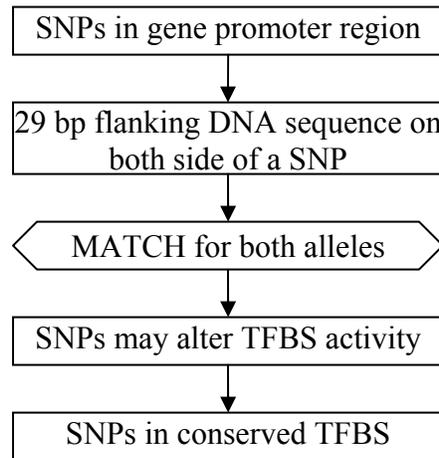
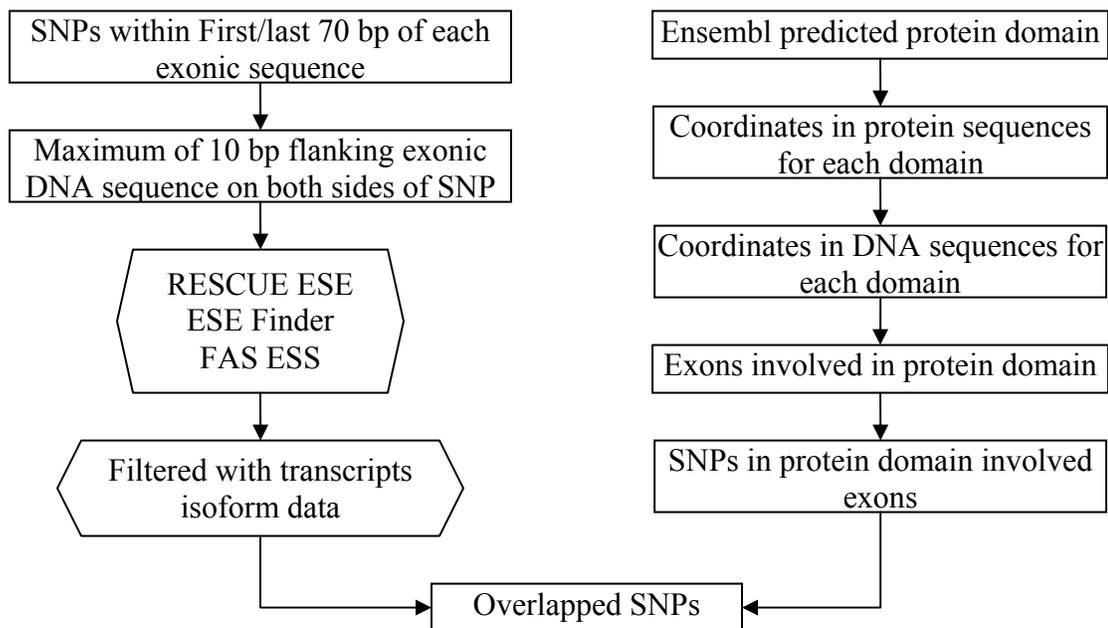


Supplementary Figure 1: Flow chart to identify SNPs that affect TFBS activity.



There are 656,662 SNPs within all gene promoter region (5kb upstream or 1kb downstream region of the transcription start site (TSS) of genes. Using the method described in text we found 64,147 of them occur in putative TFBS and having MSS or CSS difference of ≥ 0.2 between reference allele and alternative allele. Among the 64,147 SNPs, 17,476 SNPs occurs in conserved TFBSs. Of the 17,476 SNPs in conserved TFBSs, 5,534 have regulatory potential score of 0.1 or greater, and 9,454 SNPs have regulatory potential score of 0 or greater.

Supplementary Figure 2. Flow chart to identify SNPs that affect splicing activity.

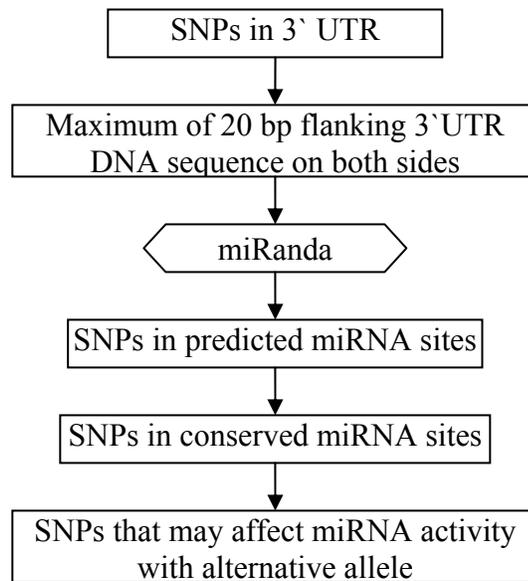


Based on the gene annotation in Ensembl database, we identified all SNPs within the first/last 70 base pair of all exons, or all SNPs within the exon if they were 140bp or shorter. We found 120,476 such SNPs in the genome. For each SNP, we extracted a maximum of 10 base pair of flanking exon DNA sequence to both sides of the SNP. For each allele of a SNP, we used RESCUE ESE and ESE Finder methods to predict possible ESE binding sites on the, at most, 21 base pair DNA sequence. ESS binding sites were predicted with FAS ESS method. For the RESCUE ESE method, we checked for perfect matches with the 238 hexamers for humans provided by the web server. For the ESE Finder method, we used all 5 position weight matrices and the default threshold provided by the ESE Finder web site (1.956, 1.867, 2.383, 2.67, 2.676 for matrices SF2/ASF, SF2/ASF, SC35, SRp40, SRp55, respectively). For the FAS-ESS method, we checked for perfect matches with the 103 hexanucleotides in “FAS-hex3” set provided by the web server. We classified a SNP as affecting splice activity if there was at least one predicted binding sites with one allele, but no predicted binding site with the alternative allele. We found a total of 28,452 such SNPs (13,420 SNPs with RESCUE ESE method, 7,769 SNPs with ESE Finder, 10,003 SNPs with FAS-ESS method). We eliminated SNPs where, based on Ensembl transcript isoform data, there was no evidence of alternative splicing. 7,267 SNPs remained.

Similar to Yuan et al.(2006 Nucleic Acids Research), we evaluated the consequence of a missing exon using Ensembl predicted protein domain information. Protein domains were predicted by a structure-based method, Pfam (Finn et. al 2006 Nucleic Acids Research) or a sequence-based method SCOP (Murzin et. al 1995 Journal of Molecular Biology,). First we extracted all predicted protein domains and found their coordinates on the protein sequence. We then mapped them to the associated transcript sequence, and then to the genomic DNA sequence in order to identify exons that contained coding information for protein domains. There were 80,433 SNPs in predicted protein domains

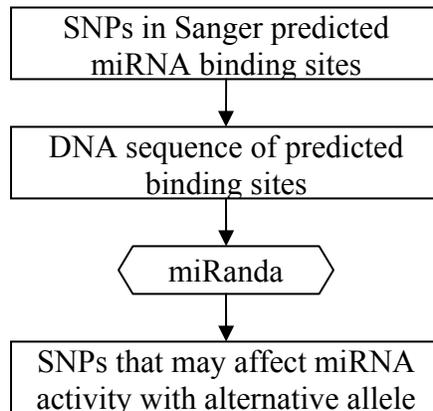
and 265,795 SNPs in these protein domain involved exons. Among the SNPs in these exons, 6,100 SNPs were in predicted ESE or ESS binding sites.

Supplementary Figure 3: Flow chart to identify SNPs that affect miRNA binding site activity.



There are 86,252 SNPs in human gene 3`UTR regions. Based on the method described in text and default miRanda parameter values (score cutoff $S \geq 140$; energy cutoff $E \leq -7.0$; gap opening: -9.0; gap extension -4.0; 5` scaling: 4), we identify 83,970 SNPs are in at least one of the predicted miRNA binding sites. Among them, 22,986 SNPs were in at least one conserved miRNA binding sites. For each of these SNPs, we compared the miRanda score between the reference allele and alternative allele and found 7,973 SNPs with score difference greater or equal to 16 between alleles.

Supplementary Figure 4: Flow chart to identify SNPs that may affect miRNA binding site activity in Sanger predicted miRNA binding sites.



miRBase Targets (<http://microrna.sanger.ac.uk/targets/v5/>) is one of the popular web resources for computationally predicted miRNA target information. This Sanger Institute web server also uses the miRanda method to predict potential target sites, but determines conserved target sites based on cross-species orthologous UTR alignments instead of whole genome alignment. There are total of 12,062 SNPs in Sanger predicted human miRNA binding sites. We extracted the human DNA sequence for all the predicted binding sites and calculated miRanda score with each of the possible alleles of a SNP using the same set of miRanda parameter values (score cutoff $S \geq 140$; energy cutoff $E \leq 7.0$; gap opening: -9.0; gap extension -4.0; 5` scaling: 4). We found 4,296 out of the 12,062 SNPs with miRanda score difference greater or equal to 16 between alleles.