GTEx small RNA processing to generate read count data.

Read processing and filtering

Prior to alignment, the four-base unique molecular identifiers (UMIs) and adapter sequences were removed from the raw sequencing reads using the UMItools and cutadapt tools, respectively. Only reads longer than 15bp after UMI and adapter removal were retained for further analysis. To remove contaminant sequences and the synthetic RNA spike-ins, reads were aligned to the UniVec database using bowtie2, filtering out any read that aligned to any of the sequences. Next, reads were aligned to canonical rRNA sequences using bowtie2 and removed from further analysis to prevent these small RNAs from overcrowding and erroneously multi-mapping during the subsequent alignment.

Generation of small RNA personal transcriptomes

The reference small RNA transcriptome used was derived from the RNAcentral v19 database, subsetting for transcripts annotated as one of the small RNA species and hence excluding transcripts annotated as longer RNA species, such as lncRNA and mRNA, and excluding miRNAs that were only annotated in GeneCards or ENA databases. To prevent multi-mapping of reads between precursor miRNA and mature miRNA sequences, transcripts annotated as precursor miRNA sequences were segregated into their own reference transcriptome file. Per donor, their annotated SNPs were intersected with the coordinates of the small RNA transcriptome. For each SNP and haplotype, a new version of the small RNA transcript was edited by replacing the reference with the alternate allele. This resulted in a personalized small RNA and precursor RNA transcriptome per donor, which served as the basis for the subsequent alignment steps.

Alignment and count generation of small RNAs

The filtered reads per sample were aligned to the personalized small RNA transcriptome using bowtie2, retaining all alignments (‘-a’ parameter), clipping a base from either of the read (‘-5 1 -3 1’ parameters) to prevent incomplete adapter/UMI removal from influencing the alignment and forcing reads to only align to the forward strand (‘--norc’ parameter). Only reads with a perfect alignment, i.e. no mismatches with the personalized transcriptome sequences (‘NM:i:0’ entry in bam file), were used for the quantification of small RNA expression. For each small RNA transcript, the number of reads was used to quantify expression, correcting for multi-mapping reads by dividing their contribution by their mapping count. Note that the correction for multimapping reads results in non-integer count values for the expression for small RNAs. For the overall quantification of small RNA expression, read counts for wildtype and mutant versions of transcripts were collapsed. Reads that did not perfectly align were subsequently aligned to the personalized precursor miRNA transcriptome and read counts were tallied as described previously. Reads that did not align perfectly were aligned to the GRCh38 reference genome for further processing and identification of novel small RNA transcripts.

Exclusion of miRXplore contamination

To exclude the possibility of cross-contamination and bleed-over of the miRXplore signal to other samples, we identified 39 small RNA transcripts that were highly specific to the pure

miRXplore spike-in samples (average log fold-change over 5 compared to all other samples, and a Benjamini-Hochberg adjusted p-value smaller than 10^{-16}). The expression of these transcripts was then used to determine the rate of contamination across samples by summing up the normalized counts of these marker genes as transcripts per million (TPM) and dividing by the mean TPM of these small RNA transcripts in pure miRXplore spike-in samples. Most samples (88%) showed very limited contamination with miRNA transcripts present in the miRXplore spike-in (less than 0.01%), but some samples showed contamination in higher ranges (1-10%). To correct the contamination, we multiplied the estimated contamination with the mean total small RNA TPM vector from miRXplore spike-in samples and total coverage, before subtracting the result from individual samples, correcting small negative values to zeroes.