4
0.2
0
Minor nucleotide pair rate (%)
AU-GC
GC-AU
A
B
C
D
E
F
G
H
I
J
Site groups sorted by sequencing depth (descending)

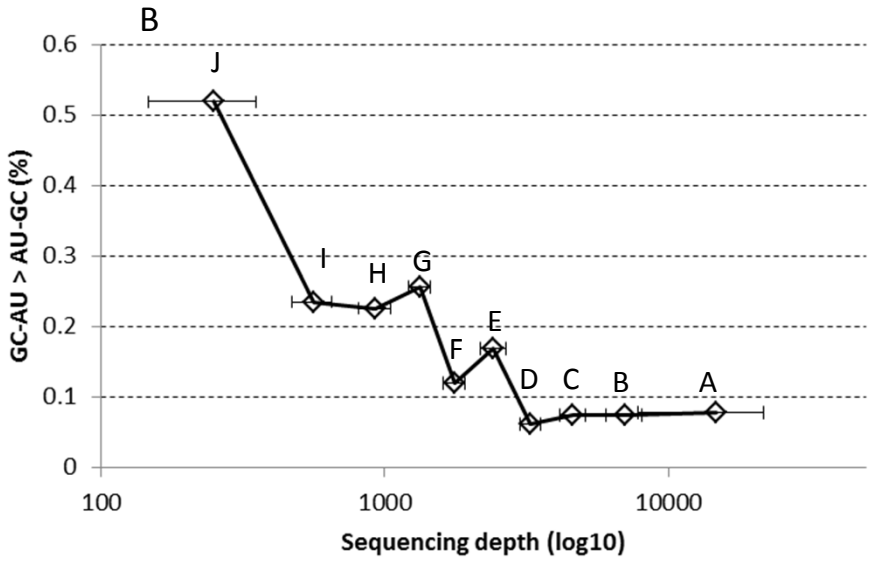

B
0.6
0.5
0.4
0.3
0.2
0.1
0
GC-AU > AU-GC (%)
J
I
H
G
F
E
D
C
B
A
100
1000
10000
Sequencing depth (log10)

Figure 4

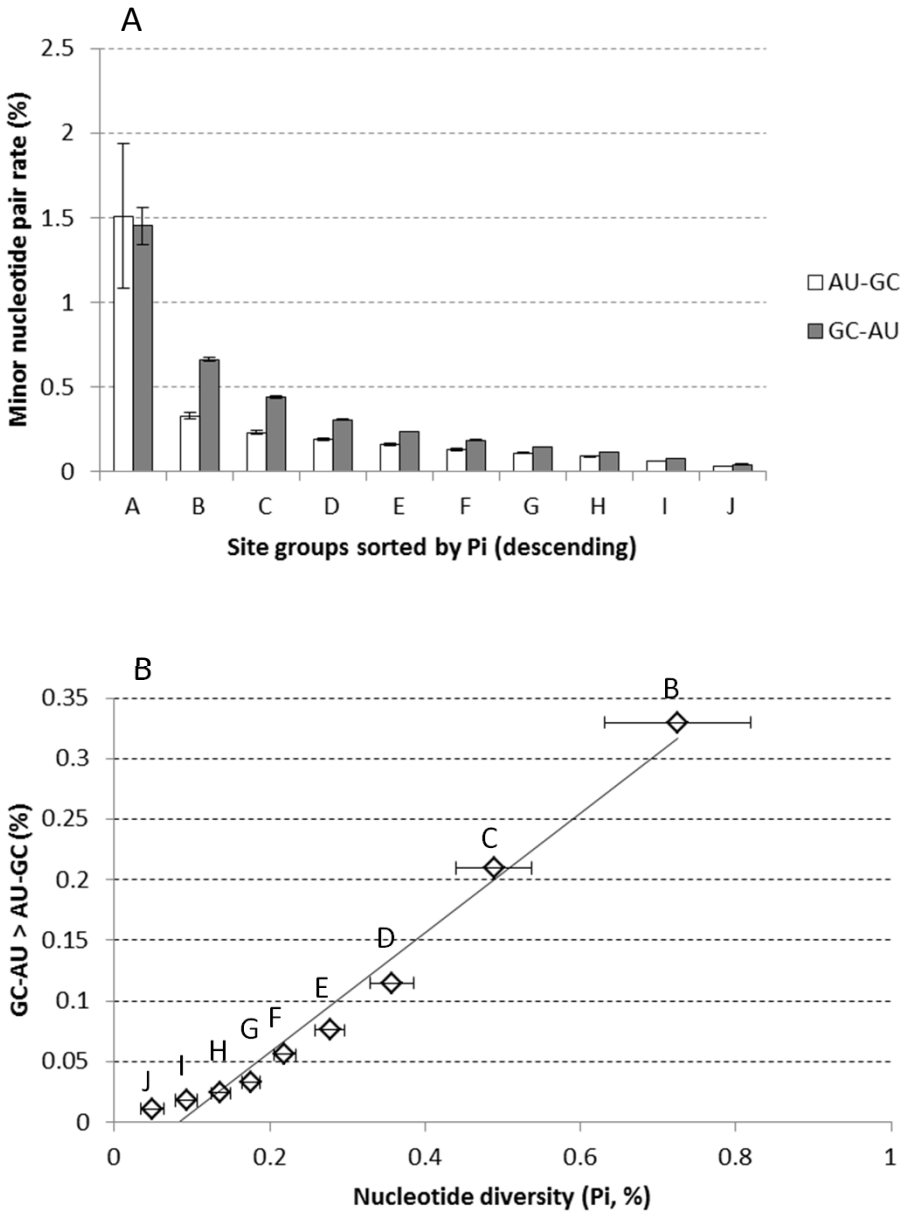

A
Minor nucleotide pair rate (%)
2.5
2
1.5
1
0.5
0
A
B
C
D
E
F
G
H
I
J
Site groups sorted by Pi (descending)
AU-GC
GC-AU
B
GC-AU > AU-GC (%)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Nucleotide diversity (Pi, %)

Figure 4

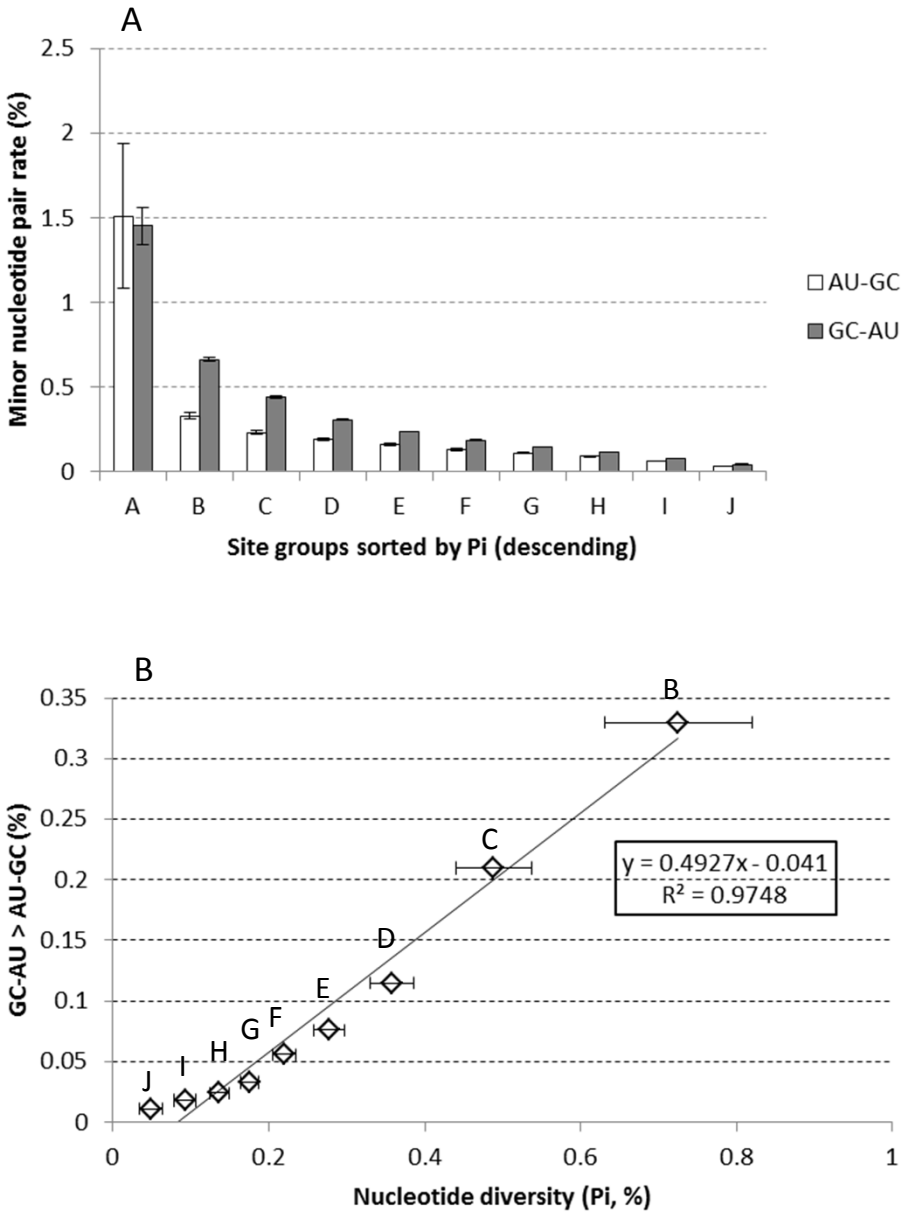
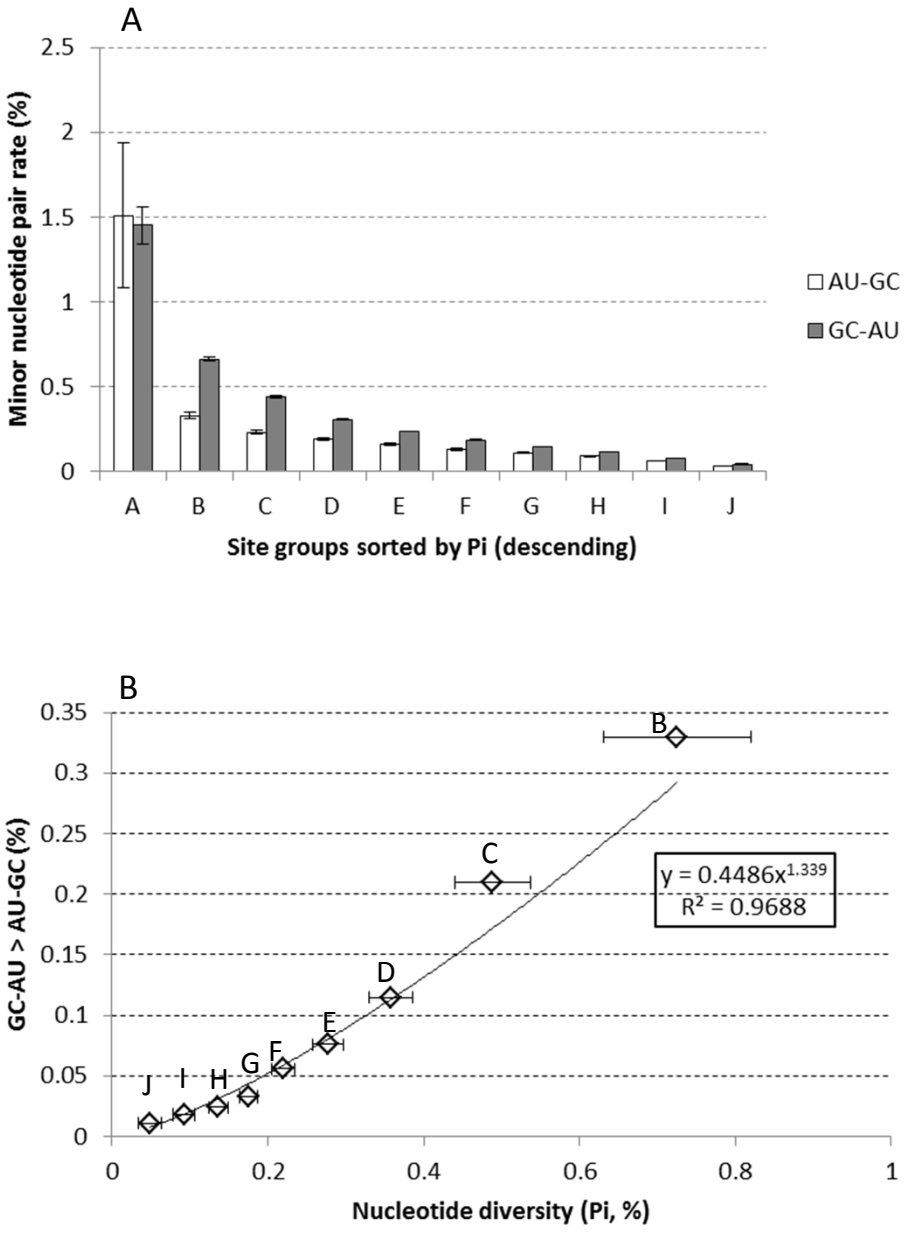
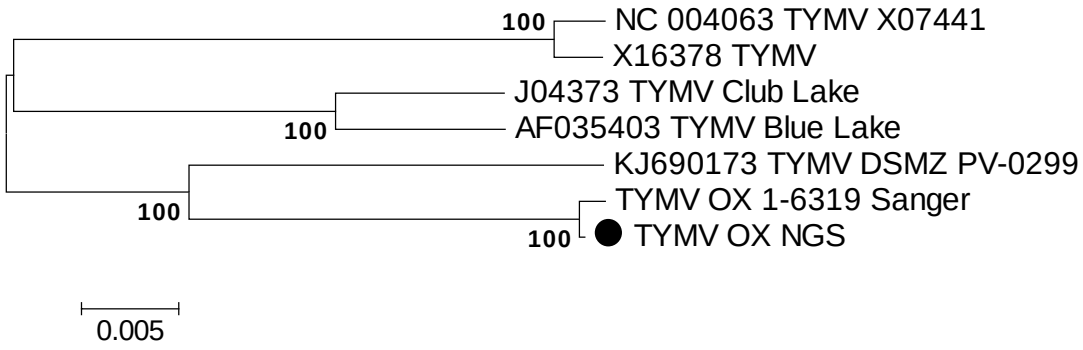

A
Minor nucleotide pair rate (%)
2.5
2
1.5
1
0.5
0
A
B
C
D
E
F
G
H
I
J
Site groups sorted by Pi (descending)
AU-GC
GC-AU
B
GC-AU > AU-GC (%)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Nucleotide diversity (Pi, %)
y = 0.4927x - 0.041
R² = 0.9748

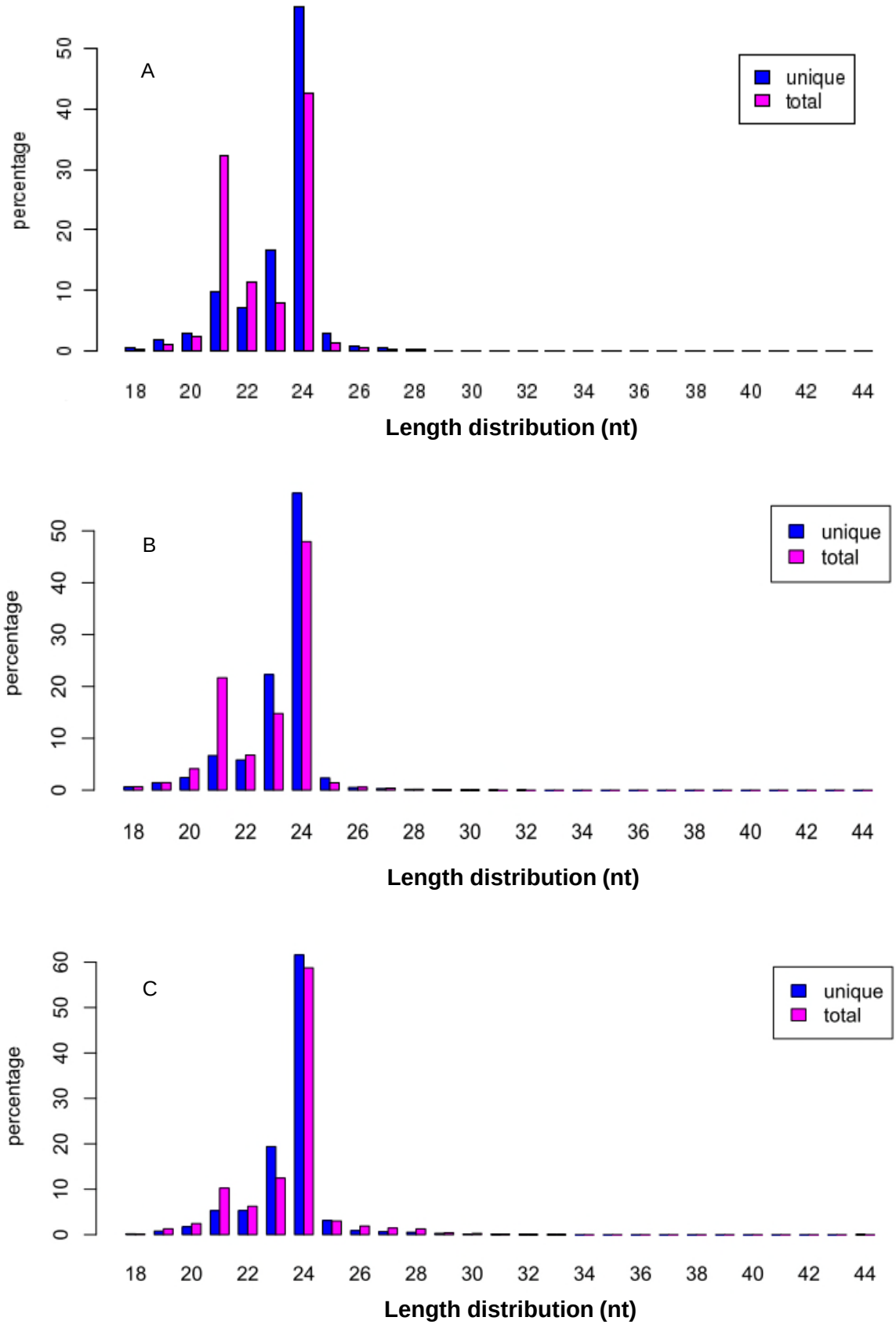
Figure 4


A
2.5
2
1.5
1
0.5
0
Minor nucleotide pair rate (%)
AU-GC
GC-AU
A
B
C
D
E
F
G
H
I
J
Site groups sorted by Pi (descending)
B
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
GC-AU > AU-GC (%)
y = 0.4486x^1.339
R² = 0.9688
0
0.2
0.4
0.6
0.8
1
Nucleotide diversity (Pi, %)

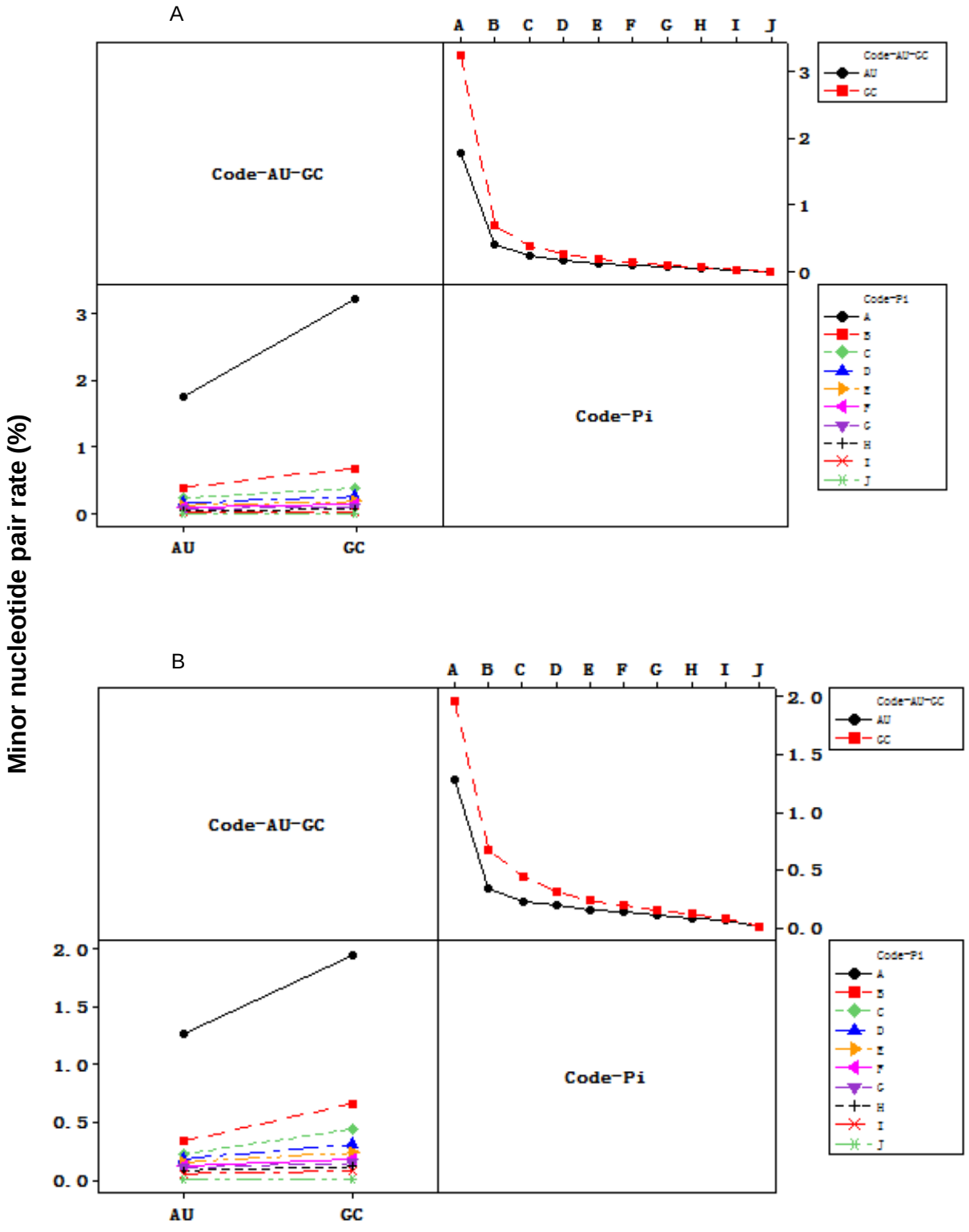
Supplementary Figure S1



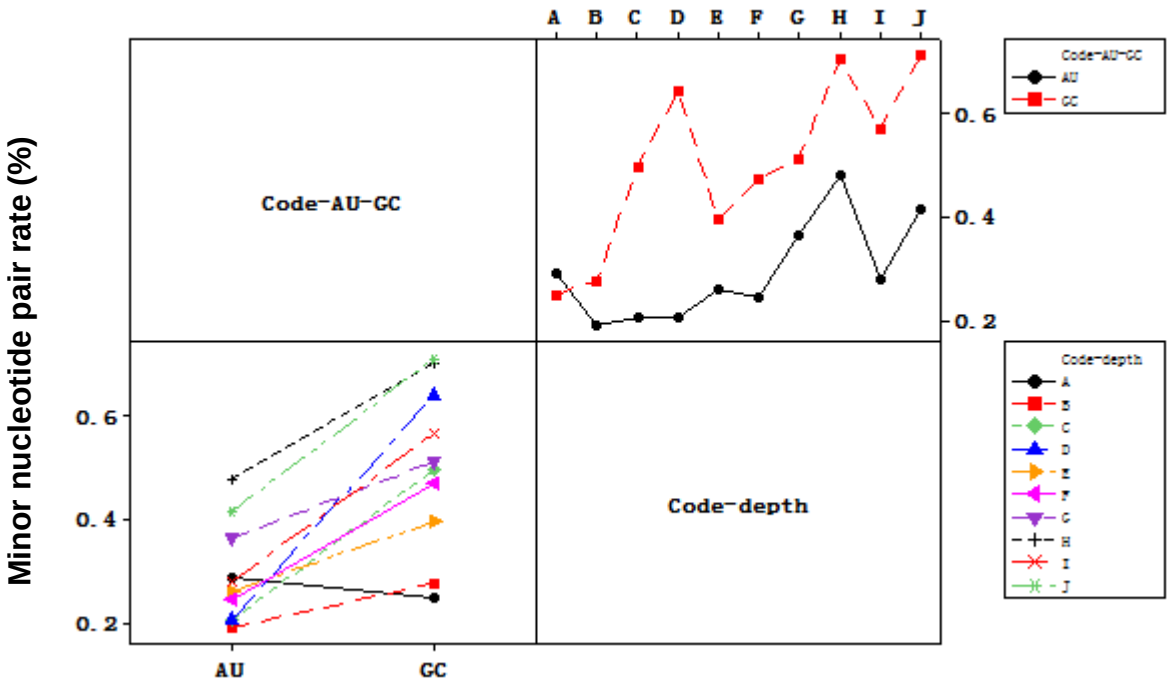
Supplementary Figure S2



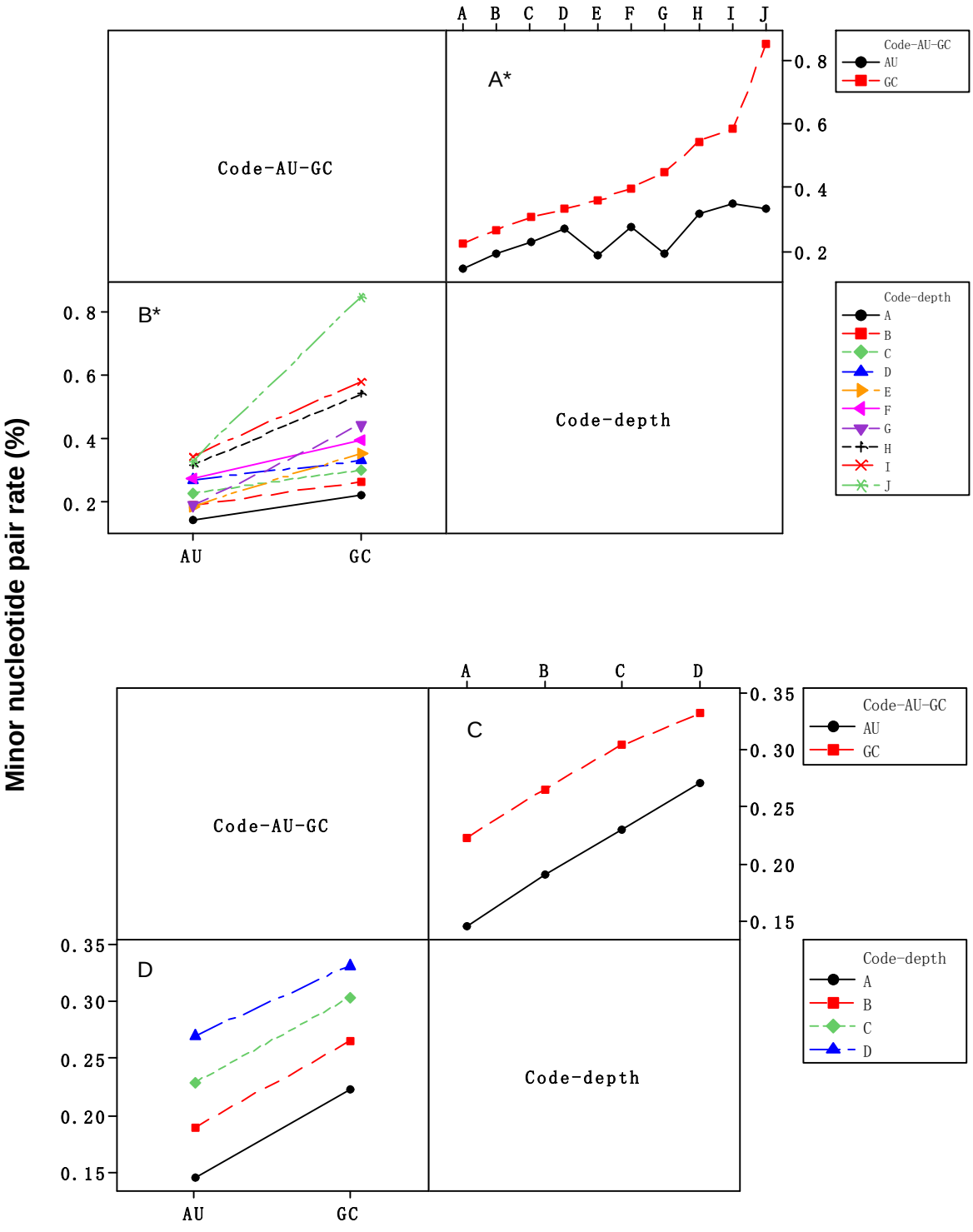
Supplementary Figure S3



Supplementary Figure S5



Supplementary Figure S4



Supplementary Figure S6