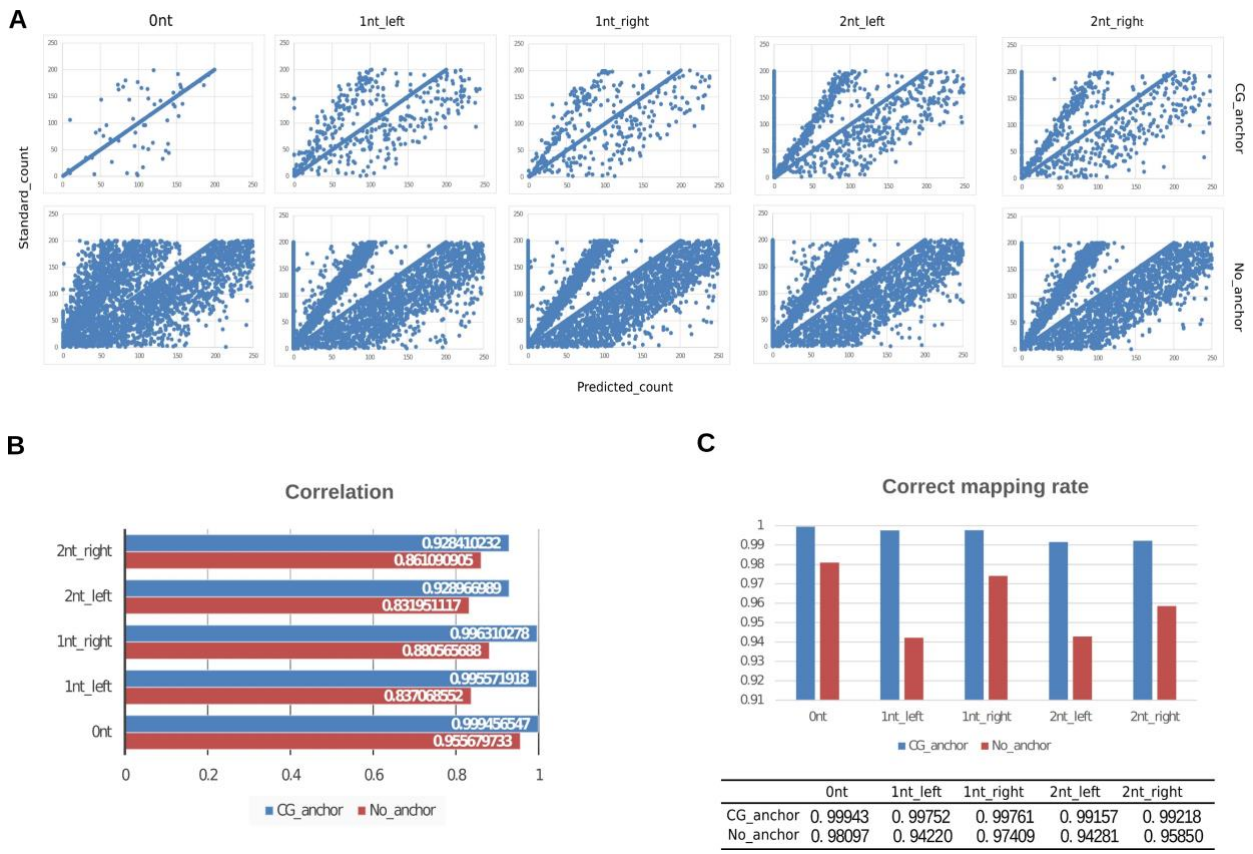
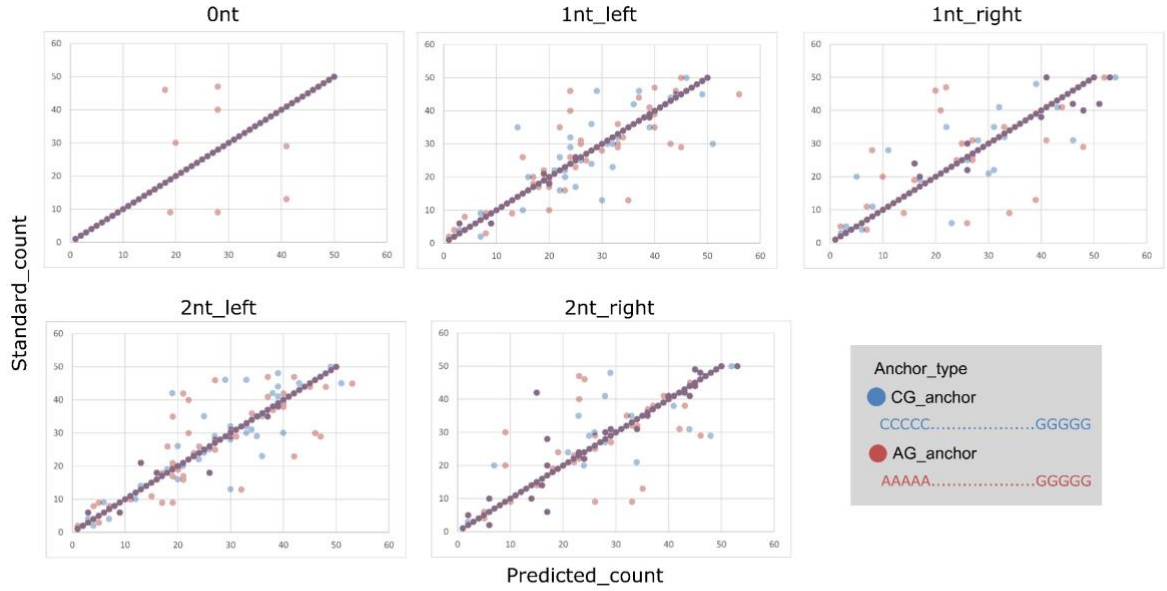


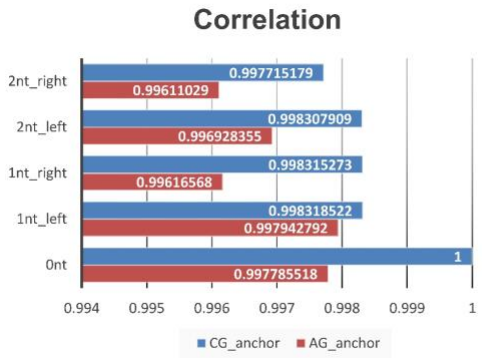
	0nt	1nt_left	1nt_right	2nt_left	2nt_right
CG_anchor	0. 99943	0. 99752	0. 99761	0. 99157	0. 99218
No_anchor	0. 98097	0. 94220	0. 97409	0. 94281	0. 95850

Figure S1. **Comparison of alignment accuracy between AASRA (CG_anchor) and the direct sncRNA-sncRNA alignment method (No_anchor) using sncRNA simulation data containing miRNAs, endo-siRNAs, piRNAs, snRNAs and tRNAs.** (**A**) Scatter plots showing alignment of the simulation sncRNA datasets with or without 1-2nt overhangs by AASRA (CG_anchor) and the direct sncRNA alignment method (No_anchor). (**B**) Bar graphs comparing the correlation coefficient (R^2) values between predicted counts (calculated by the algorithm) and standard counts (known for the simulation data) identified using AASRA (CG_anchor) and the direct sncRNA alignment method (No_anchor). (**C**) Bar graphs comparing the correct mapping rates of the simulation datasets between AASRA and the direct sncRNA alignment. The correct mapping rate is defined as the number of correctly mapped reads/ total reads.

A



B



C

Table S1. 5'-3' ending sequence combination pattern counts in the miRNA index file.

	3' A	3' C	3' G	3' T
5' A	666	784	533	946
5' C	537	749	498	797
5' G	947	987	583	1134
5' T	9050	12888	5885	8948

Figure S2. Comparison of alignment accuracy between the CG or AG anchor using miRNA simulation data containing mature and precursor miRNAs. (**A**) Scatter plots showing correlations between the predicted counts (calculated by the algorithm) and standard counts (known for the simulation data) derived from alignment using the CG anchor (blue dots) or the AG anchor (red dots). The CG anchor yielded better results than the AG anchor. (**B**) Bar graphs comparing the correlation coefficient values between the predicted counts (calculated by the algorithm) and standard counts (known for the simulation data) identified using the CG or AG anchor for alignment. (**C**) Frequency of the four nucleotides at both ends of miRNAs in the miRNA index file. miRNA sequences that start with cytosine and end with guanine are the least common and thus, the CG anchor is a better choice due to a lower frequency at both ends of sncRNAs.

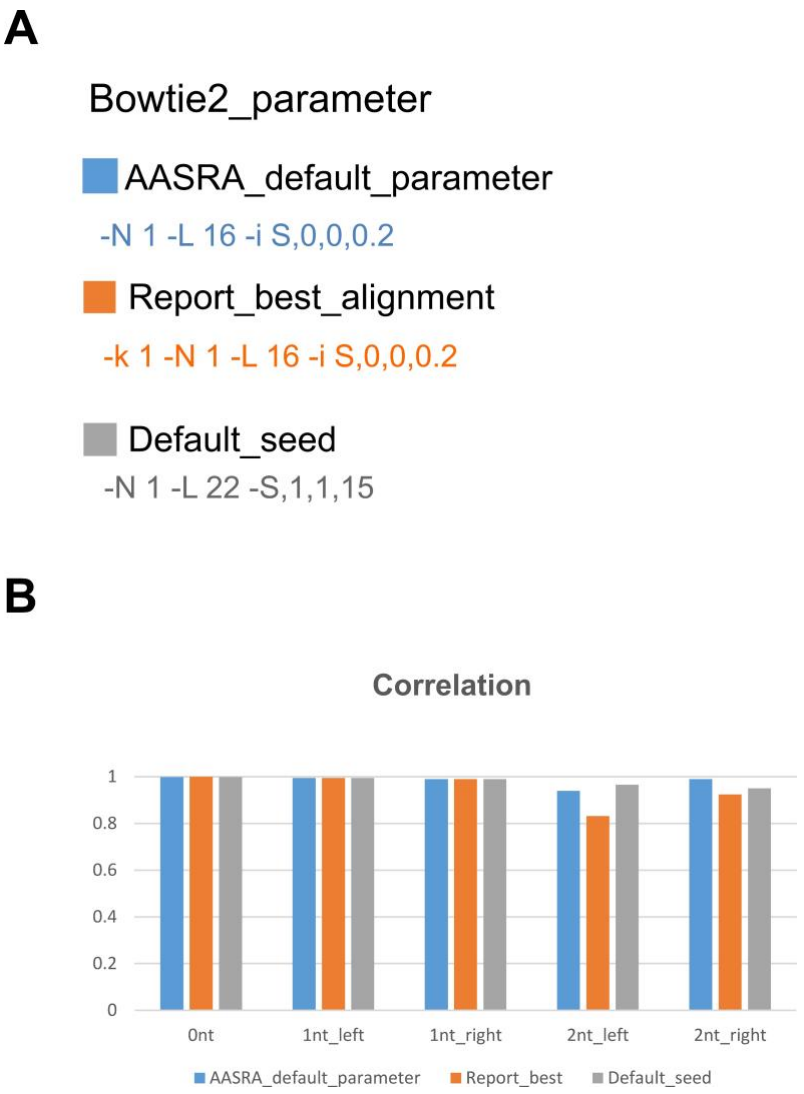


Figure S3. **Comparison of alignment accuracy affected by different Bowtie2 parameters using sncRNA simulation datasets containing miRNAs, endo-siRNAs, piRNAs, snRNAs and tRNAs with (0nt) or with 1-2nt overhangs.** (**A**) Bowtie2 commands with parameters used for comparison. AASRA default parameters were optimized based on those yielding the best-reported alignment results by Bowtie2 (Reprot_best_aligment). Bowtie default parameters (Default_seed) are tested as well. Seed length (-L) was optimized for miRNA alignment in AASRA. AASRA default allows 1 mismatch in seed sequence alignment (-N 1). (**B**) Bar graphs showing the correlation rates between predicted counts (by software calculation) and standard counts (simulation data) when 3 different Bowtie2 parameters, as illustrated in panel A, were used for alignment. Default AASRA parameter produces the best overall alignment accuracy for simulation datasets.

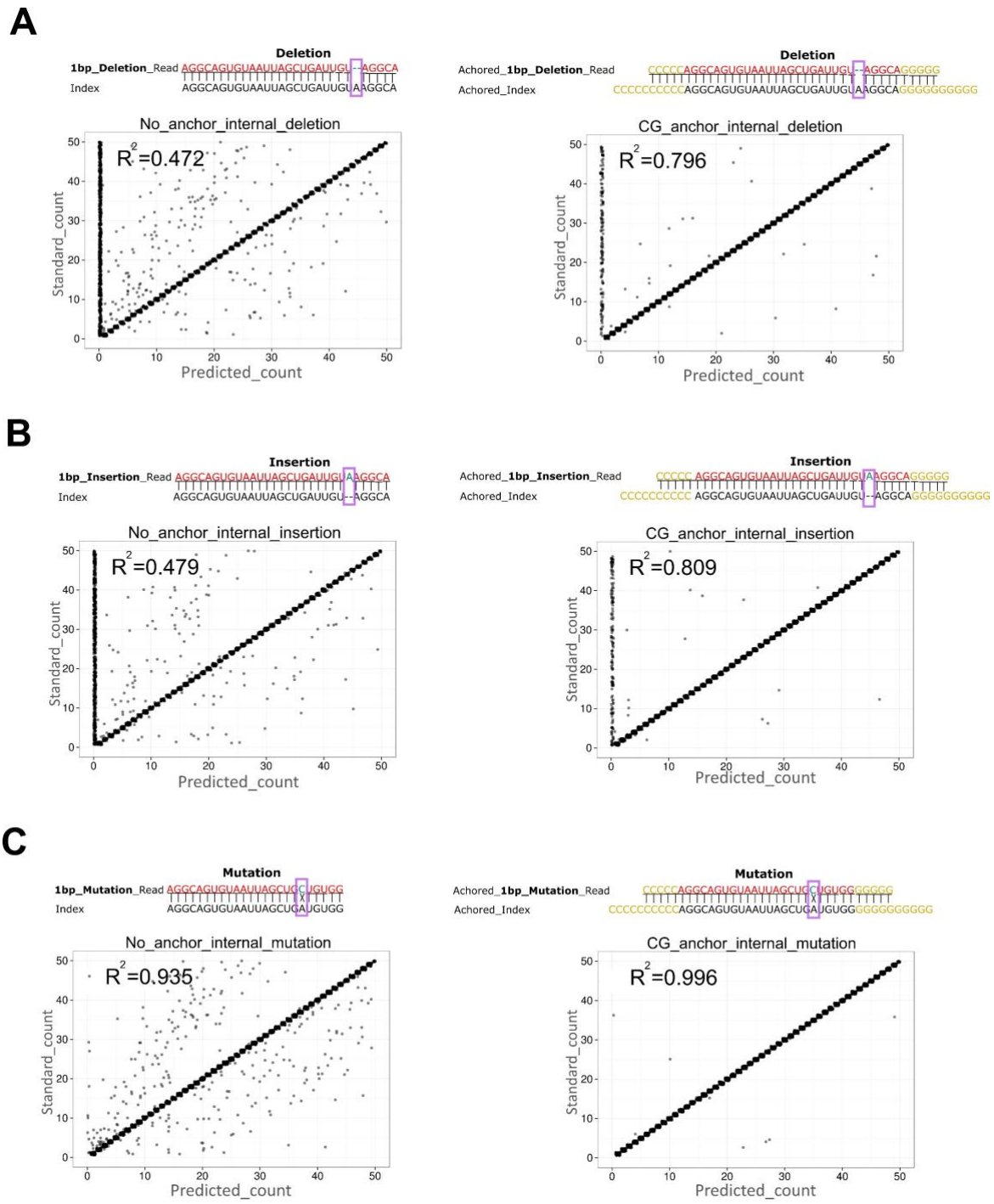


Figure S4. Comparison of alignment accuracy between AASRA (CG_anchor) and the direct sncRNA alignment method (No_anchor) using sncRNA simulation datasets containing miRNAs, endo-siRNAs, piRNAs, snRNAs and tRNAs with 1nt internal deletions **(A)**, insertions **(B)** or mutations **(C)**.

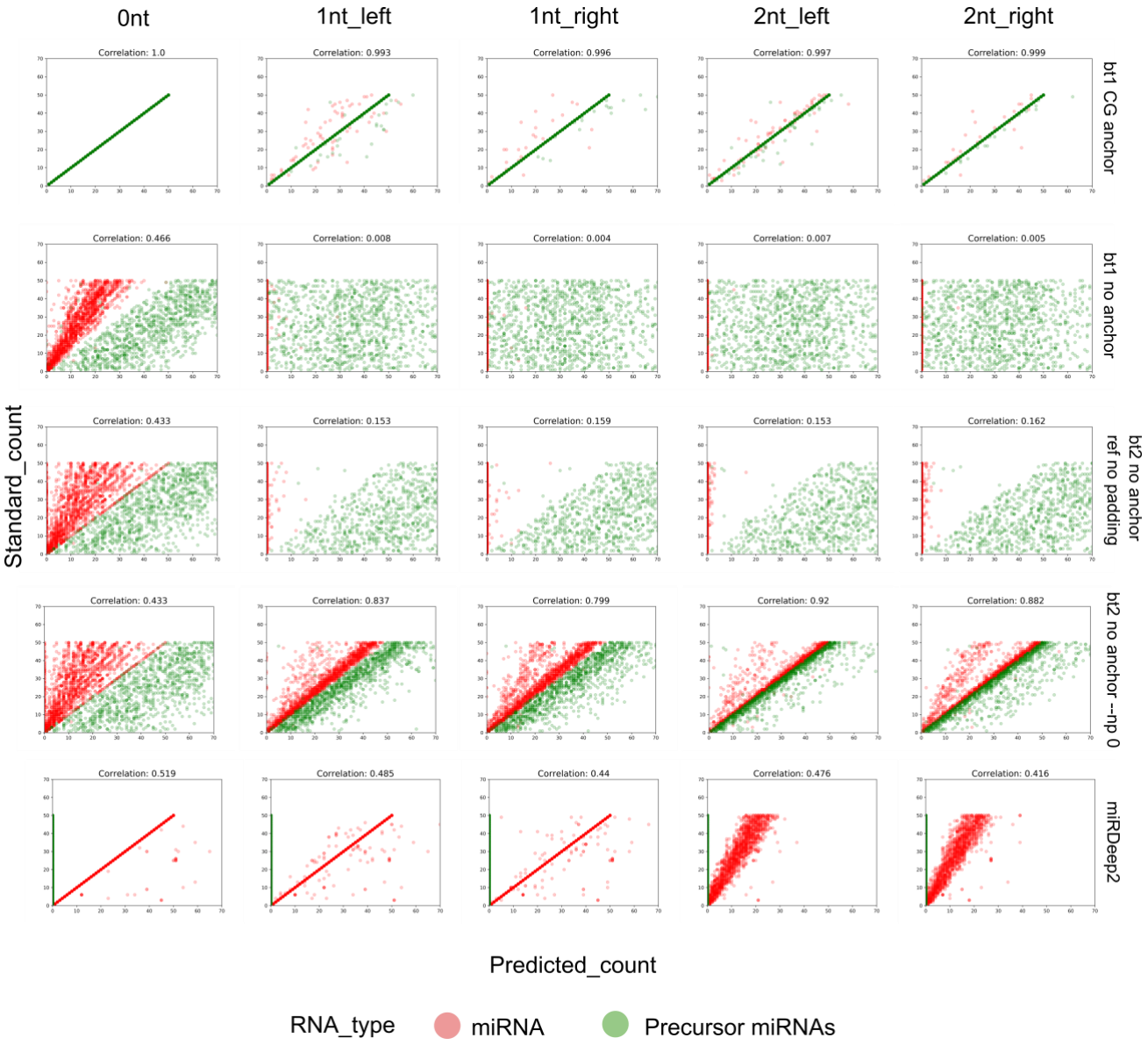


Figure S5. **Performance comparison between Bowtie1-based AASRA, Bowtie1, Bowtie2, Bowtie2 --np 0, and miRDeep2**. Simulation datasets containing both mature (red dots) and precursor (green dots) miRNA sequences with 0-2nt overhangs were aligned to the reference sncRNA dataset using Bowtie1 (-v 2) with CG anchor, Bowtie1 (-v 2), Bowtie2 default parameter (reference without padding sequence), Bowtie2 with --np 0 parameter (reference with 10nt N repeats padding sequence on 5' and 3') and miRDeep2 default parameter (quantifier.pl). Pearson product-moment correlation coefficients between the standard and software predicted counts are indicated above each figure. No sample anchoring was performed except for Bowtie1_CG_anchor group.

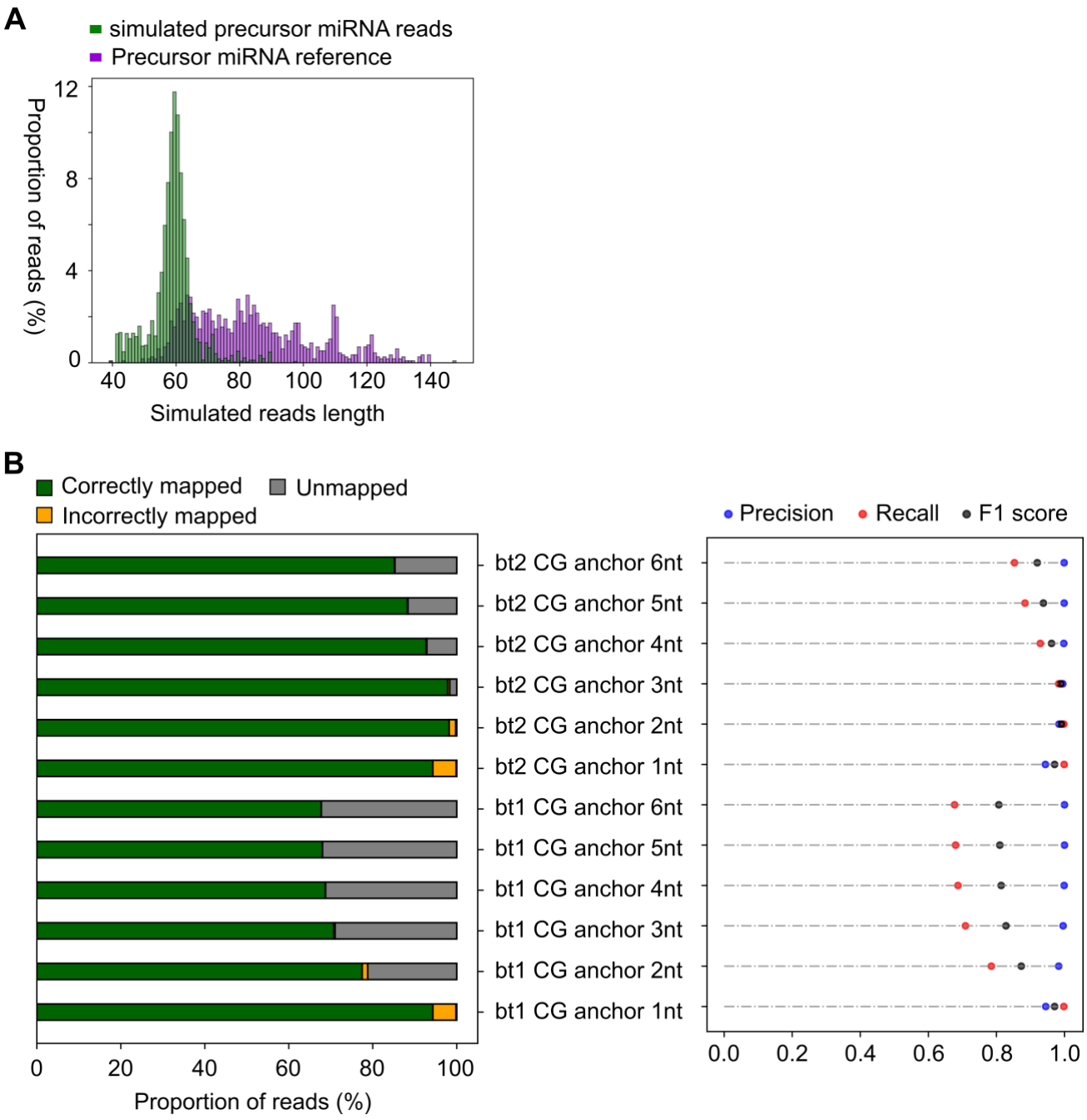


Figure S6. Performance comparison of AASRA among different anchor lengths with realistic simulated miRNA reads. (**A**) Length distribution of the precursor miRNA from the miRbase reference (purple) and realistic simulation dataset (green). (**B**) Assessment of different CG anchor length with Bowtie2 (bt2) or Bowtie1 (bt1) by percentage of correctly mapped reads (green), incorrectly mapped reads (yellow), and unmapped reads (grey) (left panel) and precision (correctly mapped/(correctly mapped + incorrectly mapped)), recall (correctly mapped/(correctly mapped + unmapped)) and F1 score (right panel).

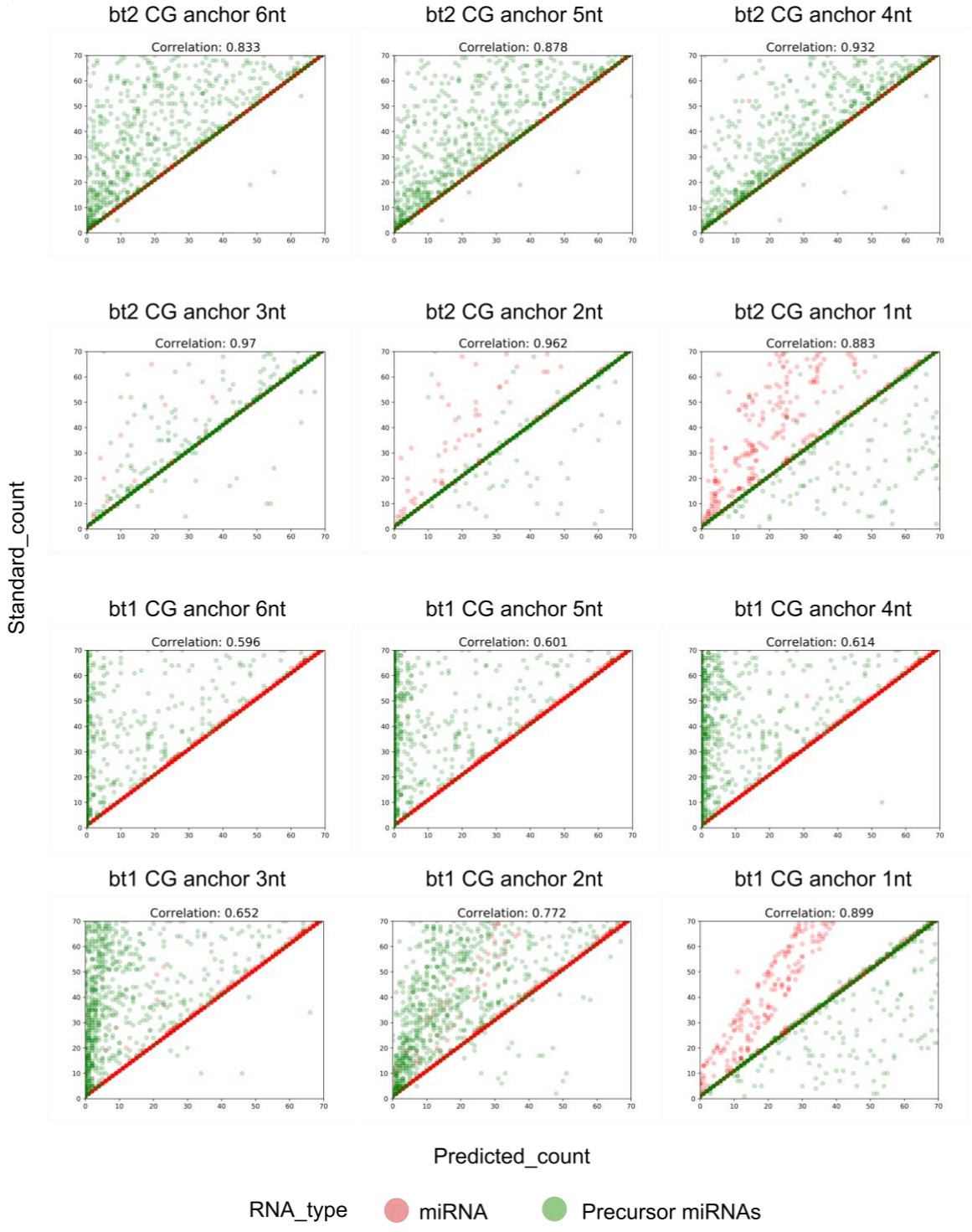


Figure S7. **Alignment accuracy assessment of AASRA among different anchor length with realistic simulated miRNA reads**. Assessment of different CG anchor length with Bowtie2 (bt2) or Bowtie1 (bt1) by pearson product-moment correlation coefficients between the standard counts and the annotation software predicted counts.

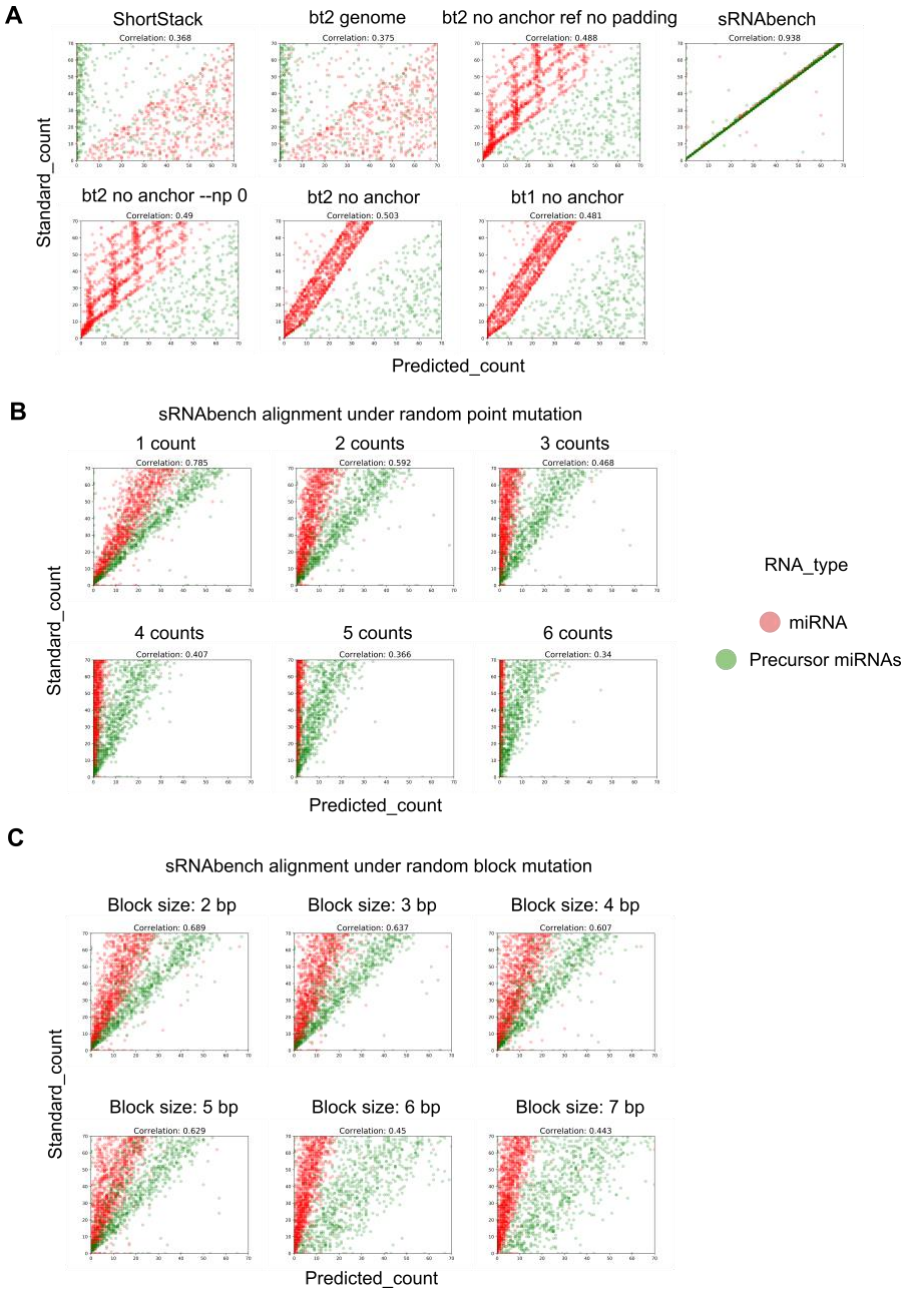


Figure S8. Alignment accuracy assessment with realistic simulated miRNA reads, point mutations and block mutations. (**A**) Simulated miRNA datasets were aligned to the reference miRNA dataset using Bowtie1 -v 2 (bt1 no anchor), Bowtie2 default parameter (bt2 no anchor), Bowtie2 default parameter without reference padding sequence (bt2 no anchor ref no padding), Bowtie2 --np 0 (bt2 no anchor –np 0) and sRNAbench. For ShortStack and Bowtie2 genome alignment (bt2 genome), simulated miRNA datasets were aligned to the mm9 genome using software default parameter. Pearson product-moment correlation coefficients between the standard and software predicted counts are indicated above each figure. **(B)** Point mutations, including insertion, deletion, mismatch, duplication and move, are randomly assigned to each read with different number of mutation (counts) per read. **(C)** Block mutations, including insertion, deletion, mismatch, duplication and move, are randomly assigned to each read with different length of block mutation per read. Pearson product-moment correlation coefficients of standard counts and the alignment software predicted counts are indicated.

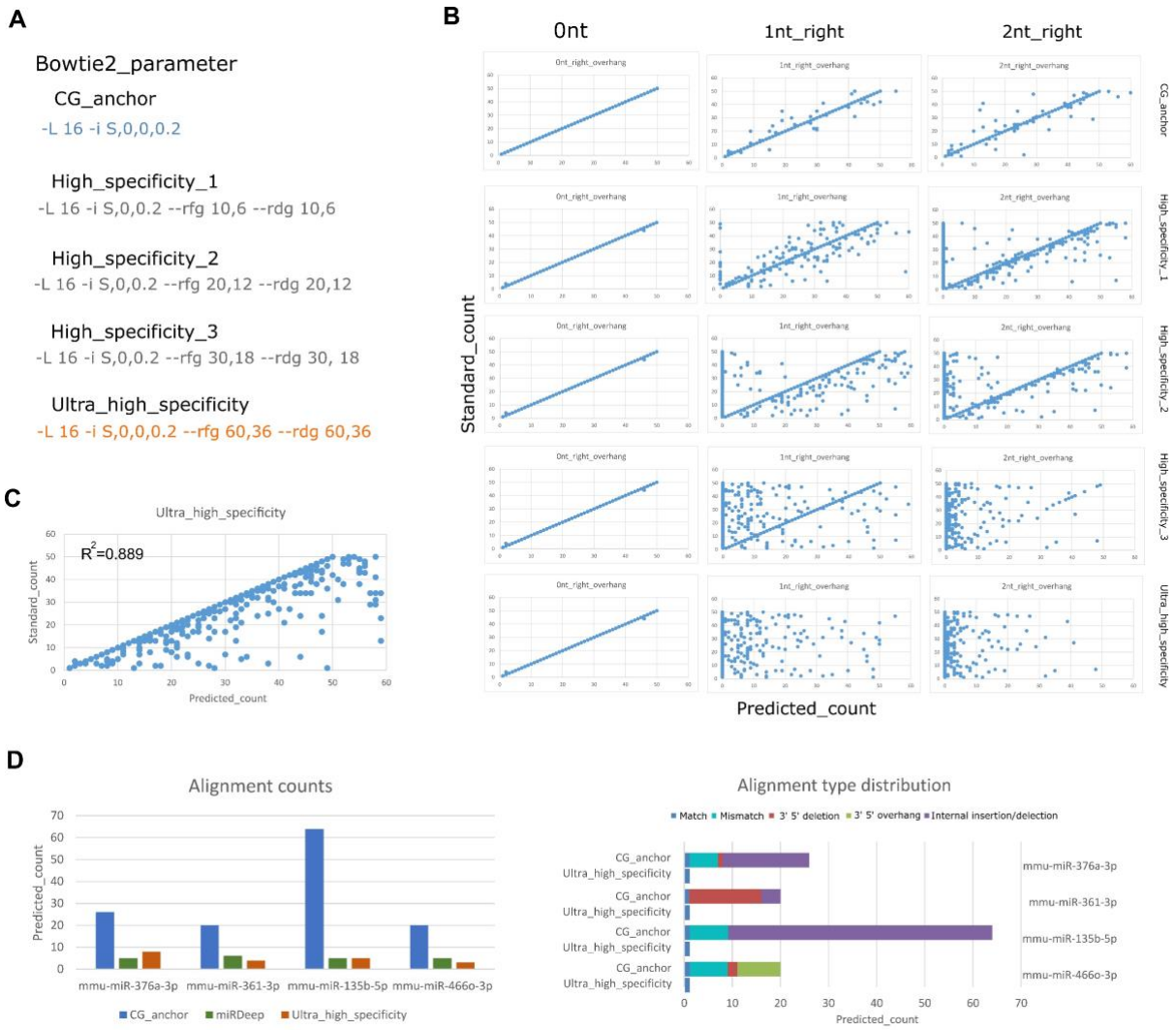


Figure S9. **The AASRA ultra-high specificity function**. (**A**) The parameters adjusted for various levels of alignment stringency. -L is the seed length of alignment; –i functions to govern the interval between two seed substrings used during multispeed alignment, controlling sensitivity and speed; --rfg --rdg controls the gap opening penalty. The highest stringency (ultra_high_specificity) setting can eliminate sequences containing 1-2nt overhangs. (**B**) Scatter plots showing alignment results at the four levels of specificity settings (High_specificty_1-3, and ultra_high_specificity) using sncRNA simulation datasets containing miRNAs, endo-siRNAs, piRNAs, snRNAs and tRNAs with (0nt) or without 1-2nt overhangs. (**C**) No effects of sequences with overhangs on the alignment results under the ultra-high specificity setting. Simulation datasets containing no (0nt) or 1-2nt overhangs were merged and used for mapping against the reference dataset. The ultra_high_specificity setting effectively eliminated the interference from the reads with overhang nucleotides (R^2=0.889). (**D**) Bar graphs showing the counts of sperm sncRNA reads aligned to four mature miRNAs using miRDeep and AASRA at default (CG_anchor) and ultra-high specificity settings (Ultra_high_specificity). Counts for different variant types of the four miRNAs identified using AASRA under the default (GC_anchor) and ultra-high specificity (Ultra_high_specificity) settings. The ultra_high_specificity setting effectively removed miRNA variants in the sncRNA-Seq data.