Supporting Result:
Circular RNAs Are Depleted of Polymorphisms at MicroRNA Binding Sites

Laurent F. Thomas and Pål Sætrom

Supplementary Methods

Nucleotide frequencies of reference alleles

We computed the probability distribution of A, C, T and G nucleotides to be reference alleles at SNPs in the 1000 genomes dataset.

TargetScan variables

Seed type and five other variables are used to compute TargetScan scores (version 6.0) [1]: AU content in flanks, 3' supplementary pairing, target abundance, seed-pairing stability, and site position in 3' UTR. TargetScan scores for site position were discarded here since circRNA are circular. The AU content variable is based on AU level in the transcript 30 bp upstream and downstream of the seed sites excluding the seed region. The 3' supplementary variable is based on Watson-Crick pairing at the 3' end of the miRNA. The target abundance variable is based on miRNA motif frequencies within 3'UTRs, and the seed-pairing stability score is based on thermodynamic parameters of the miRNA motif.

The target abundance and seed-pairing stability variables are based on the seed motif, the AU variable is based on flanks, and the 3' supplementary pairing is based on a small part of the upstream flank.

We ran TargetScan 6.0 to compute those scores at each site. For each of the four TargetScan variables, we computed the nine quantiles spliting the sites in ten groups of 10% (deciles). We computed the SNP distribution in each group by summing SNP counts at each position and dividing by the amount of sites, and smoothed the distribution with a 6-nucleotide sliding window.

AU and SNP density for microRNA motifs

The AU ratios of seed over flank for each miRNA motif in circRNAs were estimated by the average AU count at positions 1-8 of seed region averaged among all sites of a given motif in circRNAs, divided by the average AU density in the corresponding flanks. Similarly, the SNP density ratios of seed over flank for each miRNA motif were estimated by the average SNP count at positions 2-7 of the seed region averaged among all sites of a given motif, divided by the average SNP density in the corresponding flanks.

Total SNP densities of circular RNAs

Total SNP densities of circular RNAs were estimated by counting SNPs in circRNA exons divided by the circRNA length.

Supplementary Results

Using TargetScan, we computed scores for each of the miRNA target variables, grouped sites by score quantiles and computed the SNP density for each group, to see how each miRNA target variable affects SNP density distribution at potential miRNA binding sites (Supplementary Figures S1A-D).

Figure S1A showed that the low AU content in flanks (red) gave a higher SNP density in flanks (due to the higher probability of C and G being reference alleles) compared to high AU content (blue) and that all intermediate groups had intermediate SNP density levels in flanks. As expected, we did not see any particular difference of SNP density between the different AU scores at the seed region since the AU scores are based on the flanks, but we still saw a decrease of SNP density at the seed site compared to the flanks. However for high AU content (blue) in the flanks, the SNP depletion at the seed compared to the flank was only mild, as many seed sites in that group most likely have a lower AU content than their flanks.

Similar to the AU variable, we grouped sites by the TargetScan variable for 3' supplementary pairing (Figure S1B). We did not see any difference between the groups, indicating that this variable does not affect the SNP density at the seed site. The absence of noticeable depletion at sites with strong 3' supplementary pairing is possibly due to aligning sites by miRNA 5' end instead of 3' end.

Similarly, we grouped sites by the TargetScan variable for target abundance (Figure S1C), which is based on frequency of 8mers, 7mer-A1 and 7mer-m8 in 3'UTRs of reference mRNAs. This 3'UTR frequency correlates with motif frequency in circRNAs (spearman rho: 0.931, n=2043). This variable is meant to give priority to rare miRNA complementary seed motifs, as they require lower expressed miRNAs to be targeted, and is useful given miRNA and target expression profiles, but may therefore not really apply for selective pressure analysis based on SNPs. Figure S1C showed that the most common motifs (red) had the biggest SNP depletion at seed sites and that the rarest motifs (blue) had an increase at the seed site. This is because rare motifs have a lot of C and G both in their motifs and in their complementarity sequences. Specifically, target abundance raw scores correlated with counts of A and U in miRNA motifs (Pearson's coefficient $r = 0.299$, $p < 2.2 * 10^{-16}$, $n = 2043$) indicating that rare motifs have less A and U, and therefore more SNPs.

Similar to the target abundance variable, the seed-pairing stability (SPS) variable is also based on the seed motif and on its level of C and G. The SPS raw score correlated with counts of A and U in miRNA motifs (Pearson's coefficient r=0.982, $p < 2.2 * 10^{-16}$, n=2043 for 8mer and 7merM8, and r=0.908, $p < 2.2 * 10^{-16}$, n=2043 for 7merA1). Figure S1D showed a decrease of SNP density at seed sites except for motifs with many C and G (high seed stability) as they have a higher probability of overlapping SNPs due to the reference allele bias. Interestingly, flanks also had a slightly lower SNP density for low seed stability sites (red) compared to high stability sites (blue), possibly because the GC level of flanks partly depends on the GC level of seeds, *i.e.* the AU variable (AU level in flanks) is correlated with the SPS variable (GC level in seeds): Pearson's coefficient $r = -0.881$, $p < 2.2 * 10^{-16}$, n=2043).

The miRNA family subset gave similar results (data not shown).

To investigate why the rare and high seed stability miRNA motifs (*i.e.* motifs with higher CG content) have such an increase in SNP density at the seed regions compared to flanks, we computed SNP density ratios and AU ratios of seed over flank for each miRNA motif (Figure S2). We found that the increase in SNP density at rare and stable seeds is due to CpG dinucleotides in the miRNA motifs (and therefore in the seed region, as CpG also has CpG reverse complement) or to hexamers of Gs in the miRNA motif (*i.e.* hexamers of Cs in the complementary site). CpGs have highly increased occurrences of SNPs, because cytosines at CpGs are frequently methylated and spontaneous deamination of methylated cytosines creates mutagenic T:G mismatches in DNA; unless repaired, these T:G mismatches result in C to T changes after DNA replication.

Since SNP density can vary through the genome, we computed the SNP density in the 1906 autosomal circRNAs, and found an average of 1.45 SNP/hbp, including 108 circRNAs without SNPs (5.7%), and 5 circRNAs located in highly polymorphic loci, such as HLA and immunoglobulin loci (Figure S3).

If miRNA targeting of circRNAs follow similar principles as miRNA targeting of 3'UTRs [1], a circRNA with multiple seed sites is on average more likely to function as a miRNA sponge than a circRNA with one or few seed sites. We therefore identified the best seed of each circRNA as the miRNA motif(s) with the maximum number of sites in the circRNA and computed the best seed frequency as motif count by 100 bp (Figure S4A). Of the 1951 circRNAs, 4 circRNAs had a best seed density greater than 4 motifs per 100 bp, and 278 had their best seed occurring only once (14.2%). More generally, circRNAs with low best seed density may not act as miRNA sponges for a specific miRNA family, but maybe as general miRNA sponges. Reasoning that miRNA seed sites that occur multiple times in a circRNA are more likely to be functional than seed sites that occur once, we filtered out the best seed motif for each circRNA and computed their SNP density at seed complementary sites and flanks (Figure S4B). The SNP density ratio for the best seeds over their flanks was 0.76, which is a stronger depletion than without the best seed filtering and indicates that these best seeds are under stronger selective pressure than are less frequently occurring seed sites.

References

[1] Andrew Grimson, Kyle Kai-How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol. Cell*, 27(1):91–105, JUL 6 2007.

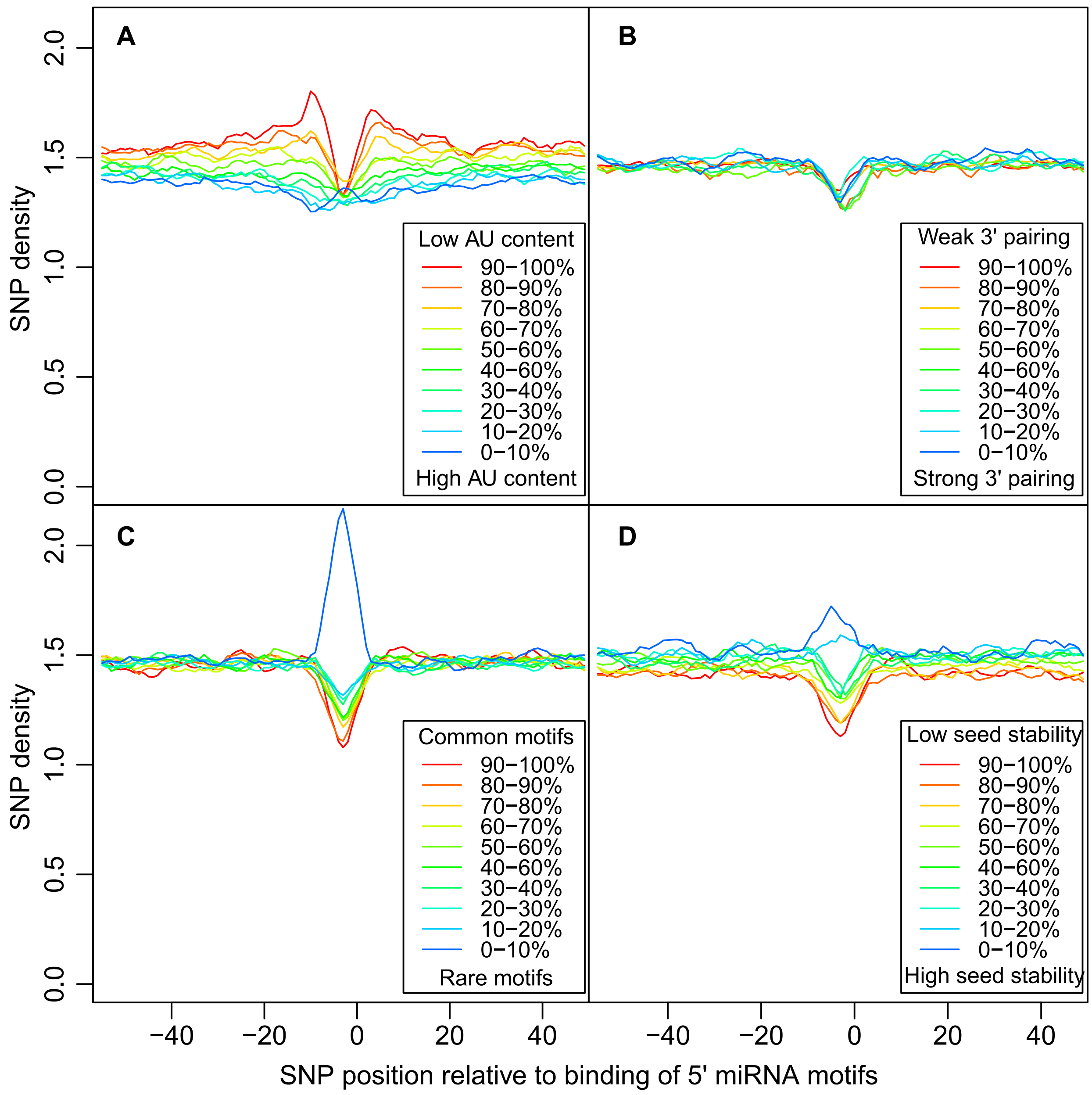

Figure S1: SNP density around miRNA complementary seed sites in circRNA transcripts for four TargetScan variables. Panels A-D show respectively SNP densities for sites grouped by TargetScan's scores for AU content, 3' supplementary pairing (3'SP), target abundance (TA), and seed pairing stability (SPS). The AU and SPS variables affect SNP density in flanks, the TA and SPS variables affect SNP density in seed, and the 3'SP variable does not affect SNP density.

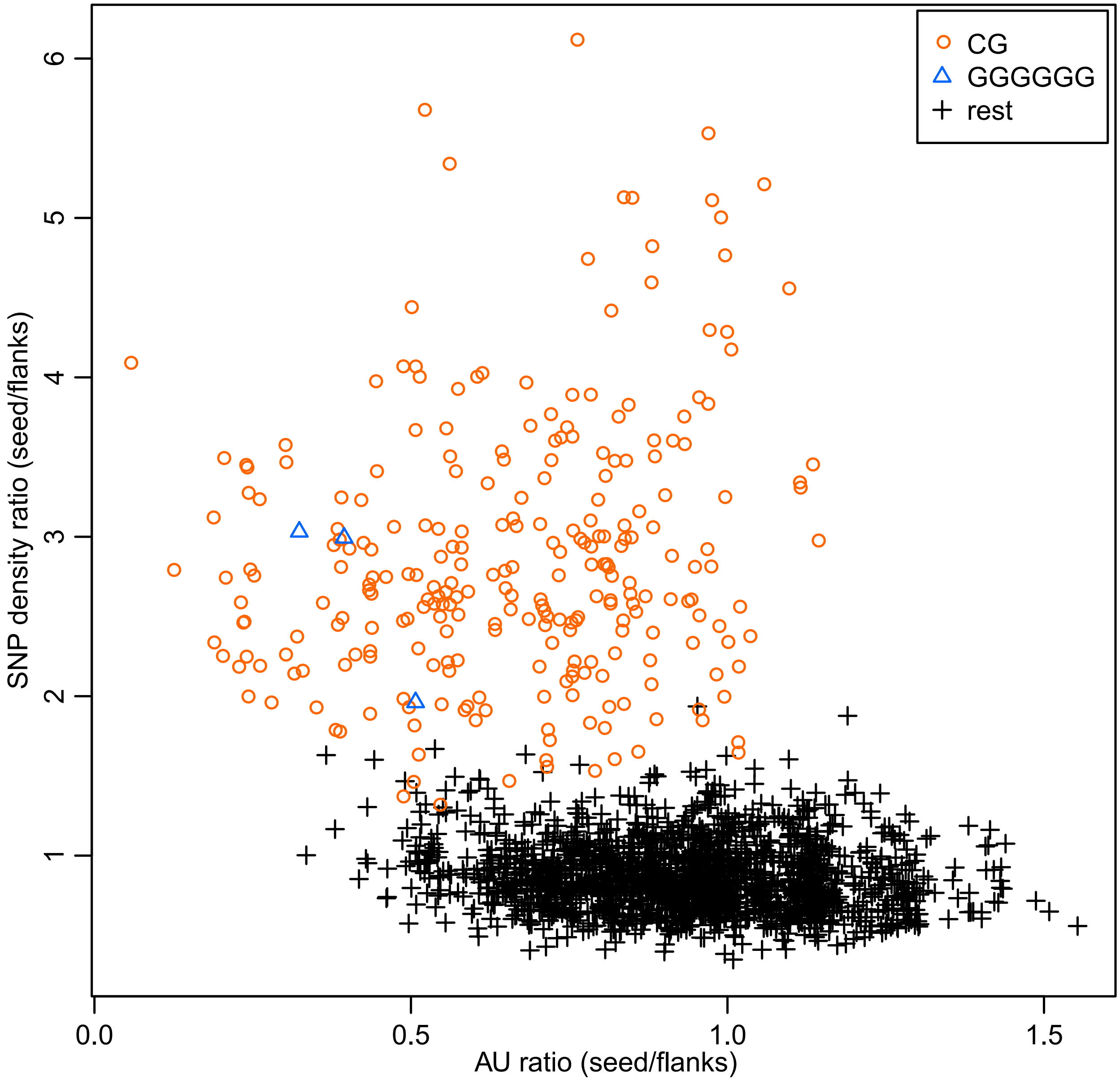

Figure S2: Increased SNP density in seed site for motifs with CpG dinucleotides. The x- and y-axes show respectively the AU ratios of seed over flanks and the SNP density ratios of seed over flanks averaged for miRNA motif (one point per miRNA motif). Orange circles show motifs with CpG dinucleotides within nucleotides [2-7], blue triangles show motifs with at least 6 consecutive G, and black crosses represent the other motifs. The orange and blue groups are separated from the rest as they have a higher SNP density level in their seed compared to their flanks. This increase is not particularly attributable to the AU ratio but to the CpG dinucleotides in motifs.

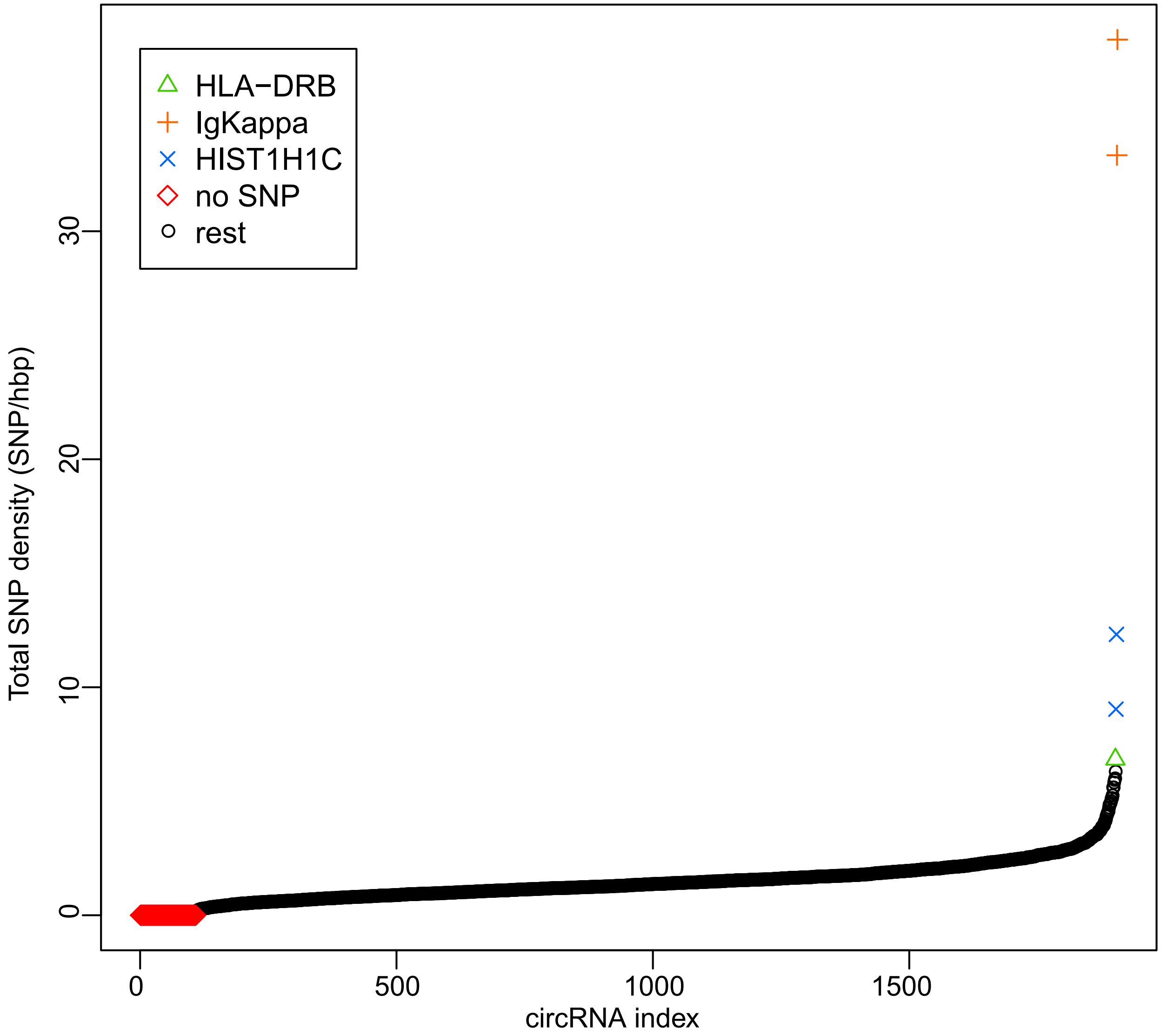

Figure S3: Total SNP density in each circRNA. The x-axis shows circRNAs ordered by SNP density and the y-axis shows the SNP density in circRNA in SNP count by hundred bp (SNP/hbp). CircRNAs without SNPs are depicted by a red diamond, and loci with extremely high SNP densities are denoted by a green triangle (HLA locus), an orange plus sign (Immunoglobulin Kappa locus) and a blue cross (Histone H1 (HIST1H1C) locus). Other circRNAs are depicted by a black circle.

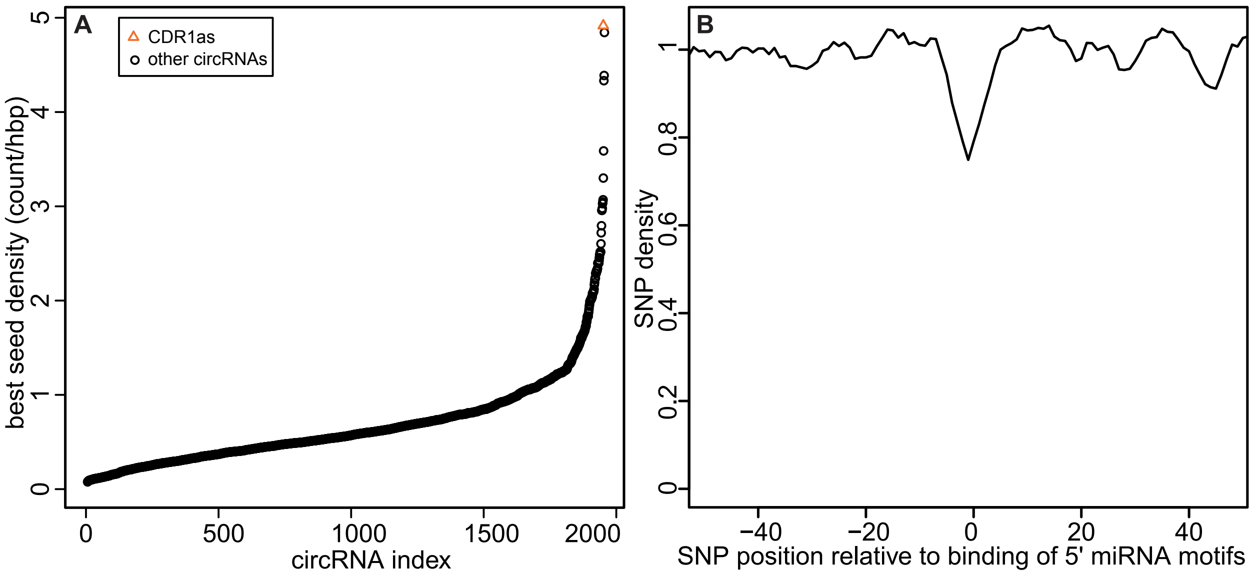

Figure S4: Best seed of circRNAs. Panel A shows the density of best miRNA seed in each circRNA. The x-axis shows circRNAs ordered by best seed density and the y-axis shows the best seed density in circRNA in seed count by hundred bp (seed/hbp). The orange triangle shows the best seed density of CDR1as, the only human circRNA validated as a miRNA sponge. The black circles represent the remaining circRNAs. Panel B shows SNP density around best seed complementary sites of each circRNA transcript. A 6-nucleotide window was used to smooth the distribution and the distribution was normalised by dividing by the median, so that flank density is one. The point at -1 shows the average SNP density at positions 2-7 of the best seed complementary sites compared to flanks, and has a SNP density ratio of 0.76.