Supplementary Information

Additional File 1: Figures S1-S6

Human *Alu* elements promote the establishment and enhancement of piRNA-protein-coding gene targeting relationships

Chong He[1*], Hao Zhu[123*]

[1] Bioinformatics Section, School of Basic Medical Sciences, Southern Medical University, Guangzhou, 510515, China

[2] Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Southern Medical University, Guangzhou, 510515, China

[3] Guangdong Provincial Key Lab of Single Cell Technology and Application, Southern Medical University, Guangzhou, 510515, China

[*] Corresponding author: biohe@foxmail.com; zhuhao@smu.edu.cn

Figure S1

Figure S1 The *Alu*(TE)-association of sense piRNA/piRNA-similar sequence pairs. In Fig. 2, The *Alu*(TE)-asscociation of antisense piRNA/piRNA-similar sequence pairs has been shown. Here, indices corresponding to those in Fig. 2 are shown for sense piRNA/piRNA-similar sequence pairs.

Figure S2

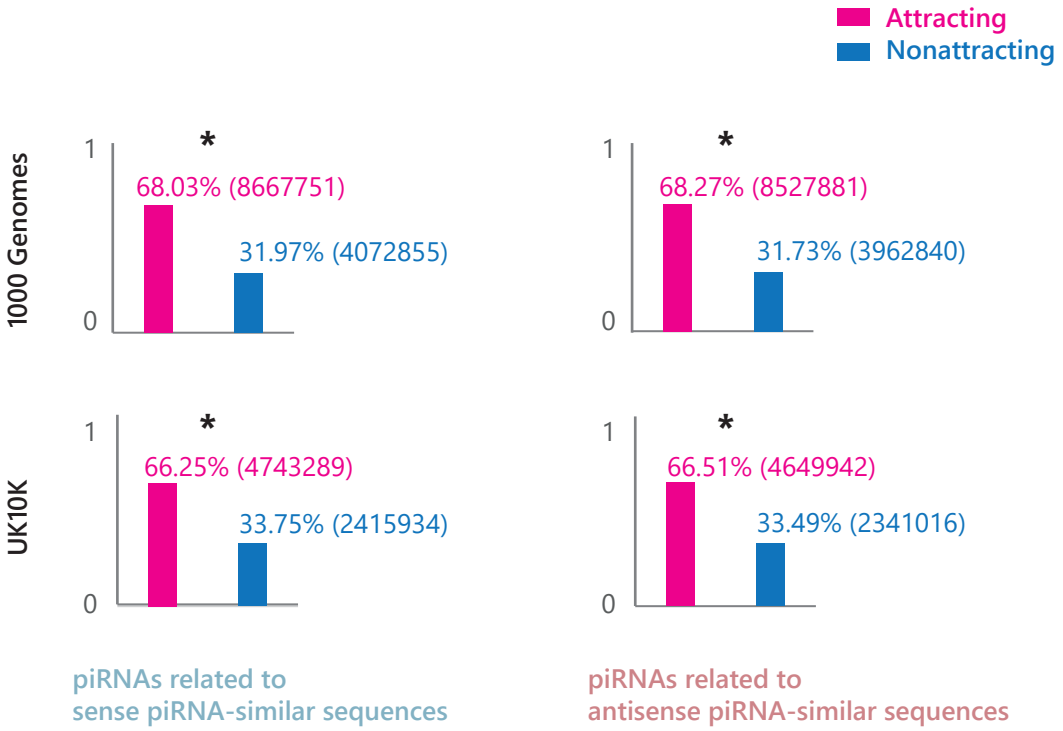

Figure S2 The proportion of occurrences of attracting SNVs in genomic locations of piRNAs. An asterisk represents that the observed proportion of occurrences of attracting SNVs significantly deviates from the expectation of 1/3 ($p < 0.05$, binomial test). SNVs at all possible genomic locations of piRNAs related to piRNA-similar sequences were counted. Similar to SNVs in piRNA-similar sequences, the excess of attracting SNVs was detected both in sense and antisense piRNA/piRNA-similar sequence pairs. This phenomenon persisted in both the 1000 Genomes data and the UK10K data.

Figure S3

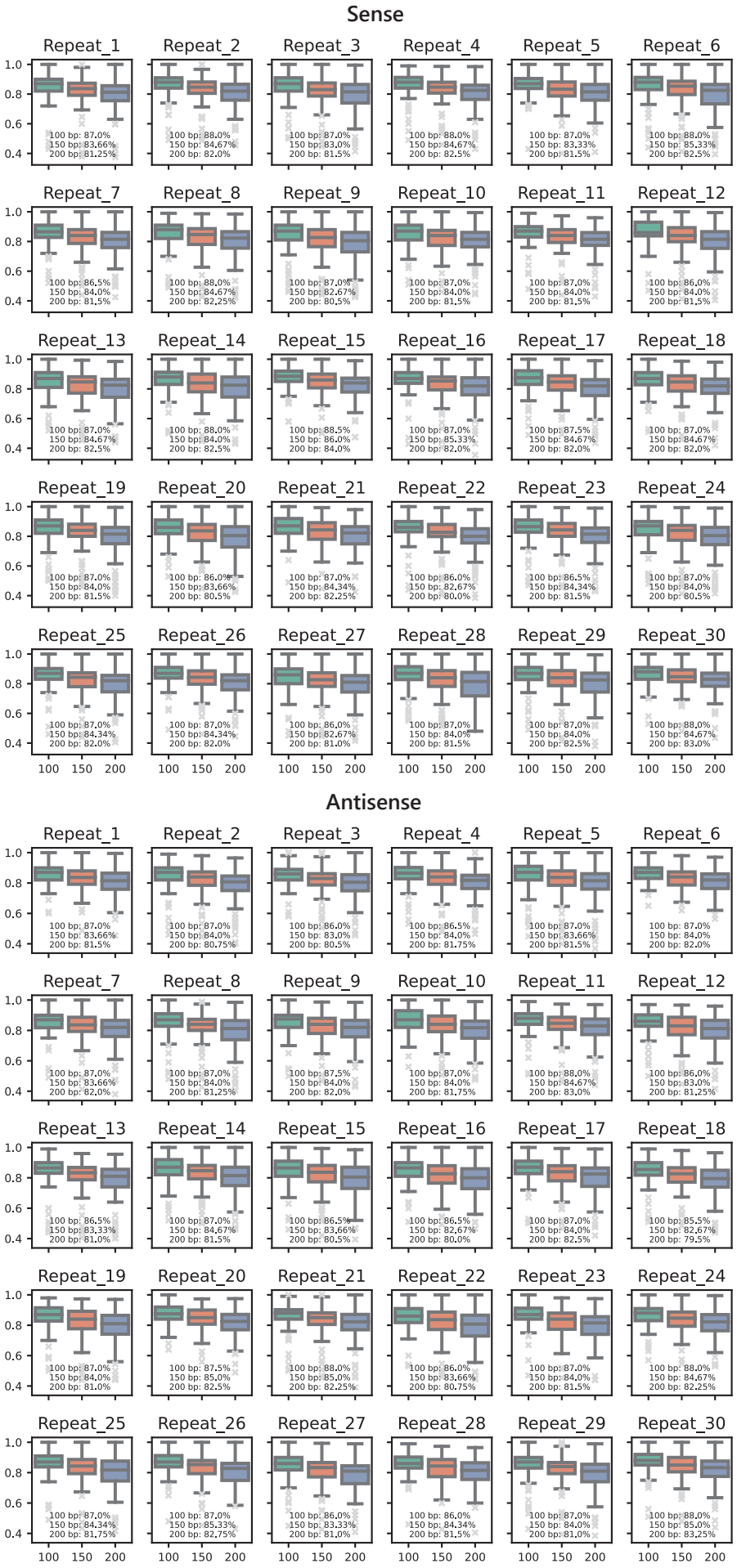

Figure S3 The medians of maximum similarity have been ploted in Fig. 5. Here, we show the distribution of maximum similarity for each repeat. As is the custom, each box plots the first, second (median) and third quatiles of maximum similarity. The text in each boxplot shows the median values of sequence similarity that are ploted in Fig. 5.

Figure S4

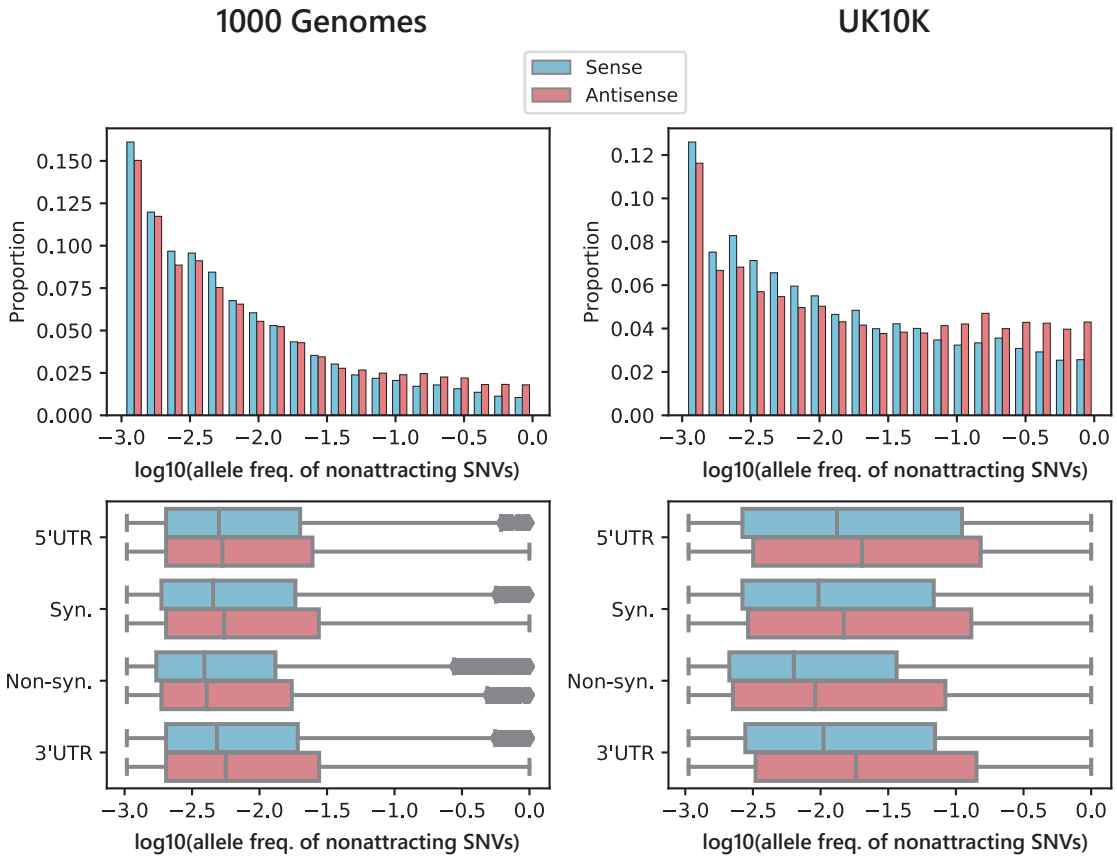

Figure S4 Allele frequencies of nonattracting SNVs rather than attracting SNVs are shown here. Allele frequencies were restricted to ≥ 0.001. In the top panels, each histogram shows the overall distribution of allele frequencies for sense/antisense nonattracting SNVs, pooling all categories of variants together (at a log10 scale). In the bottom panels, the distribution of allele frequencies for sense/antisense nonattracting SNVs is shown separately for each category of SNVs (at a log10 scale). For all categories of variants, whether the strand relationship is sense or antisense, the allele frequencies of attracting SNVs are higher than those of nonattracting SNVs (compare with Fig. 6; $p < 0.05$, Mann–Whitney U test). Addtionally, different from the observation in attracting SNVs, antisense nonattracting alleles did not show a tendency towards low frequencies in comparison with sense nonattracting alleles.

Figure S5

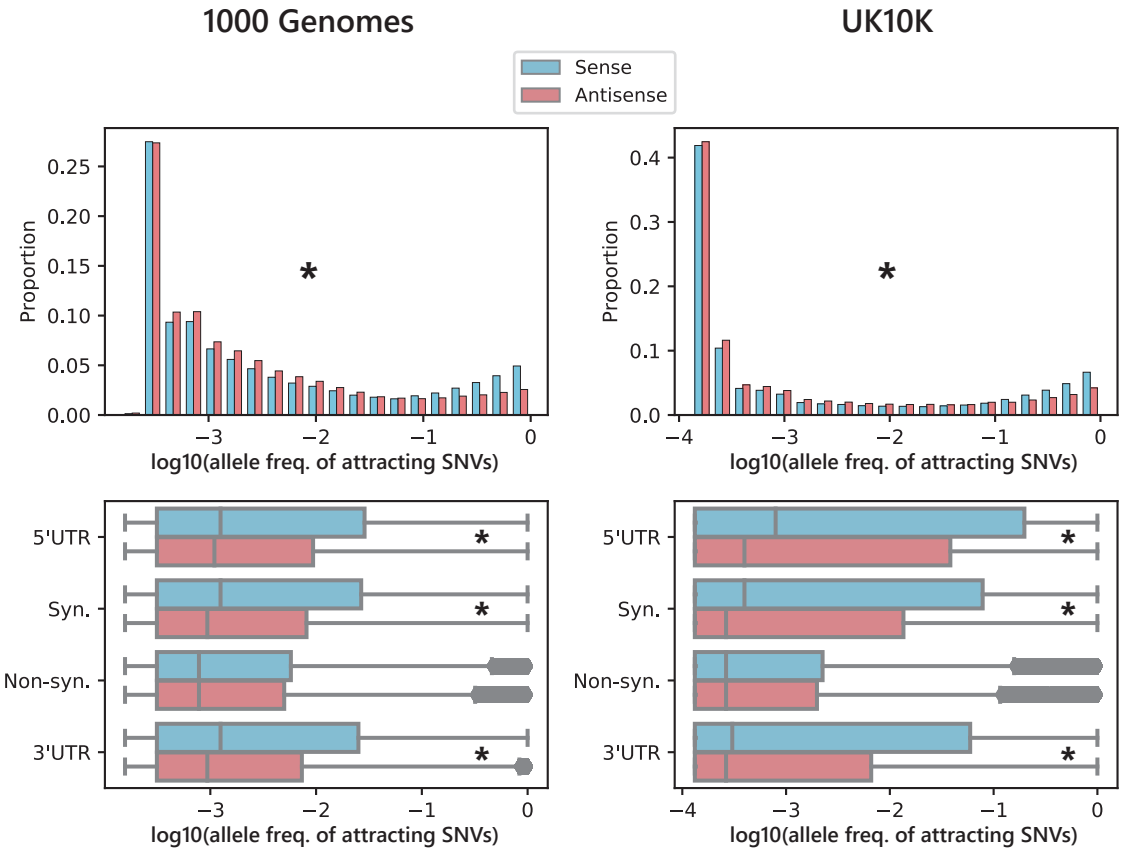

Figure S5 Results without restricting allele frequencies are shown here. Similar to Fig. 6, in the top panels, each histogram shows the overall distribution of allele frequencies for sense/antisense attracting SNVs, pooling all categories of variants together (at a log10 scale). In the bottom panels, each two boxplot shows the distribution of allele frequencies for sense/antisense attracting SNVs (at a log10 scale) separately for different categories of variants. These results are generally consistent with the results shown in Fig. 6 but less pronounced. In nonsynonymous variants, the allele frequencies under an antisense strand relationship do not show significant difference from those under a sense strand relationship.

Figure S6

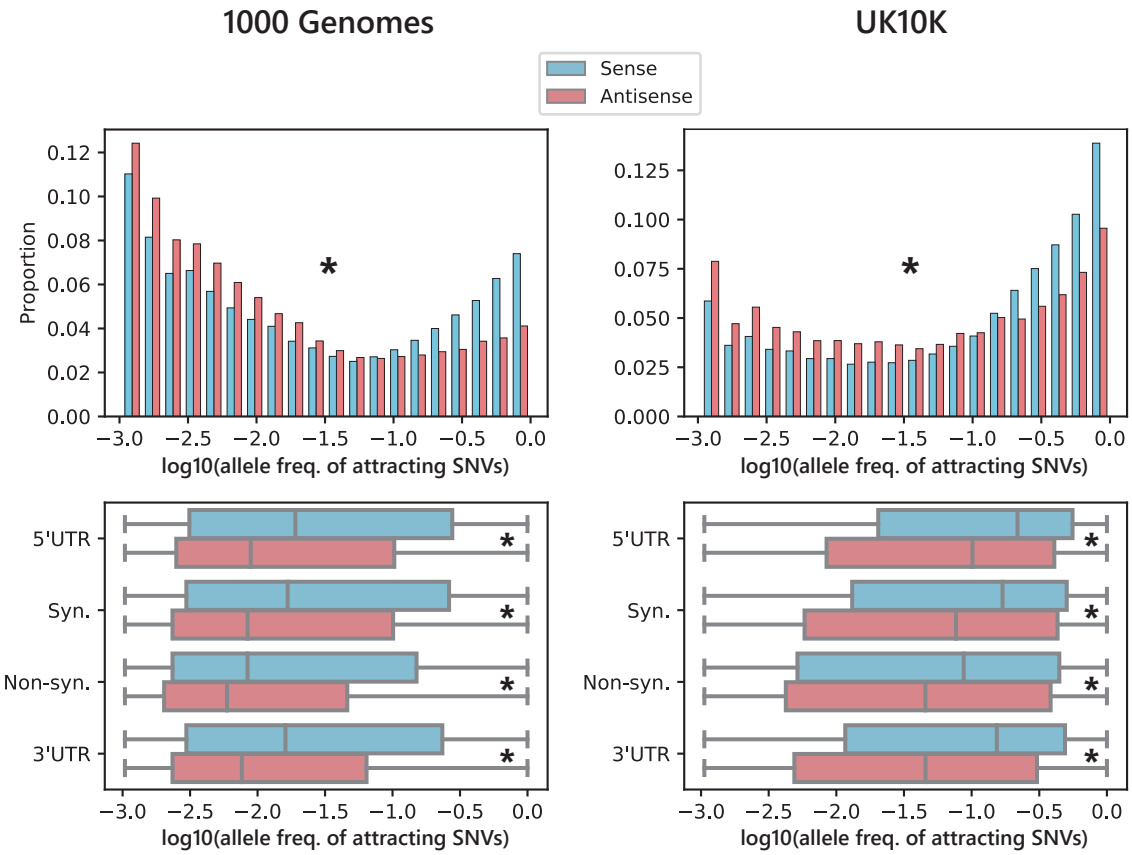

Figure S6 Results with using the rule described in Zhang et al., 2018 are shown here. Allele frequencies were restricted to ≥ 0.001. Similar to Fig. 6, in the top panels, each histogram shows the overall distribution of allele frequencies for sense/antisense attracting SNVs, pooling all categories of variants together (at a log10 scale). In the bottom panels, each two boxplot shows the distribution of allele frequencies for sense/antisense attracting SNVs (at a log10 scale) separately for different categories of variants. These results are very similar to the results shown in Fig. 6.