**Quantifying evolution of SNPs that affect RNA secondary structure in *Arabidopsis thaliana* genes**

Galen T. Martin[1], Christopher J. Fiscus[1] and Brandon S. Gaut[1, 2]

[1]Department of Ecology and Evolutionary Biology, UC Irvine, USA

[2] Author for correspondence:

Brandon S. Gaut

321 Steinhaus Hall

University of California, Irvine

Irvine, CA 92697

bgaut@uci.edu

**Keywords:** secondary structure, RNA biology, purifying selection, gene expression, bioclimate

**Running Head:** The evolution of SNPs that affect RNA secondary structure

**ABSTRACT**

Single-stranded RNA molecules have the ability to form intramolecular bonds between nucleotides and create secondary structures. These structures can have phenotypic effects, meaning mutations that alter secondary structure may be subject to natural selection. Here we examined the population genetics of these mutations within *Arabidopsis thaliana* genes. We began by identifying derived SNPs with the potential to alter secondary structures within coding regions, using a combination of computational prediction and empirical data analysis. We then examined nucleotide diversity and allele frequencies of these "pair-changing mutations" (pcM) in 1,001 *A. thaliana* genomes to infer selective pressures on these sites. The pcM SNPs at synonymous sites had an 8.8% reduction in nucleotide diversity relative to non-pcM SNPs at synonymous sites, and they were found at lower allele frequencies. We used demographic modeling to estimate selection coefficients from the pcM allele frequency spectrum, finding that pcMs had estimated selection coefficients ~3 to 5x smaller than nonsynonymous mutations. Additionally, we explored associations of pcMs with bioclimatic variables and gene expression. They exist at higher frequencies in higher latitudes and colder environments, with mean annual temperature explaining 51% of the variation in pcM frequency among subpopulations. Alleles containing pcMs had an average reduction of 137.4 normalized counts compared to genes with the ancestral secondary structure (mean expression = 3215.7 normalized counts). Overall, we conclude that structure-altering mutations are subject to intermediate levels of selection.

## INTRODUCTION

RNA molecules are single stranded (ssRNA), which gives them the ability to form Watson-Crick bonds, as well as less stable bonds (Varani and McClain 2000) between bases in the same molecule. This intramolecular base pairing, termed secondary structure, largely determines the three-dimensional shape of the molecule. The capacity for an RNA sequence to form secondary structures affects the function of transcribed regions of genomes in many ways (Vandivier et al. 2016). For example, secondary structures influence function by modulating translation (Kozak 1988; Svitkin et al. 2001), mRNA splicing (Buratti and Baralle 2004), ribozyme activity (Steitz and Moore 2003), localization (Bullock et al. 2010), and protein-RNA interactions (Williams and Marzluff 1995). Additionally, they affect the epigenetic fate of genes by influencing their RNA stability (Li et al. 2012), complement of small interfering RNAs (siRNAs) and DNA methylation (Martin et al. 2022). The ultimate impact of a transcribed genomic region on phenotype (Duan et al. 2003) and fitness (Innan and Stephan 2001) is therefore shaped by its capacity to form secondary structures; for example, mutations that affect mRNA structure in humans have been implicated in disease (Halvorsen et al. 2010). Yet, the evolutionary dynamics of mutations affecting secondary structures in mRNAs have received little attention in the evolutionary biology literature, with most such studies focusing on non-coding RNAs (Nowick et al. 2019).

One interesting and unexplored aspect of selection on secondary structure is its potential to contribute to adaptation. In protein coding genes, positive selection can, in theory, act on both the amino acid sequence and the mRNA secondary structure. However, selection is typically considered through the lens of mutational effects on proteins. Metrics like $d_N/d_S$ (the ratio of nonsynonymous to synonymous substitutions in genes) are used to identify loci under positive or purifying selection, but this approach does not account for possible selection on mRNA secondary structure, which is likely to affect both the numerator and the denominator of dn/ds. Consider, too, that fitness optima of secondary structures may change with the environment. For example, Ferrero-Serrano et al. (Ferrero-Serrano et al. 2022) recently demonstrated that two experimentally-validated structure-changing SNPs [often termed "riboSNitches" (Halvorsen et al. 2010)] caused different folding dynamics in cold vs. warm

environments. Thus, the fitness optima of secondary structures may vary with temperature and perhaps other environmental variables.

Selection on secondary structure could also have important methodological consequences for measuring selection with molecular data. This is because interpretation of $d_N/d_S$ ratios assumes that synonymous mutations are selectively neutral (M. Kimura 1968). Since the 1980s, evolutionary biologists have known that this is not entirely correct because codon usage is non-random (Ikemura 1981), and more recent studies have demonstrated strong non-neutral fitness effects from synonymous mutations (Lawrie et al. 2013; Lebeuf-Taylor et al. 2019). With regard to the fitness effects of secondary structure, it is known that (*i*) secondary structures within mRNA coding regions are more stable than expected under randomized codon usage (Seffens and Digby 1999), (*ii*) the location of synonymous substitutions is not random with respect to secondary structure stability (Chamary and Hurst 2005), (*iii*) codon usage is constrained towards weaker structure around miRNA-binding sites (Gu et al. 2012), and (*iv*) synonymous variants disrupting computationally-predicted secondary structure exist at reduced frequencies in human populations (Gaither et al. 2021), implying that purifying selection acts on these variants. Approaches like $d_N/d_S$ and the McDonald-Kreitman test (McDonald and Kreitman 1991) typically rely on the assumption that synonymous changes at degenerate sites are selectively neutral, but they have been shown to be sensitive to even weak selection on synonymous substitutions (Rahman et al. 2021). Depending on the strength and prevalence of RNA-level selection, accounting for secondary structure could be important for distinguishing neutral synonymous variants from less neutral variants.

Another reason such mutations are evolutionarily interesting is through the possibility of different selective effects between the RNA and protein “life stages” of gene expression. Nonsynonymous mutations can alter both amino acid sequence and secondary structure stability, potentially leading to a conflict between selection for protein function (protein-level selection) and mRNA stability (RNA-level selection) (Wegler et al. 2020). For example, a derived missense substitution may enhance the effectiveness of a protein but have an overall deleterious effect by compromising mRNA fitness through a less favorable secondary structure, potentially leading to improper splicing, translation, or reduced stability (Vandivier et al. 2016).

The frequency and importance of this potential pleiotropic antagonism depend on the relative strength of selection acting on mutations affecting secondary structure compared to mutations that affect amino acid sequence. These conflicts, if they exist, will constrain the efficacy of positive selection (Fraïsse et al. 2019).

Finally, while secondary structures serve important functions, particularly strong secondary structures have unique properties that may have negative effects on mRNA stability. Stable genic hairpins can cause genes to behave like pre-microRNA (miRNA) transcripts (Li et al. 2012), which form hairpin structures that are targeted by Dicer-like enzymes (Vergani-Junior et al. 2021) and are subsequently degraded into small RNAs. Like in pre-miRNA loci, elevated numbers of small interfering RNAs map to these structured genes (Li et al. 2012; Martin et al. 2023), putatively because their hairpins are degraded by Dicer-like enzymes. In turn, regions of miRNA-like secondary structure within genes correspond to high densities of small RNA mapping as well as high levels of small RNA-associated methylation, which often repress gene expression and function (Li et al. 2012). Given that small RNA mapping and repressive methylation levels are typically associated with silenced sequences, such as transposable elements, it is intriguing that many genes (up to 70% of annotated *Zea mays* genes) contain hairpin secondary structures (Martin et al. 2023). It is possible that these regions represent evolutionary conflicts between crucial secondary structure and downstream epigenetic effects.

In this study, we focus on the population genetics of mutations that are likely to affect secondary structure within the genes of the *Arabidopsis thaliana* (Arabidopsis) 1001 Genomes dataset (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium 2016). We begin by establishing a method to identify SNP variants that are within the ancestral secondary structures of expressed coding regions. There are generally two approaches to identify these structures, and neither is perfect. The first is empirical measurement. X-ray crystallography can accurately determine the structure of a transcript (Zhang and Ferré-D'Amaré 2014), but it is prohibitively expensive and infeasible to perform on a genome-wide level. Sequencing approaches such as double-stranded RNA (dsRNA) sequencing (Zheng et al. 2010; Li et al. 2012), structure-seq (Ding et al. 2014), and SHAPE-seq (Kwok et al. 2013; Liu et al. 2021) have also been used to estimate secondary

structures. These methods can be error prone, depend on coverage, and capture only a single possible secondary structure in a moment in time. The second approach is computational prediction [e.g., (Halvorsen et al. 2010; Lorenz et al. 2011; Zhang et al. 2020)], which is widely used but does not always recapitulate known structures from X-ray crystallography (Zhang et al. 2020). Nonetheless, newer prediction methods have become more accurate and have distinct advantages, such as the ability to integrate information across many possible secondary structures (Zhang et al. 2020).

Here we adopt a combined approach that uses both computational prediction and sequencing data to identify SNPs that may affect secondary structures in expressed coding region of *A. thaliana* and then calculate the population frequencies of these variants across the 1001 global Arabidopsis accessions to address four sets of questions. First, is there evidence that these mutations are under selection? That is, do they have evolutionary histories similar or dissimilar to putatively neutral synonymous mutations? Second, if they appear to be under selection, what is the inferred strength of selection compared to synonymous and nonsynonymous SNPs that are not within secondary structures? Third, is there any evidence to suggest that RNA-level selection meaningfully conflicts with protein-level selection or that they affect one possible phenotype, gene expression? Finally, previous work has suggested that secondary-structure altering SNPs may be associated with environment variables, particularly temperature. Is this true on a genome-wide scale?

## RESULTS

*Identifying unpaired mutations (upM) and pair changing mutations (pcM) mutations*

Identifying causative mutations that change the overall conformation of an RNA molecule is a complex and unsolved problem (Ferrero-Serrano et al. 2022). We developed a method to identify derived mutations with a high likelihood of being ancestrally paired (that is, hydrogen bonded to another nucleotide base in the ancestral *A. thaliana* genome) within secondary structures. We refer to such mutations as "**p**air **c**hanging **m**utations (pcM)", while those that are not ancestrally paired within stem:loop structures are "**u**n**p**aired **m**utations (upM mutations)." To classify these variants, we followed three steps (**Figure 1**). First, we polarized ancestral SNPs from the arabidopsis 1001 Genomes Project (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium 2016) using an *Arabidopsis lyrata* outgroup (see Methods). We used these ancestral SNPs to create an ancestral pseudo-transcriptome from the TAIR10 assembly (Berardini et al. 2015) by replacing derived alleles present in the *Col-0* reference genome with the inferred ancestral SNP. Second, we extracted mRNA sequences from the pseudo-ancestral reference and inferred base-pairing potentials within these sequences using LinearPartition (Zhang et al. 2020). LinearPartition calculates a partition function for a complete RNA sequence, and it sums equilibrium constants for all possible secondary structures for a sequence (i.e, not just the most likely structure). It outputs a base-pairing matrix that conveys the estimated probability that two bases pair. We focused only on bases with a high (>0.90) probability of pairing. Finally, we overlapped LinearPartition analyses with empirical data, namely dsRNA data generated by previously (Zheng et al. 2010). The data were generated from 6-week-old *Col-0* (flower bud clusters, leaves, and all aerial portions), and we used the data with the intention of distinguishing true paired bases from less-likely paired bases. In the sequencing data, we found that dsRNA overlapping regions were, on average, 26.9 nt long, and spanned a total of $1.9 \times 10^6$ nt, representing 1.7% of the total mRNA database.

Given both LinearPartition and dsRNA data, we defined derived pcM mutations as the subset of SNPs: i) that had a LinearPartition probability >0.9, ii) that were detected as paired in dsRNA data, and iii) whose presumed paired base did not also contain a complementary SNP.

For example, if the identified base contained an A to G mutation at position 1 and was found to be ancestrally paired with a T at position 20, the SNP at position 1 was not counted if position 20 contained a T toC substitution. This approach yielded three sets of derived SNPs: *i*) pair-changing mutation (pcM), which were predicted to alter base-pairing based on both empirical and computational evidence, *ii*) 'equivocal' SNPs that were predicted to alter base-pairing based on only one of the two prediction methods, and *iii*) unpaired mutations (upM) that had no evidence for being within secondary structures. After applying these rules, the pcM set consisted of 8,469 inferred mutations across 5,141 Arabidopsis genes (**Table 1**), representing a subset of 201,965 ancestrally paired bases. The upM set contained ~200-times more mutations than the pcM set (**Table 1**). Note, however, that the pcM SNPs likely do not reflect *all* of the bases involved in secondary structures, due to inevitable false negatives in our conservative approach. To address this concern, we also compared results from less conservative pcM/upM definitions (e.g., changing the pairing-probability cutoff and relying only on LinearPartition by not considering dsRNA overlap) to the pcM sets. The less conservative datasets yielded quantitatively similar results (Figs S3,4), and so for simplicity we focus on analyses of the pcM subset below.

*Prevalence and distribution of upM and pcM mutations*

If mutations that affect secondary structure are selectively important, one naive expectation is that their distribution across the genome differs from those mutations that do not affect structure. To profile and compare distributions among SNP types, we categorized derived SNPs based on their predicted impact based on SnpEff annotations (Cingolani et al. 2012). upMs were more likely to occur in untranslated regions (UTRs) and splice sites than pcMs, which were more common in coding regions (**Table 2**). pcMs were about half as likely to be found in the UTRs as expected under random distributions based on the length of features within genes—i.e., UTRs comprised ~15% of genic space but only ~8% of pcMs were within UTRs. Further, the percentage of 5' UTR pcM mutations (4.98%) was slightly lower than expected given the percentage of paired bases within 5' UTRs (6.86%).

These observations could be caused by two factors: i) paired bases could be less frequent in non-coding regions, making identification of pcM mutations less likely due to the distribution of structure, or ii) purifying selection that maintains ancestral secondary structure is stronger in these regions, so that pcM mutations are purged. To test the former neutral hypothesis (i), we applied a permutation test based on the set of 201,965 ancestrally paired bases (i.e., sites that are paired but do not have a derived SNP) and the 8,469 pcMs with observed SNPs. If the pcM distribution is due to the distribution of secondary structure in genes, then distribution of pcMs within genes should not be significantly different from the distribution of ancestrally paired bases (e.g., ~8% of ancestrally paired bases should exist in UTRs). To perform each permutation, we chose a random subset of n = 8,469 sites from the complete paired site dataset of n= 201,965. For each permutation, we counted the percentage of randomly-assigned sites in each genic component (5' UTR, coding region, and 3' UTR).We used these permutations to build distributions of random–pcM frequencies in each component. From this permutation test, we observed a significant difference ($P < 0.01$) from the anticipated divergence between paired site distribution and pcM distribution among genic components.

*Reduced nucleotide diversity at paired versus unpaired sites*

Nucleotide diversity at synonymous sites ($\pi_S$) tends to be higher than at nonsynonymous sites ($\pi_N$), which is generally interpreted as the result of purifying selection (Ingvarsson 2010; Osada 2015). One goal of this paper is to determine whether selection on pcM mutations is observable and comparable to selection on missense mutations. To do so, we focused only on synonymous pcM (syn_pcM) sites to avoid the confounding effects of selection on nonsynonymous substitutions. We therefore compared nucleotide diversity at segregating synonymous sites between the syn_pcM (n=3,214) and syn_upM (n=631,838) categories. We hypothesized that, if syn_pcM mutations are neutral, syn_pcM diversity ($\pi_{syn_pcM}$) should be equivalent to syn_upM diversity ($\pi_{syn_upM}$). Alternatively, if selection on paired bases is strong and similar to that of protein-level selection, $\pi_{syn_pcM}$ should be similar to nonsynonymous diversity ($\pi_N$) (n=864,006). We rejected the neutral hypothesis, because $\pi_{syn_pcM}$ was significantly lower than $\pi_{syn\text{-}upM}$, suggesting purifying selection on synonymous pcM sites.

However, $\pi_{syn_pcM}$ was also significantly higher than $\pi_N$ sites, putting it at an intermediate level (medians: $\pi_{syn_upM}$ = 8.2e-3, $\pi_{syn_pcM}$ = 7.1e-3, $\pi_N$ = 3.7e-3)(**Figure 2a**; t-test $P < 0.001$).

We also sought to determine whether selection on secondary structure interferes with selection for amino acid sequence. One factor that may influence this relationship is variation in selection across the length of genes: if different types of selection operate in distinct gene regions, potential interference between the two may be rare, irrespective of the relative strength of each selection type. For example, secondary structure is known to be particularly important at start codons and intron splice sites (Li et al. 2012; Vandivier et al. 2016), while amino acid sequence is more important towards the middle of the protein (Bricout et al. 2023).

To explore spatial distributions, we measured π at the three different types of sites across the length of gene coding sequences (**Figure 2b**). The distributions differed visually among site categories. $\pi_N$ was lowest at the middle of the coding sequence, while $\pi_{upM}$ was lowest towards the edges. The signal for $\pi_{syn_pcM}$ was noisy, likely owing to the low *n* of this category, but it dipped towards the 3' end of the coding sequence. We were curious whether these differences in the distribution of π related to specific feature types, such as translation start sites (start codons), stop codons, and splice junctions. We thus measured π in each category as a function of distance to each of these feature types but found that the correlation was not significant in each case (linear model $P>0.05$; **Figure S1**). Overall, patterns of π differed superficially for different categories, suggesting that selection at the protein and RNA levels may predominate in different regions of genes.

*Intermediate levels of selection on structural mutations*

Given results based on π, we predicted that mutations that putatively change secondary structure are generally more deleterious than synonymous changes that do not affect secondary structure. To examine this prediction further, we calculated the site frequency spectra (SFS) of various classes of mutations and then used the SFS to infer the strength of selection. Focusing first on the calculated SFS for different site categories, the distribution for pcM sites was skewed towards lower frequency alleles compared to upM sites, with more singletons and fewer intermediate and fixed mutations. The set of pcM mutations is a small

subset of total SNPs, so we tested for statistical significance between SFSs in two ways: a Kolgorov-Smirnov test and a permutation-based approach (**Figure 3c-d; Figure S2**). Both methods found that the SFS for pcM mutations at both synonymous and nonsynonymous sites differed significantly from the SFS of syn_upM mutations. The effect was stronger among nonsynonymous mutations: that is, nonsynonymous pcM (non_pcM) mutations had a stronger skew toward rare variants than syn_pcM mutations (permutation nonsyn. $P \simeq 0$ vs syn. $P = 0.01$). Finally, we note that the SFS of mutations in the equivocal class (i.e., those which were identified by either the computational or sequencing approach, but not both) fell between the pcM and upM distributions (**Figure 3a-b**).

We evaluated the robustness of these results by investigating datasets based on alternative definitions of pcM and upM. First, we considered sites identified as likely to be paired by LinearPartition, without filtering by dsRNA overlap. For both synonymous (n = 19,252) and nonsynonymous (26,021) pcM, allele frequencies remained significantly different between pcM and upM (**Figure S3**). Second, we loosened cutoffs for base-pairing probabilities inferred by LinearPartition, while continuing to filter by dsRNA cutoff. With a base-pairing probability threshold of 0.50 (opposed to the original 0.90), both synonymous (n = 20,596) and nonsynonymous (57,349) pcM allele frequencies remained significantly different between pcM and upM (**Figure S4**).

We used SFS information to infer the strength of selection on pcM mutations using ∂a∂I (Gutenkunst et al. 2009; Kim et al. 2017). To do so, we first inferred a simplified demographic history of European Arabidopsis populations using the synonymous SFS at unpaired sites. We found that demographic models with a recent bottleneck and a two-epoch model were similarly robust to the data (maximum log composite likelihoods = -3.23 and -3.26 respectively), with modern/ancestral population size ratio estimates between 21.5x and 43.1x with population size expansions ~0.3 to 0.6*2NAncestral generations ago. These inferences of recent population size-change fits well with previous demographic inference of Arabidopsis, which is thought to have experienced several bottlenecks and expanded from refugia after the last glacial maximum ~20 KYA (François et al. 2008). We then used both demographic models to estimate distributions of fitness effects (DFE) from the SFS of pcM SNPs and compared them to

the DFE of upM missense SNPs. By doing so, we aimed to partition the fitness effects of changes in secondary structure (syn_ pcM sites) from *non*-structure related fitness effects (nonsynon_upM). Given our *a priori* expectations about amino acid changes having greater effects than mRNA secondary structure changes, we expected the DFE to be weaker among synonymous pcM mutations than among nonsynonymous mutations, and this was the case (**Figure 3**). From the inferred gamma DFE distributions, we estimated that synonymous pcM sites had a mean scaled selection coefficient ($\gamma$, $2N_{Ancestral}S$) = 15.81 for the two-epoch model, and $\gamma$ = 318.48 for the bottleneck model. These mean effect sizes were ~3 to 5-fold smaller than the effect sizes for nonsynonymous sites, which had corresponding estimates of $\gamma$ = 52.21 for the two-epoch model and $\gamma$ = 1570.28 for the bottleneck model. Note, however, that the mean $\gamma$ estimate for synonymous pcM mutations was different from neutrality ($\gamma < 1$).

*The potential for pleiotropy between mRNA and protein stages of gene expression*

Our analyses indicate that pcM mutations have quantifiable fitness effects. But is selection on secondary structure strong enough to interfere with selection at the protein level? To estimate how frequently this type of interference may occur, we performed simple simulations of selection effects from protein and RNA secondary structure in mutations of genic coding regions. The simulations consisted of four steps. First, we began with 24,820 genes representing the length of genes that have Ensembl-based *A. lyrata* orthologs (mean length = 1275.4 nt). Second, for each gene we assumed that mutations at 60% of potential sites represented non-synonymous changes (as in (Motoo Kimura 1968)). Third, since our empirical analyses indicated that 0.43% of nonsynonymous sites had pcM mutations (**Table 1**), we randomly assigned this proportion of nonsynonymous mutations to affect secondary structure. At these nonsynonymous sites, we assigned two fitness effects: a "protein-level" fitness effect ($\gamma_{protein}$) and a "RNA-level" fitness effect ($\gamma_{structure}$). The two fitnesses were assigned by drawing a value from the corresponding nonsynonymous (all sites) and syn_pcM DFE gamma distributions. Finally, we tallied a number of metrics from each simulation, such as the expected number of genes with non-neutral $\gamma > 1$ nonsynonymous mutations, the number of genes

expected to have nonsynonymous sites subject also to constraints on pairing, and the number of sites where RNA-level selection, as reflected by $\gamma_{structure}$, was larger than $\gamma_{protein}$.

From these observed/inferred values and assumptions, we estimated that there were ~5.06 x $10^4$ sites across ~20,000 genes where the selection coefficient for the amino acid change was higher than that for a pcM mutation. Such sites were found among 73% of the genes. In contrast, a smaller but still substantial number of ~2.59 x $10^4$ sites encompassed the opposite case, where $\gamma_{structure} > \gamma_{protein}$. These sites were found in 54% (or 13,403) genes. While these estimates are subject to numerous caveats of our approach, they do suggest that even the small fitness effects observed among pcM mutations in this dataset can affect sites across many genes. Although the average γ for syn_pcM mutations was an order of magnitude smaller than that of nonsynonymous mutations, pleiotropic interference was not proportional to this average difference. That is, RNA-level selection overwhelmed protein-level selection only half as frequently as the inverse, rather than one-tenth as frequently as implied by the DFE.. Of course, the evolutionary importance of pleiotropic effects depends on the efficacy of selection, and these differences may only infer meaningfully at extremely large population sizes.

*Structure-changing SNPs in the geographic landscape*

We have shown that pcMs are under purifying selection. Given previous work on two experimentally-validated riboSNitches (Ferrero-Serrano et al. 2022), we tested the hypothesis that the frequency of pcMs covaried with temperature. To do this, we first summarized environmental variation across individuals using principal component analysis (PCA). The first three PCs explained 36.02%, 31.33%, and 12.93% of the variation in climate, respectively. We included these first three PCs in mixed-linear models to predict the number of all pcMs, syn_pcMs and nonsyn_pcMs. We did not detect any significant associations between all pcMs and nonsyn_pcMs, but syn_pcMs were significantly associated with PC 2 ($p = 0.00099$). PC 2 also appears to be geographically defined, with a strong correlation with latitude (Spearman's rho = 0.82, $p = <2.2 \times 10^{-16}$), and a moderate correlation with longitude (Spearman's rho = 0.34, $p <2.2 \times 10^{-16}$ , **Figure 5a**). To understand the relative contribution of bioclimatic variables to PC 2, we examined the loadings of each bioclimatic variable on this axis. The bioclimatic variables

related to temperature (BIO1 to BIO11) disproportionately contributed to PC 2 compared to the precipitation-related bioclimatic variables; e.g., the absolute value of cumulative loadings across temperature variables was 2.48 on PC2 compared to 1.24 for precipitation variables (**Figure 5b)**. Supporting this observation, annual mean temperature (BIO 1) explained 69% of variation in PC 2 in a simple linear model, while annual precipitation (BIO 12) only explained ~9% (**Figure 5c**). Taken together, our analyses suggest that pcM has an environmental basis that is primarily explained by differences in temperature across the Eurasian range of *A. thaliana*.

We also assessed frequencies of derived pcM alleles in geographically distinct subpopulations. To define subpopulations, we used admixture groups determined from previous analyses of the 1001 Genomes (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium 2016). For each subpopulation, we estimated the mean frequency of derived syn_pcM alleles. To compare population frequencies to climate data, we extracted climatic variables for each individual in each subpopulation based on its geographical coordinates and then calculated the mean of each climatic variable in each subpopulation. We found that climatic variables related to temperature were negatively correlated with the frequency of pcM alleles within subpopulations (**Figure S6a-b**). The effect was especially strong for BIO1 (mean annual temperature; linear model $P = 0.03$, $R^2 = 0.51$) and BIO6 (minimum temperature of the coldest month; $R^2 = 0.54$, $P = 0.024$). Other temperature-related variables, such as isothermality (BIO3) and maximum temperature (BIO5) were uncorrelated ($P > 0.05$)

*Expression effects of structure-changing SNPs*

One potential explanation for the different selective signatures of pcM vs upM SNPs is that structural mutations change the RNA abundance of gene transcripts, either by modulating the stability of the RNA molecule or the efficiency of translation (Li et al., 2012; Vandivier et al., 2016). We thus predicted that gene expression differed between alleles containing derived pcM mutations with a gene versus alleles that do not carry the pcM mutation. To investigate this effect on a genome-wide level, we used the expression data from the 1001 Genomes dataset and constructed a linear model with mixed effects similar to Muyle et al. (Muyle et al. 2021)

(see Methods). The model measured within-gene expression differences between alleles and tested for the significance of the allelic state (pcM vs. no pcM) across all genes, ignoring genes without pcM SNPs. Genome-wide, our model detected that pcM alleles had significantly lower levels of expression compared to non-pcM alleles (mean difference = 137.4 normalized counts; P = 0.001)(Figure S7A). The effect was not always consistent among genes, however, because 58.5% of allelic genes had lower expression in pcM alleles, while 41.4% had higher expression in these genes.

**DISCUSSION**

Using a combination of bioinformatic and experimental data, we have identified derived mutations in *Arabidopsis thaliana* that are likely to affect base-pairing in secondary structures. Such mutations have long been known to have functional effects (Wan et al. 2014); for example, they contribute prevalently to human genetic disease (Halvorsen et al. 2010; Lin et al. 2020). A subset of these SNPs may act as riboSNitches that affect both secondary structure and gene expression (Ferrero-Serrano et al. 2022) in a temperature dependent manner (Su et al. 2018). Here we have identified derived mutations that may change pairing between bases within CDS regions, based on both bioinformatic predictions and experimental data, and studied their evolutionary dynamics, including allelic expression and environmental associations.

By classifying mutations at putative ancestrally unpaired sites (upM) versus paired sites (pcM), we found that signatures of purifying selection maintain ancestral base pairing, leading to reduced diversity at paired sites (**Figure 2A**) and skewed allele frequencies (**Figure 3AB**). We also compared the strength of selection on pcMs compared to nonsynonymous unpaired mutations (upMs). Perhaps not surprisingly, selection on secondary structure is estimated to be weaker than selection on amino acid changes; depending on the demographic model, we have found that the γ for these derived pcM alleles was, on average, about 20.3% to 30.3% that of derived nonsynonymous substitutions. However, selection on pcM sites was detectably stronger than on non-pairing synonymous sites (**Figure 4**). Our finding is consistent with previous empirical work showing the existence of strongly deleterious synonymous variants in

*Drosophila* (Lawrie et al. 2013) and with population genetic estimates in humans that showed reduced frequencies of alleles predicted to change secondary structure (Gaither et al. 2021). Overall, our results suggest that secondary structure is a weak but prevalent factor shaping the evolution of Arabidopsis genes.

**Caveats and Limitations of Identifying Pair-Changing SNPs**

Although our analyses generally establish selection on pcM mutations in Arabidopsis, we must also recognize several important caveats concerning our methodology. First, we do not have precise or accurate knowledge about the individual effects of each allele on the broader secondary structure of an mRNA. To help focus on bonafide SNP variants that affect structure, we excluded alleles with potentially compensatory mutations within the same gene. Nonetheless, the identification of riboSNitches remains a complex and unresolved challenge (Corley et al. 2015; Ferrero-Serrano et al. 2022), with only a score having been experimentally validated across species (Corley et al. 2015; Assmann et al. 2023). Our goal was therefore to identify a set of segregating SNPs with evidence that they may disrupt ancestral secondary structures.

Our process for identifying ancestral secondary structures relied on several assumptions. Specifically, we based our analyses on a pseudo-ancestral genome that is based on the *Col-0* reference and an *A. lyrata* outgroup. This approach is likely to mis-assign ancestral states for a subset of sites, introducing uncertainty. Similarly, our analyses are likely biased toward studying secondary structures that are present in the *Col-0* reference, because we used dsRNA data from *Col-0* in our identification pipeline. We have also assumed specific criteria to classify base-pairing sites, including a cutoff probability > 0.9 with LinearPartition and empirical dsRNA evidence. Although our general conclusions hold with relaxed criteria, such as lower LinearPartition base-pairing probabilities, we suspect that our strict criteria misidentifies many *bona fide* pcMs as either equivocal or upMs. However, the net effect of these misclassifications should underemphasize differences among site classes (pcM, equivocal and upM) making comparisons among categories inherently conservative.

Finally, there are other methods to estimate base-pairing, including different bioinformatic tools and additional empirical data, that could, in theory, be applied to pcM identification. We chose LinearPartition because it performed reliably and efficiently relative to other secondary structure prediction programs (Zhang et al. 2020), including in extensive comparisons to RNAfold (Lorenz et al. 2011) focused on secondary structure prediction (Martin et al. 2023). With respect to empirical data, we chose to use the dsRNA data, even though Arabidopsis is one of only two plant species with available genome-wide structure-seq information (Ding et al. 2014; Deng et al. 2018). The Arabidopsis structure-seq data was based on the Col-0 reference, so it could not be easily applied to our inferred ancestral genome and many genes were not represented in the structure-seq dataset.

**Paired mutations and their inferred selective effects**

It has long been known that synonymous mutations are not entirely neutral (Ikemura 1981), but our findings present evidence that mutations influencing ancestral secondary structure exert a detectable impact on fitness. The population genetic evidence for these impacts are three-fold: i) an 8.8% reduction of nucleotide diversity ($\pi$ ) at syn_pcM sites compared to syn_upM sites (**Figure 2A**), ii) a skewed SFS relative to syn_upM sites (**Figure 3A**), and iii) an inferred DFE that differs from syn_upM sites. As importantly, we have compared DFEs between syn_pcMs and nonsynonymous, non-pairing SNPs (non_upMs), revealing that selection on syn_pcMs falls somewhere between neutrality and selection on nonsyn_upM sites. Of course, all of these analyses have strengths and weaknesses, particularly the inference of the DFE. Our *dadi*-based inferences rely on demographic models that fit the data reasonably well but could be incorrect. However, we inferred similar relative magnitudes of selection based on two different models, suggesting that the inferences are reasonably robust to demographic inferences. More generally, DFE inference approaches, including *dadi* and other methods that rely on analysis of McDonald-Kreitman tables, do not deal adequately with the effects of linked selection. Although the effects on DFE inference can be complex, linked selection should tend to make inferences among site classes more similar than expected.

Our results - i.e., that secondary structure imposes constraint on synonymous sites - is similar to results from experimental evolution in microbial systems (Chursov et al. 2013; Bailey et al. 2021) and other work, using phylogenetic approaches, have shown that slower rates of evolution of synonymous sites correlate with both gene expression and the strength of secondary structure (Park et al. 2013). Our results do differ somewhat from work in Drosophila, however. Lawrie et al (2013) studied the population genetics of synonymous sites and found that many were under strong purifying selection, particularly in highly conserved and highly expressed genes (Lawrie et al. 2013). Yet, in a preliminary analysis they were unable to correlate strength of selection with secondary structure. One intriguing possibility is that Drosophila and Arabidopsis differ with respect to the significance of secondary structure in their genes. Testing this hypothesis may not be easily feasible, given the inherent ambiguity in defining structures within species. Nonetheless, considering the intertwined relevance of secondary structure and small RNAs (Li et al. 2012), coupled with the distinct small RNA production dynamics between plant and animal genomes, this prospect merits further investigation.

Our work has provided a preliminary framework for studying the locational biases of pcMs as well as the interplay between selection at the RNA versus protein life stages of a gene. With regard to location, previous work in Arabidopsis has shown that the interruption of secondary structures in 3' UTRs destabilize mRNAs (Zhang et al. 2024). This work is consistent with our finding that the 3' regions of genes have marked reduction of $\pi$ in 3' regions of genes (**Figure 2B**), including UTRs, suggesting that selection against pair-changing mutations is particularly strong in these regions. This may not be a universal feature of plant genes, however, because rice and maize seem to have stronger folding dynamics in 5' UTRs (Deng et al. 2018; Martin et al. 2023) compared to Arabidopsis (Deng et al. 2018). With respect to the RNA versus protein life stages, we performed draws from the syn_pcM and non_upM DFE distributions to assess how often the magnitude of selection against secondary structural changes was higher than for a nonsynonymous change. The results were surprising, because it showed that, at the site level, many nonsynonymous mutations may have larger fitness effects due to secondary structure effects. Further disentangling the numerous (perhaps contradictory)

pressures shaping genome evolution will require integrating structural dynamics into molecular and population genetic analysis.

**Potential Explanations for Associations between paired mutations and climate**

In a landmark study, Su et al (Su et al. 2018) presented a compelling demonstration of the potential importance of secondary structures within plant genes. They subjected rice (*Oryza sativa* L.) seedlings to a high temperature, eliciting a heat-shock response, and then found that the RNAs of ~14,000 genes unfolded over the experimental temperature range. As a class, these genes also demonstrated shifts in gene expression, which they attributed primarily to more rapid degradation of unfolded RNAs (as opposed to reduced translation rates). Their work suggested that SNPs that modify secondary structures could be more or less tolerated, based on temperature and climate. That is, selection against pcMs may not be as strong for plants that reside in regions of moderate (as opposed to high) temperatures. Their results also suggest that RNA folding is a vital component of gene expression, so that one expects correlations between the presence of secondary structure altering SNPs and shifts in gene expression.

To explore these threads in our work, we first examined the distribution of pcM variants across the sampling landscape of the Arabidopsis 1001 Genomes dataset. We investigated the association of alleles and climate across individuals and subpopulations (i.e., previously inferred admixture groups). Both provide genome-wide evidence that derived pcM mutations are more common at locations with lower temperatures, as measured by bioclimatic variables (**Figure 5**). This is the first demonstration of this genome-wide pattern and provides an opportunity to consider the evolutionary forces that contribute to such a pattern. We can think of two reasonable explanations: local adaptation and deleterious load. Previous work has argued convincingly that some associations between pcMs and climate likely represent local adaptation events (Ferrero-Serrano and Assmann 2019), but we favor the deleterious load interpretation for our genome-wide pattern, for two reasons. First, derived, deleterious pcM mutations may be less strongly selected against in low temperature environments where strong-folding may not be as critical. Second, deleterious mutations tend to accumulate at the edges of geographic ranges, due to smaller population sizes and allele surfing, particularly in

expanding regions of a species' range (Travis et al. 2007; Excoffier et al. 2009; Angert et al. 2020). Visually, we find that higher pcM counts occur in the Northern and Eastern edges of the sampled range (**Figure 5C)**, perhaps representing expanding edges from Ice Age refugia.

Finally, we sought to determine whether alleles with pcMs exhibit different levels of gene expression than those without and found significantly lower expression in pcM alleles compared to genes with ancestral secondary structure. This genome-wide result, across all sampled genes, mimics similar results in microbial systems where the disruption of secondary structures reduces gene expression (Bailey et al. 2021). However, there was also wide variation across genes, because >40% of genes showed the opposite pattern (i.e., pcM alleles had higher expression). One explanation for this variation is experimental noise - indeed, we think it remarkable that there was a significant overall trend given that alleles in most genes undoubtedly differ by more than just the presence/absence of a pcM, although it is feasible that the disruption of particularly strong secondary structures. One biological explanation for higher expression of pcM alleles is that the mutations that disrupt especially strong secondary structures may also interrupt RNA-interference (Li et al. 2012). No matter the cause, we have shown that pair-changing mutations are common, that they are under moderate levels of purifying selection, and that one potential cause is that perturbation of secondary structures alters gene expression dynamics.

## MATERIALS and METHODS

*Identification of derived upM and pcM mutations*

We used the 1001 Genomes Project v.3.1 (https://1001genomes.org/data/GMI-MPI/releases/v3.1/) SNP calls (Lamesch et al. 2012a) and variant annotations (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium 2016) for all analyses. We filtered the dataset to retain biallelic SNPs and assigned ancestral and derived states by aligning the *A. lyrata* v1.0 genome assembly (Hu et al. 2011) to the *A. thaliana* TAIR10 reference (Lamesch et al. 2012b) using AnchorWave 1.0 (Song et al. 2022). This approach enabled us to polarize 5,613,812 of 12,883,854 (43.6%) SNPs.

To identify sites that may contribute RNA secondary structures, we first constructed a pseudo-ancestral genome by replacing derived sites in the TAIR10 assembly with their corresponding ancestral alleles from the polarized VCF using GATK FastaAlternateReferenceMaker v3.7 (McKenna et al. 2010). Then, we extracted the longest mRNA (coding) sequence for each protein-coding gene from the pseudo-ancestral reference using bedtools2 getfasta 2.27.1 (Quinlan and Hall 2010) before estimating RNA folding for each sequence with LinearPartition v1.0 (Zhang et al. 2020). We then selected SNP sites that overlapped with positions with base-pairing probability > 0.9 as determined by LinearPartition for further analysis. We verified putative pairing sites by determining the overlap with dsRNA sequencing data generated from wildtype flower buds (NCBI Gene Expression Omnibus GSE23439) (Zheng et al. 2010). Since the dsRNA data was mapped to the TAIR9 assembly, we converted to TAIR10 assembly coordinates (Lamesch et al. 2012a) using CrossMap 0.6.4 (Zhao et al. 2014). We considered putative pair changing mutations (pcM) with both computational and empirical evidence (i.e. high base pairing probability in LinearPartition analysis and dsRNA coverage). We further filtered these sites by finding overlap with potential compensating mutations using the base pairing probability files from LinearPartition; to do so, SNPs were excluded if the paired base position also contained an alternative allele with base-pairing compatibility with the derived allele. All overlaps of genomic features were calculated using the GenomicRanges R package 1.48.0 (Lawrence et al. 2013).

*Nucleotide diversity, allele frequency, and DFE analysis*

We estimated nucleotide diversity (π) for synonymous, nonsynonymous, pcM, equivocal, and unpaired (upM) sites using VCFtools v0.1.16 (Danecek et al. 2011). First, we extracted subsets of the full 1,001 genomes VCF file for each category (syn. pcM, nonsyn. pcM, etc.) using samtools tabix and the annotations from our paired/unpaired site identification above and the 1,001 genomes SnpEff file. We calculated π with the per-site method in each gene using VCFtools. We measured distance between sites and various genic features (starts, stops, and intron junctions) using GenomicRanges in R.

Site frequency spectra were calculated using a custom R script with vcfR v1.15.0 (Knaus and Grünwald 2017), data.table v1.15.2 (Barrett et al. 2024), and tidyverse 2.0 (Kuhn and Wickham 2020) R packages. Permutation tests for differences between spectra were done by sampling # upM sites, building a SFS for each sample, calculating the difference between the sample SFSand the true upM SFS, and repeated this procedure for 10,000 iterations. We considered sites 2 kbp away from focal sites to be distal and all other sites to be proximal.

To estimate the distributions of fitness effects for different mutation types, we used the fitdadi python package (Gutenkunst et al. 2009; Kim et al. 2017). We first inferred demography using the synonymous upM SFS. We analyzed various demographic models with various starting parameters to optimize for the most robust demographic models. Our final demographic models for the DFE used the following parameters: two-epoch: nu = 26.9, T = 0.113; and bottlegrowth-1d: nuB = 0.00634688, nuF = 2.60504245, and T=0.0762755. We estimated the DFE using the synonymous pcM SFS and the nonsynonymous pcM+upM SFS separately, modeling DFE as a gamma distribution. We plotted DFEs in python using matplotlib (Hunter May-June 2007). We estimated the mean of each gamma distribution by multiplying the shape*scale parameter of each.

*Pleiotropy simulation*

To investigate the potential for conflicts between protein and RNA-level selection, we started with a collection of 27,206 genes multiplied the CDS lengths of each *A. thaliana* gene with an *A. lyrata* ortholog (downloaded from Ensembl) by 0.60 to approximate the number of

nonsynonymous sites across the genome. For each gene, we assigned each site as either paired or unpaired based on the probability data from **Table 1**. We then assigned each site a "protein-level" fitness effect (γ) by pseudo-randomly drawing a value from the nonsynonymous DFE gamma distribution (shape = 0.47, scale = 111). This selection coefficient assignment was pseudo-random, because the maximum value of assignments was capped at 1,000 ($2N_AS$). We then assigned paired sites an "RNA-level" fitness effect by the same method, but this time sampling from gamma distribution representing the DFE of synonymous pcM SNPs (shape = 0.22, scale = 73). We evaluated the accuracy of our DFE simulations by comparing the means of these sampled DFE to the "true" means estimated from the gamma distributions (shape*scale). We repeated the simulation several times, finding that the derived solutions changed minimally.

*Geospatial and climatic correlations*

We studied the association between environmental variables and the number of synonymous pcM SNPs at both subpopulation and individual scales. For the environmental variables, we used the 19 WorldClim 2 bioclimatic variables at 2.5 minute resolution, which summarize past climate averages from 1970-2000 (Fick and Hijmans 2017). Bioclimatic values for each accession were extracted using the collection coordinates reported by the 1001 Genomes Project (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium 2016) and the raster 3.6-26 R package (Hijmans 2023). For the subpopulation-based approach, we considered the 10 subpopulations defined by (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium 2016) and fit both simple linear models and generalized linear models in R (R Core Team 2023) to predict the mean allele frequency across synonymous pcM alleles using the mean of each bioclimatic variable for each subpopulation.

For the individual-scale approach, we first did a principal component analysis of the bioclimatic variables using the prcomp function in R (R Core Team 2023) and then fit mixed linear models using the lmekin function from the coxme 2.2-20 R package (Therneau 2024) to

test for association between between the first three PCs and the number of pcM alleles across pcM sites per accession. A centered relatedness matrix calculated from all biallelic SNPs using gemma 0.98.5 (Zhou and Stephens 2012) was included as a random effect in the models. We corrected the p-values using the Bonferroni method and assessed significance at α = 0.05.

*Expression analyses*

Expression data in the form of log normalized counts was downloaded from the NCBI Gene Expression Omnibus (GSE80744) (Kawakatsu et al. 2016), which includes expression data for 24,175 genes in the 727 Salk accessions from the 1,001 Genomes dataset. pcM overlap with genes was determined using the GenomicRanges library in R (Lawrence et al., 2013). Allelic state for each accession was determined using the 1,001 Genomes VCF file. Genes with no pcM allele were excluded from the analysis, and only two allelic states were considered: whenever an accession contained one or more pcMs in the gene, it was considered a pcM allele, irrespective of whether the pcM was the same SNP between alleles (e.g., if an accession contained a pcM at one position within a gene, and another accession contained a pcM at a different position, both were put into the same category of "pcM alleles"). The mixed-effect linear model was analyzed using the R package lme4 (Bates et al. 2015), and included pcM allelic state as a fixed effect and gene identity as a random effect, expressed as:

$$log(Gene\ expression + 1) \sim pcM\ presence\ +\ (1|Gene)$$

**DATA AVAILABILITY:**

All of the data used in this study are freely available. Custom scripts can be found at github.com/GalenTMartin/structure_selection and extra files can be found at https://figshare.com/projects/RNA_structure_evolution/221926. The NCBI Gene Expression Omnibus numbers for dsRNA and gene expressions were GSE23439 and GSE80744.

**ACKNOWLEDGEMENTS:**

We thank Edwin Solares for computing infrastructure support. This work was supported by NSF grant IOS-2414478 to BSG.

**REFERENCES:**


1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, 1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166:481–491.

Angert AL, Bontrager MG, Ågren J. 2020. What Do We Really Know About Adaptation at Range Edges? Annu. *Annu. Rev. Ecol. Evol. Syst.* 51:341–361.

Assmann SM, Chou H-L, Bevilacqua PC. 2023. Rock, scissors, paper: How RNA structure informs function. *Plant Cell* 35:1671–1707.

Bailey SF, Alonso Morales LA, Kassen R. 2021. Effects of synonymous mutations beyond Codon bias: The evidence for adaptive synonymous substitutions from microbial evolution experiments. *Genome Biol. Evol.* [Internet] 13. Available from: http://dx.doi.org/10.1093/gbe/evab141

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T. 2024. data.table: Extension of `data.frame`. Available from: https://CRAN.R-project.org/package=data.table

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models Usinglme4. *J. Stat. Softw.* 67:1–48.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* 53:474–485.

Bricout R, Weil D, Stroebel D, Genovesio A, Roest Crollius H. 2023. Evolution is not Uniform Along Coding Sequences. *Mol. Biol. Evol.* [Internet] 40. Available from: http://dx.doi.org/10.1093/molbev/msad042

Bullock SL, Ringel I, Ish-Horowicz D, Lukavsky PJ. 2010. A'-form RNA helices are required for cytoplasmic mRNA transport in Drosophila. *Nat. Struct. Mol. Biol.* 17:703–709.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell. Biol.* 24:10505–10514.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.

Chursov A, Frishman D, Shneider A. 2013. Conservation of mRNA secondary structures may filter out mutations in Escherichia coli evolution. *Nucleic Acids Res.* 41:7854–7860.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6:80–92.

Corley M, Solem A, Qu K, Chang HY, Laederach A. 2015. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic Acids Res.* 43:1859–1868.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

Deng H, Cheema J, Zhang H, Woolfenden H, Norris M, Liu Z, Liu Q, Yang X, Yang M, Deng X, et al. 2018. Rice In Vivo RNA Structurome Reveals RNA Secondary Structure Conservation and Divergence in Plants. *Mol. Plant* 11:607–622.

Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. 2014. In vivo genome-wide

profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505:696–700.
Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* 12:205–216.
Excoffier L, Foll M, Petit RJ. 2009. Genetic Consequences of Range Expansions. *Annu. Rev. Ecol. Evol. Syst.* 40:481–501.
Ferrero-Serrano Á, Assmann SM. 2019. Phenotypic and genome-wide association with the local environment of Arabidopsis. *Nat Ecol Evol* 3:274–285.
Ferrero-Serrano Á, Sylvia MM, Forstmeier PC, Olson AJ, Ware D, Bevilacqua PC, Assmann SM. 2022. Experimental demonstration and pan-structurome prediction of climate-associated riboSNitches in Arabidopsis. *Genome Biol.* 23:101.
Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS. *Int. J. Climatol.* 37:4302–4315.
Fraïsse C, Puixeu Sala G, Vicoso B. 2019. Pleiotropy Modulates the Efficacy of Selection in Drosophila melanogaster. *Mol. Biol. Evol.* 36:500–515.
François O, Blum MGB, Jakobsson M, Rosenberg NA. 2008. Demographic history of european populations of Arabidopsis thaliana. *PLoS Genet.* 4:e1000075.
Gaither JBS, Lammi GE, Li JL, Gordon DM, Kuck HC, Kelly BJ, Fitch JR, White P. 2021. Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. *Gigascience* 10:giab023.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.
Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on Synonymous Sites for Increased Accessibility around miRNA Binding Sites in Plants. *Mol. Biol. Evol.* 29:3037–3044.
Halvorsen M, Martin JS, Broadaway S, Laederach A. 2010. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* 6:e1001074.
Hijmans RJ. 2023. raster: Geographic Data Analysis and Modeling. Available from: https://CRAN.R-project.org/package=raster
Hunter JD. May-June 2007. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9:90–95.
Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43:476–481.
Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.* 151:389–409.
Ingvarsson PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in Populus tremula. *Mol. Biol. Evol.* 27:650–660.
Innan H, Stephan W. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. *Genetics* 159:389–399.
Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al. 2016. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana

Accessions. *Cell* 166:492–505.
Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics* 206:345–361.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.
Kimura M. 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles*. *Genet. Res.* 11:247–270.
Knaus BJ, Grünwald NJ. 2017. vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* 17:44–53.
Kozak M. 1988. Leader length and secondary structure modulate mRNA function under conditions of stress. *Mol. Cell. Biol.* 8:2737–2744.
Kuhn M, Wickham H. 2020. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. Available from: https://www.tidymodels.org
Kwok CK, Ding Y, Tang Y, Assmann SM, Bevilacqua PC. 2013. Determination of in vivo RNA structure in low-abundance transcripts. *Nat. Commun.* 4:2971.
Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012a. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40:D1202–D1210.
Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012b. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40:D1202–D1210.
Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9:e1003118.
Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in D. melanogaster. *PLoS Genet.* 9:e1003527.
Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. 2019. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife* [Internet] 8. Available from: http://dx.doi.org/10.7554/eLife.45952
Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. 2012. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* 24:4346–4359.
Lin J, Chen Y, Zhang Y, Ouyang Z. 2020. Identification and analysis of RNA structural disruptions induced by single nucleotide variants using Riprap and RiboSNitchDB. *NAR Genom Bioinform* 2:lqaa057.
Liu Z, Liu Q, Yang X, Zhang Y, Norris M, Chen X, Cheema J, Zhang H, Ding Y. 2021. In vivo nuclear RNA structurome reveals RNA-structure regulation of mRNA processing in plants. *Genome Biol.* 22:11.
Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6:26.
Martin G, Solares E, Muyle A, Bousios A, Gaut BS. 2022. Diverse patterns of secondary structure across genes and transposable elements are associated with siRNA production and epigenetic fate. *bioRxiv* [Internet]:2022.10.17.512609. Available from: https://www.biorxiv.org/content/10.1101/2022.10.17.512609v1
Martin GT, Solares E, Guardado-Mendez J, Muyle A, Bousios A, Gaut BS. 2023. miRNA-like secondary structures in maize (Zea mays) genes and transposable elements correlate with

small RNAs, methylation, and expression. *Genome Res.* [Internet]. Available from: http://dx.doi.org/10.1101/gr.277459.122

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–654.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.

Muyle A, Ross-Ibarra J, Seymour DK, Gaut BS. 2021. Gene body methylation is under selection in Arabidopsis thaliana. *Genetics* [Internet] 218. Available from: http://dx.doi.org/10.1093/genetics/iyab061

Nowick K, Walter Costa MB, Höner Zu Siederdissen C, Stadler PF. 2019. Selection Pressures on RNA Sequences and Structures. *Evol. Bioinform. Online* 15:1176934319871919.

Osada N. 2015. Genetic diversity in humans and non-human primates and its evolutionary consequences. *Genes Genet. Syst.* 90:133–145.

Park C, Chen X, Yang J-R, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.* 110:E678–E686.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Rahman S, Pond SLK, Webb A, Hey J. 2021. Weak selection on synonymous codons substantially inflates *dN/dS* estimates in bacteria. *Proceedings of the National Academy of Sciences* 118:e2023575118.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org/

Seffens W, Digby D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* 27:1578–1584.

Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. 2022. AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 119. Available from: http://dx.doi.org/10.1073/pnas.2113075119

Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* 28:411–418.

Su Z, Tang Y, Ritchey LE, Tack DC, Zhu M, Bevilacqua PC, Assmann SM. 2018. Genome-wide RNA structurome reprogramming by acute heat shock globally regulates mRNA abundance. *Proc. Natl. Acad. Sci. U. S. A.* 115:12170–12175.

Svitkin YV, Pause A, Haghighat A, Pyronnet S, Witherell G, Belsham GJ, Sonenberg N. 2001. The requirement for eukaryotic initiation factor 4A (eIF4A) in translation is in direct proportion to the degree of mRNA 5' secondary structure. *RNA* 7:382–394.

Therneau TM. 2024. coxme: Mixed Effects Cox Models. Available from: https://CRAN.R-project.org/package=coxme

Travis JMJ, Münkemüller T, Burton OJ, Best A, Dytham C, Johst K. 2007. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Mol. Biol. Evol.* 24:2334–2343.

Vandivier LE, Anderson SJ, Foley SW, Gregory BD. 2016. The Conservation and Function of RNA

Secondary Structure in Plants. *Annu. Rev. Plant Biol.* 67:463–488.
Varani G, McClain WH. 2000. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep.* 1:18–23.
Vergani-Junior CA, Tonon-da-Silva G, Inan MD, Mori MA. 2021. DICER: structure, function, and regulation. *Biophys. Rev.* 13:1081–1090.
Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505:706–709.
Wegler C, Ölander M, Wiśniewski JR, Lundquist P, Zettl K, Åsberg A, Hjelmesæth J, Andersson TB, Artursson P. 2020. Global variability analysis of mRNA and protein concentrations across and within human tissues. *NAR Genom. Bioinform.* 2:lqz010.
Williams AS, Marzluff WF. 1995. The sequence of the stem and flanking sequences at the 3' end of histone mRNA are critical determinants for the binding of the stem-loop binding protein. *Nucleic Acids Res.* 23:654–662.
Zhang H, Zhang L, Mathews DH, Huang L. 2020. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics* 36:i258–i267.
Zhang J, Ferré-D'Amaré AR. 2014. New molecular engineering approaches for crystallographic studies of large RNAs. *Curr. Opin. Struct. Biol.* 26:9–15.
Zhang T, Li C, Zhu J, Li Y, Wang Z, Tong C-Y, Xi Y, Han Y, Koiwa H, Peng X, et al. 2024. Structured 3' UTRs destabilize mRNAs in plants. *Genome Biol.* 25:54.
Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30:1006–1007.
Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, Cao K, Wang L-S, Gregory BD. 2010. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* 6:e1001141.
Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–824.

**FIGURE LEGENDS**

**Figure 1.** Schematic representation of upM/pcM identification method. Three SNPs are shown at positions 6, 12, and 27 within a hypothetical gene: one that is equivocal (6, *left*), one that is an unpaired mutation (upM) (position 12, *middle*) and one that is a pair-changing mutation (pcM) (position 27, *right*). Bases with high (>0.90) LinearPartition base-pairing probabilities are shown as parentheses. Both dsRNA and LinearPartition evidence were required to designate a base as pcM.

**Figure 2.** Nucleotide diversity at synonymous paired (pcM), synonymous unpaired (upM), and nonsynonymous sites. (A) Violin plot of nucleotide diversity at different site types, with the black dots representing mean diversity. Statistical contrasts are provided above the violin plots, based on t-tests. (B) π calculated in windows across the length across all analyzed CDS regions. The x-axis represents length-standardized windows across the span of all analyzed genes from the 5' end (transcription start site, TSS) to the 3' end (transcription termination site. TTS). Missense refers to nonsynonymous bases not identified as pair-changing mutations; paired refers to synonymous pcM mutations; unpaired refers to synonymous upM mutations.

**Figure 3.** (A) Unfolded site frequency spectra (SFS) showing derived allele frequencies of synonymous alleles categorized by their inferred effect on ancestral secondary structure. (B) SFS of nonsynonymous alleles in each category. Permutation distributions for differences between paired and unpaired SFS at synonymous sites (C) (black dot shows missense for scale) and nonsynonymous (D). Violins represent the distribution of differences in random samples (same *n* as paired sites) from the unpaired data, while points show the true differences. Differences (y-axis) were calculated as the percentage of alleles in each pcM SFS bin subtracted from the percentage in the same upM SFS bin (e.g., {0-0.1] in pCM minus {0-0.1] in upM etc.)

**Figure 4.** The inferred distribution of fitness effects (DFE) for mutation types under two demographic models. Distributions of the scaled selection coefficient, γ ($2*N_{ancestral}S$) are governed by shape and scale parameters estimated from dadi under a gamma distribution and were inferred from the SFS of synonymous upM (i.e., the subset of mutations unlikely to have an effect at either the protein or RNA level).

**Figure 5.** pcM associations with geospatial and environmental variables. (A) Map of Eurasian 1,001 Genomes accessions across their geographical range, color-scaled by PC 2 values. Accessions are clearly differentiated by latitude, with those in the north having higher values than more southern accessions. (B) Loading scores of environmental variables on PC 2. Greater

values indicate that variables contribute more to PC 2, and variables are colored by class (temperature vs precipitation). Temperature-associated variables tend to have greater loading values than precipitation. (C) Linear models of PC 2 as a function of environmental variables. BIO1 (mean annual temperature) explains more of the variation in PC 2 than BIO 12 (annual precipitation).

**TABLES**

**Table 1**. SNPs categorized by effect on amino acid sequence and secondary structure.

| | upM | pcM | Equivalent[1] |
|---|---|---|---|
| Nonsynonymous | 3,790 | 864,006 | 22,231 |
| Synonymous | 3,214 | 631,838 | 16,038 |

[1]*SNPs detected by the computational method (LinearPartition) or dsRNA coverage, but not both.*

**Table 2**. SnpEff annotations for upM vs pcM SNPs.

| SNP effect | Number upM | Percentage upM in SnpEff category[1] | Number pcM | Percentage pcM in SnpEff category[1] | Percentage difference (upM - pcM) |
|---|---|---|---|---|---|
| synonymous variant | 631838 | 27.23% | 3214 | 37.95% | -10.72% |
| missense variant | 864006 | 37.23% | 3790 | 44.75% | -7.52% |
| disruptive inframe deletion | 2825 | 0.12% | 19 | 0.22% | -0.10% |
| inframe insertion | 2437 | 0.11% | 11 | 0.13% | -0.02% |
| inframe deletion | 3449 | 0.15% | 14 | 0.17% | -0.02% |
| frameshift variant+start lost | 450 | 0.02% | 3 | 0.04% | -0.02% |
| frameshift variant+stop gained | 222 | 0.01% | 2 | 0.02% | -0.01% |
| frameshift variant+stop lost | 377 | 0.02% | 2 | 0.02% | -0.01% |
| initiator codon variant | 231 | 0.01% | 1 | 0.01% | 0.00% |
| stop retained variant | 1315 | 0.06% | 4 | 0.05% | 0.01% |
| disruptive inframe insertion | 808 | 0.03% | 2 | 0.02% | 0.01% |
| start lost | 1232 | 0.05% | 2 | 0.02% | 0.03% |
| stop lost | 1532 | 0.07% | 2 | 0.02% | 0.04% |
| splice acceptor variant | 3389 | 0.15% | 8 | 0.09% | 0.05% |
| splice donor variant | 3537 | 0.15% | 3 | 0.04% | 0.12% |
| stop gained | 15577 | 0.67% | 41 | 0.48% | 0.19% |
| frameshift variant | 24822 | 1.07% | 64 | 0.76% | 0.31% |
| 5' UTR premature start codon gain variant | 18977 | 0.82% | 42 | 0.50% | 0.32% |
| 5' UTR variant | 161299 | 6.95% | 422 | 4.98% | 1.97% |
| splice region variant | 94490 | 4.07% | 55 | 0.65% | 3.42% |
| 3' UTR variant | 276907 | 11.93% | 681 | 8.04% | 3.89% |
| intron variant | 210835 | 9.09% | 87 | 1.03% | 8.06% |

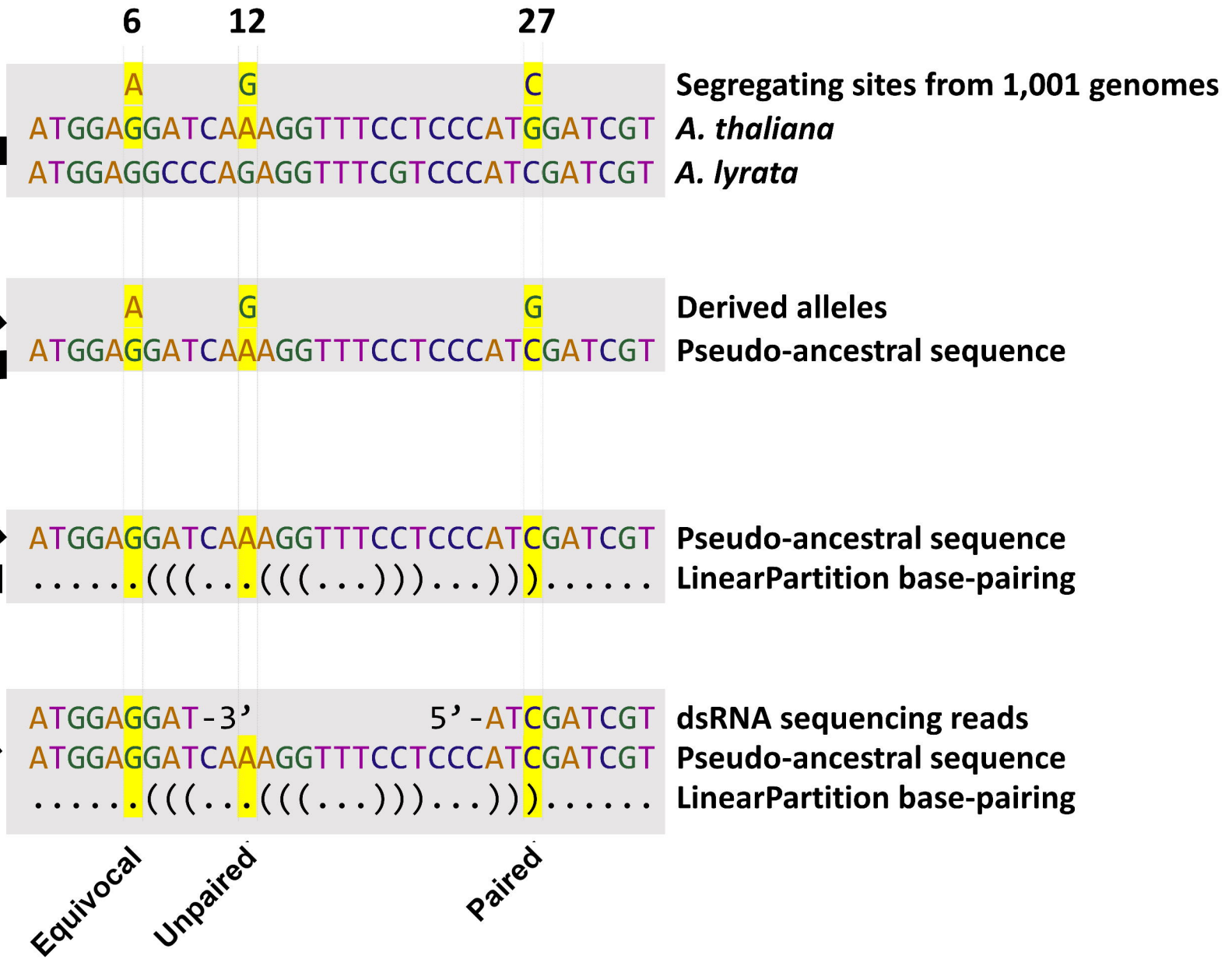
| | | |
|---|---|---|
| **Total** | **232055 5** | **8469** |

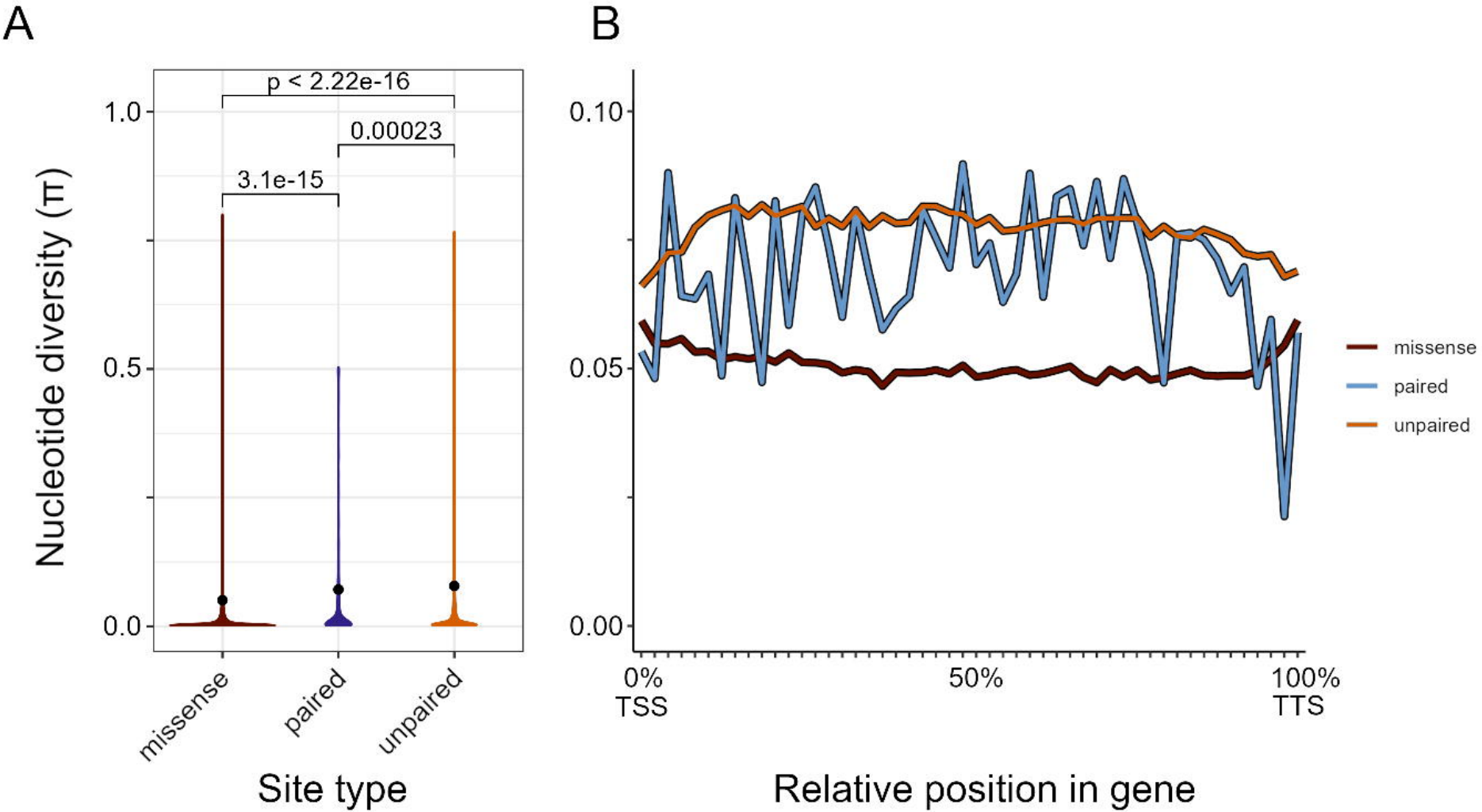
[1]Percentage of SNPs in category (upM or pcM) that had a particular SNP effect.

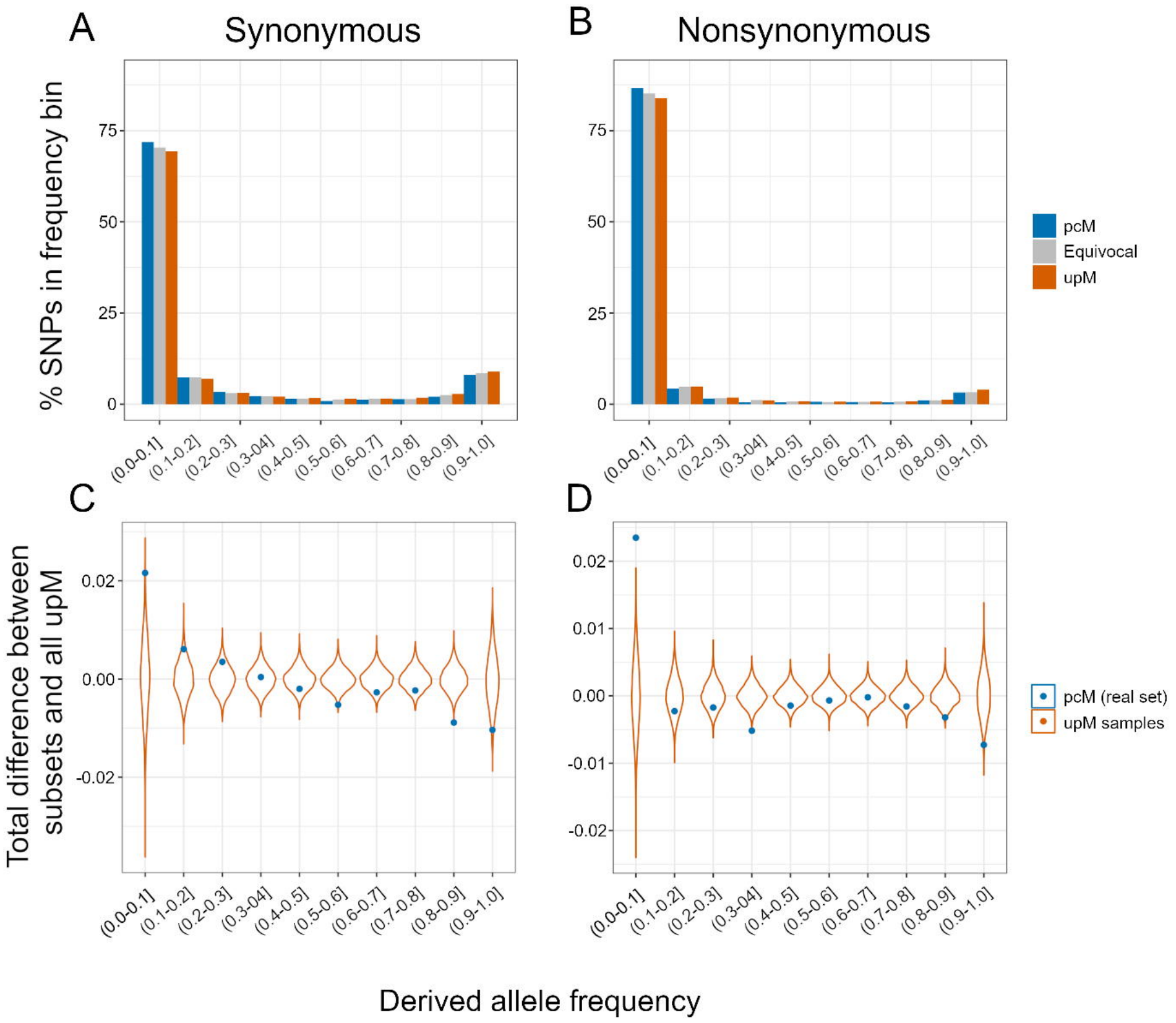
**1.** Polarize SNPs from outgroup alignment, then replace derived alleles in reference with ancestral allele

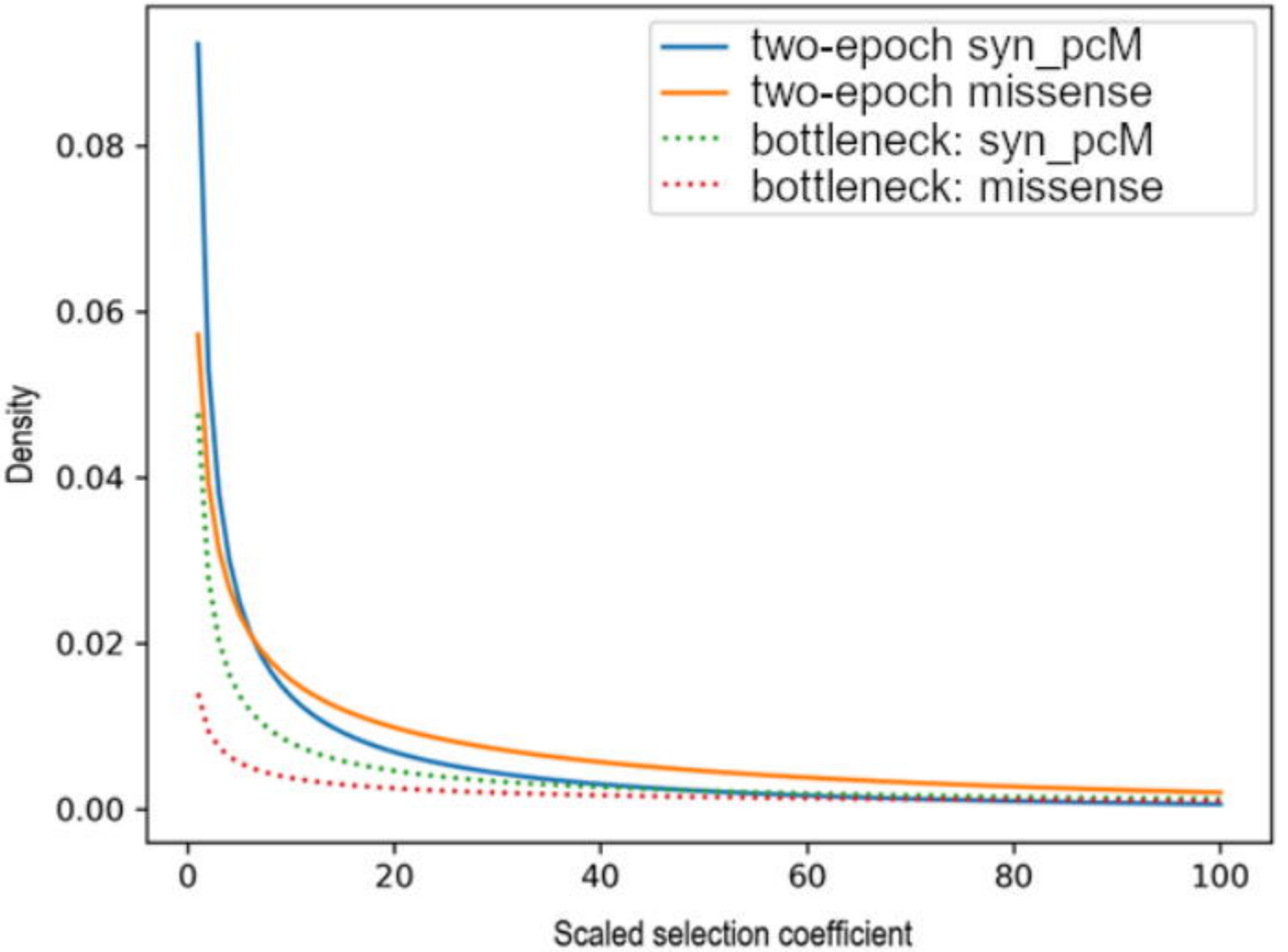
**2.** Use pseudo-ancestral sequence to computationally infer base-pairing

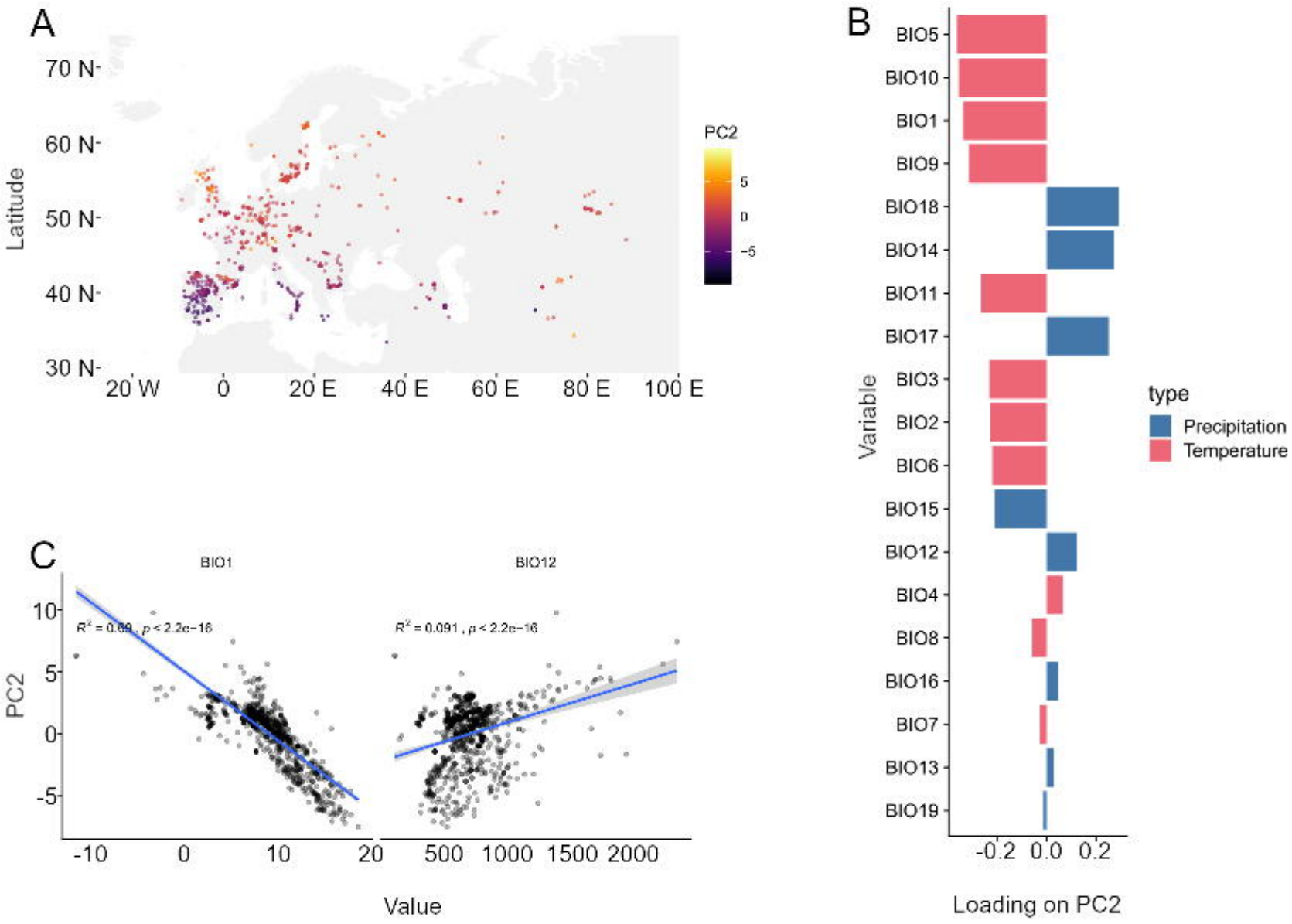
**3.** Look for overlaps between dsRNA sequences and computationally inferred paired bases




A
p < 2.22e-16
0.00023
3.1e-15
1.0
0.5
0.0
Nucleotide diversity (π)
missense
paired
unpaired
Site type
B
0.10
0.05
0.00
0%
TSS
50%
100%
TTS
Relative position in gene
missense
paired
unpaired


A
Synonymous
B
Nonsynonymous
% SNPs in frequency bin
75
50
25
0
(0.0-0.1]
(0.1-0.2]
(0.2-0.3]
(0.3-04]
(0.4-0.5]
(0.5-0.6]
(0.6-0.7]
(0.7-0.8]
(0.8-0.9]
(0.9-1.0]
pcM
Equivocal
upM
C
D
Total difference between subsets and all upM
0.02
0.00
-0.02
0.01
-0.01
pcM (real set)
upM samples
Derived allele frequency


two-epoch syn_pcM
two-epoch missense
bottleneck: syn_pcM
bottleneck: missense
Density
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Scaled selection coefficient


A
70 N
60 N
50 N
40 N
30 N
Latitude
20 W
0
20 E
40 E
60 E
80 E
100 E
PC2
5
0
-5
B
BIO5
BIO10
BIO1
BIO9
BIO18
BIO14
BIO11
BIO17
BIO3
BIO2
BIO6
BIO15
BIO12
BIO4
BIO8
BIO16
BIO7
BIO13
BIO19
Variable
-0.2
0.0
0.2
Loading on PC2
type
Precipitation
Temperature
C
BIO1
BIO12
$R^2 = 0.69$, $p < 2.2e-16$
$R^2 = 0.091$, $p < 2.2e-16$
PC2
10
5
0
-5
-10
0
10
20
500
1000
1500
2000
Value