Supplemental Material for:

RASER: Reads Aligner for SNPs and Editing sites of RNA

Jaegyoon Ahn and Xinshu xiao

Supplemental Figures 1-8
Supplemental Table 1

Figure S1

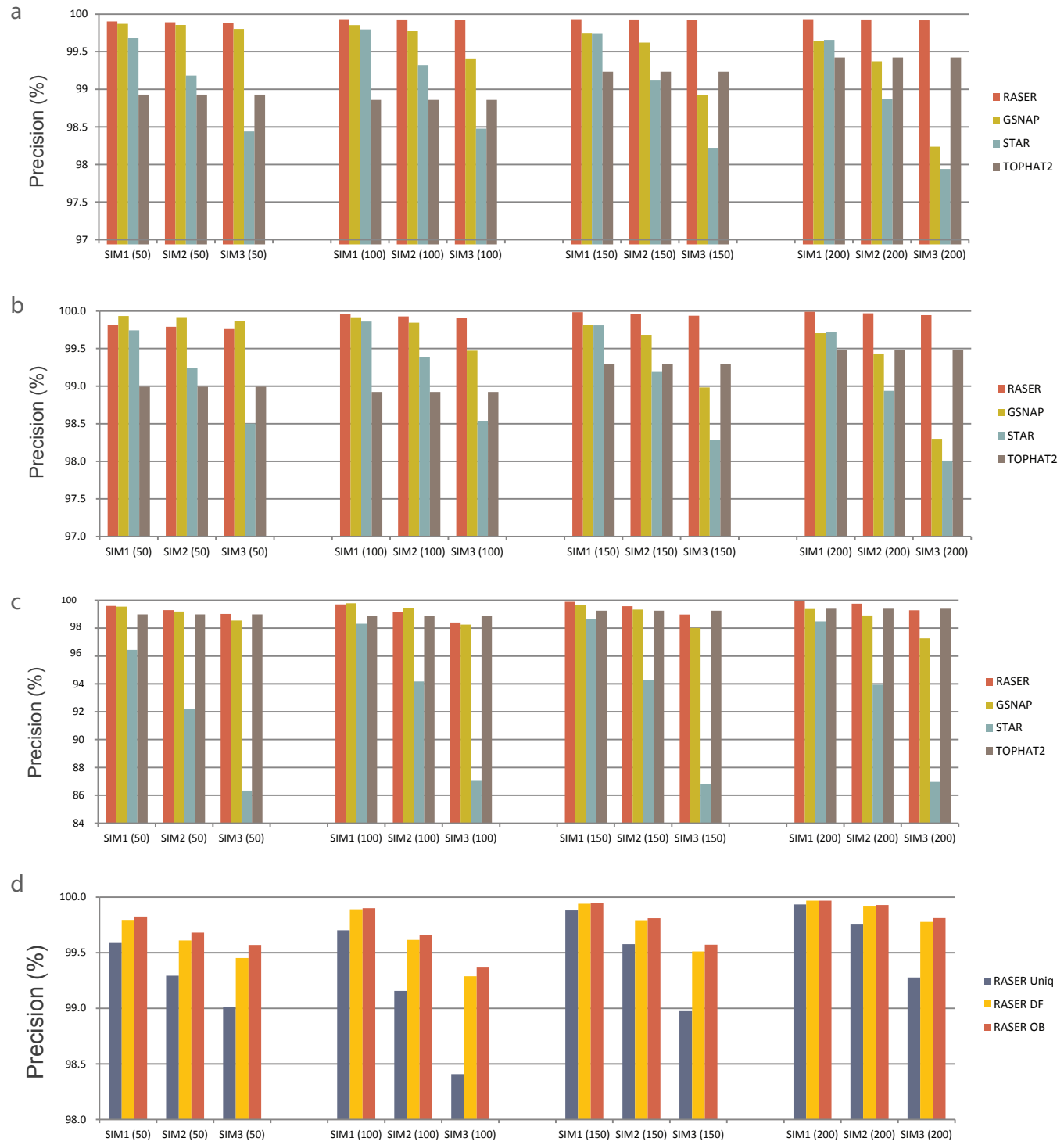


Figure S1. Comparison of mapping performance. (Unless otherwise noted, for all panels, "obviously best" mappings for RASER and unique mappings for other aligners were used.) (a) Comparison of precision (defined as percentage of correctly mapped reads among all mapped reads) using simulated data sets of varying levels of sequencing errors (SIM1 < SIM2 < SIM3) and read length (in parenthesis of the x axis label). The "obviously best" scheme was used for RASER. Unique mapping was required for results from the other aligners. The mapping was deemed correct if more than 50% of the nucleotides in a read were correctly mapped to their original genomic loci. (b) Similar as (a), but unique mapping by RASER is reported. (c) Similar as (b), but the mapping was deemed correct if more than 80% of the nucleotides in a read were correctly mapped to their original genomic loci. (d) Comparison of precision of RASER uniq (unique mappings only), RASER DF (double filtering scheme), and RASER OB (obviously best scheme). The mapping was deemed correct if more than 80% of the nucleotides in the reads were correctly mapped to its original genomic loci. (e) Comparison of recall of the 3 schemes of RASER as in (d). (f) Comparison of precision of NOVOALIGN and other aligners, excluding spliced junction reads from simulated reads (since NOVOALIGN does not map spliced reads). The mapping was deemed correct if more than 50% of the nucleotides in a read were correctly mapped to their original genomic loci. (g) Similar as (f), but the mapping was deemed correct if more than 80% of nucleotides in a read were correctly mapped. (h) Similar as (f), for comparison of recall of NOVOALIGN and other aligners. (i) Comparison of multiple mapping rate (defined as percentage of multiply mapped reads among all reads).

Figure S1

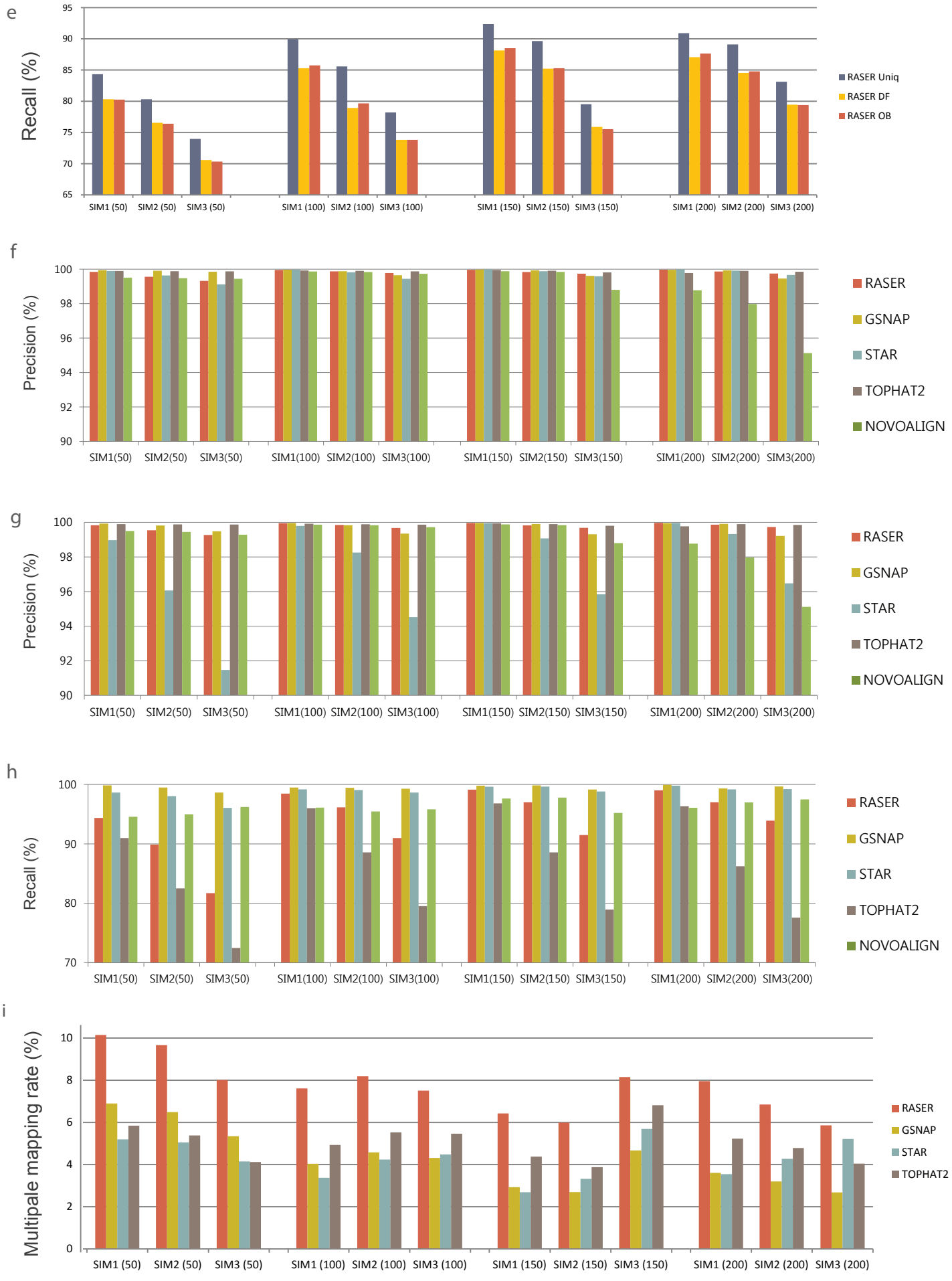


Figure S1. Comparison of mapping performance (continued)

Figure S2

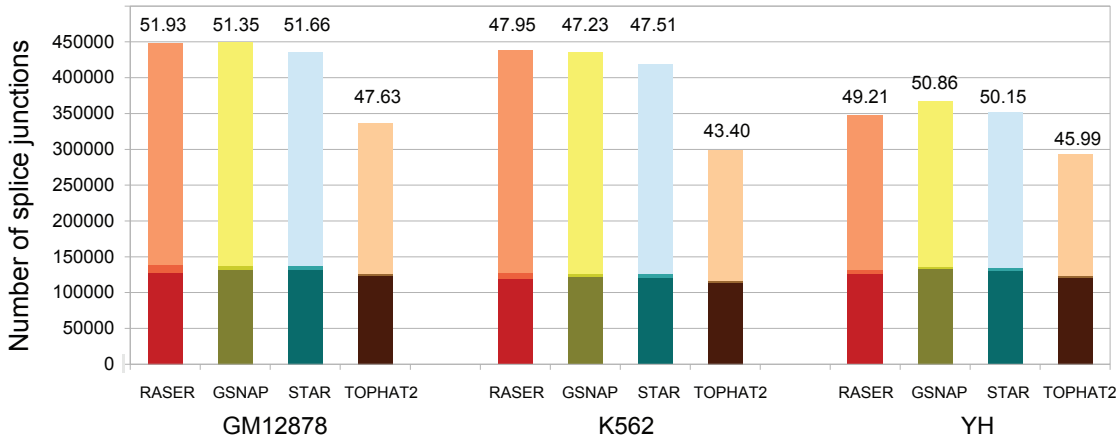


Figure S2. Identification of spliced junctions. Numbers of spliced junctions identified in real RNA-Seq data sets are shown. Results of each aligner are illustrated with one bar for each dataset, where the bottom (darkest color), middle (medium darkness) and top (lightest color) portions correspond to the number of perfectly matched junctions, partially matched junctions and novel junctions, respectively. Perfectly matched junctions were defined as those RNA-Seq-identified junctions that are exactly the same as Gencode v19 annotation. Partially matched junctions were those that differed from Gencode v19 annotation by less than 5 bases. Novel junctions were those that are not included in the Gencode v19 annotation. To define a junction based on RNA-Seq for all aligners, we require its read coverage being at least 3 and preference was given to canonical GT-AG or GC-AG splice sites. The numbers (%) above the bars represent the recall rate, which is defined as (number of perfectly or partially matched junctions / total number of junctions in Gencode v19 annotation).

Figure S3

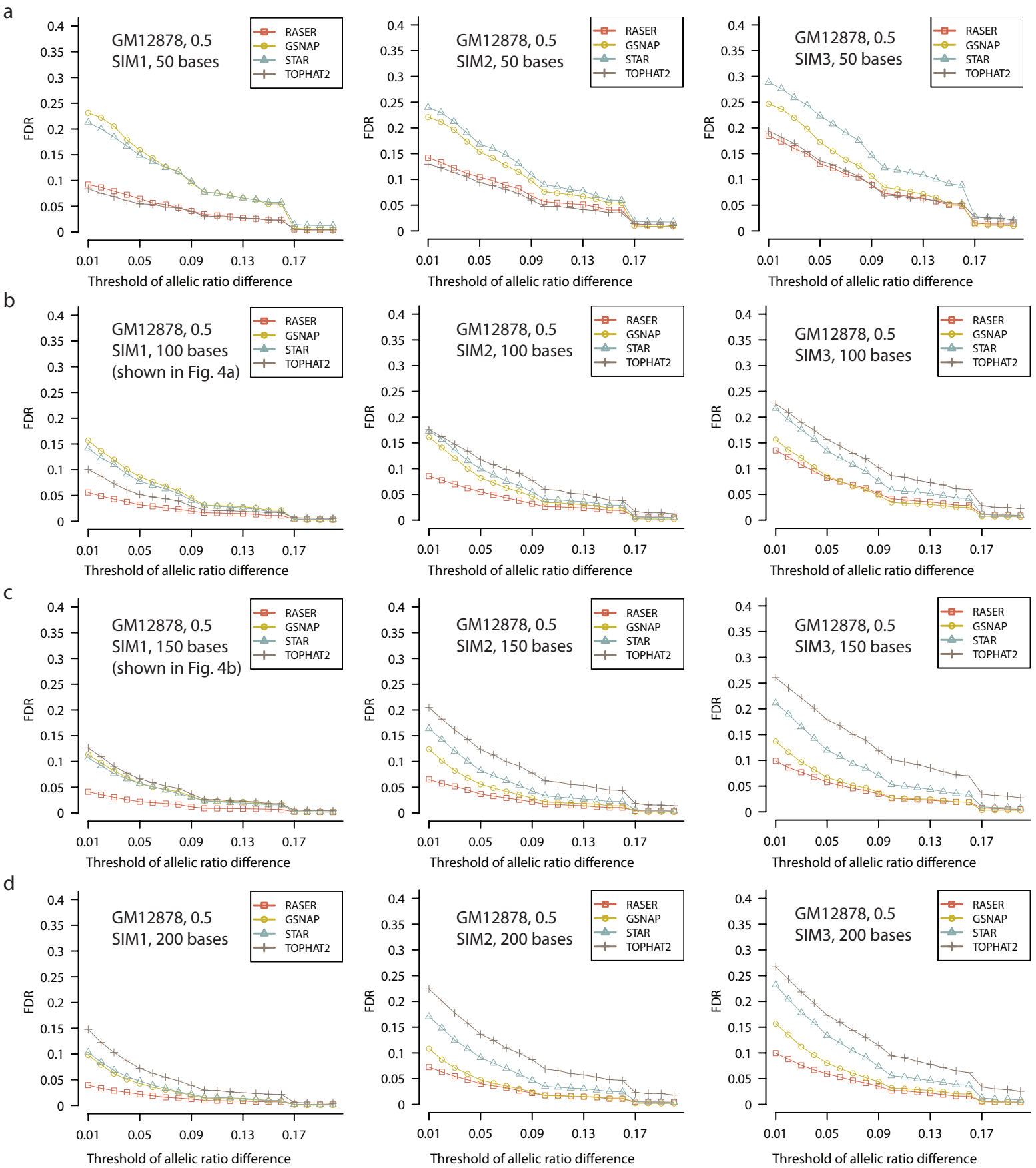


Figure S3. False discovery rate in the quantification of allelic ratios of SNPs expressed in RNA-Seq. SNPs in the GM12878 and YH samples which had corresponding whole genome sequencing data were implanted into the simulation data of 50, 100, 150 and 200 bases in length. SIM1, SIM2 and SIM3 data were used for the plots in the first, middle and last column, respectively. For each read that contains a known SNP, the probability for the SNP to have the reference or alternative allele was set to be 0.5, 0.75 and 0.25. This allelic ratio is defined as (number of reads containing the reference allele / total number of reads covering the SNP). The allelic ratio difference of this SNP is defined as the absolute difference between its observed allelic ratio and the simulated allelic ratio. At different thresholds of allelic ratio difference (0 to 0.2, x-axis), the FDR of SNP quantification is defined as FP / (FP+TP), where FP and TP are numbers of false and true positives, respectively. At each threshold, a given SNP is defined as a true positive if its allelic ratio difference is less than the threshold; otherwise, this SNP is a false positive. RASER was used with the

Figure S3

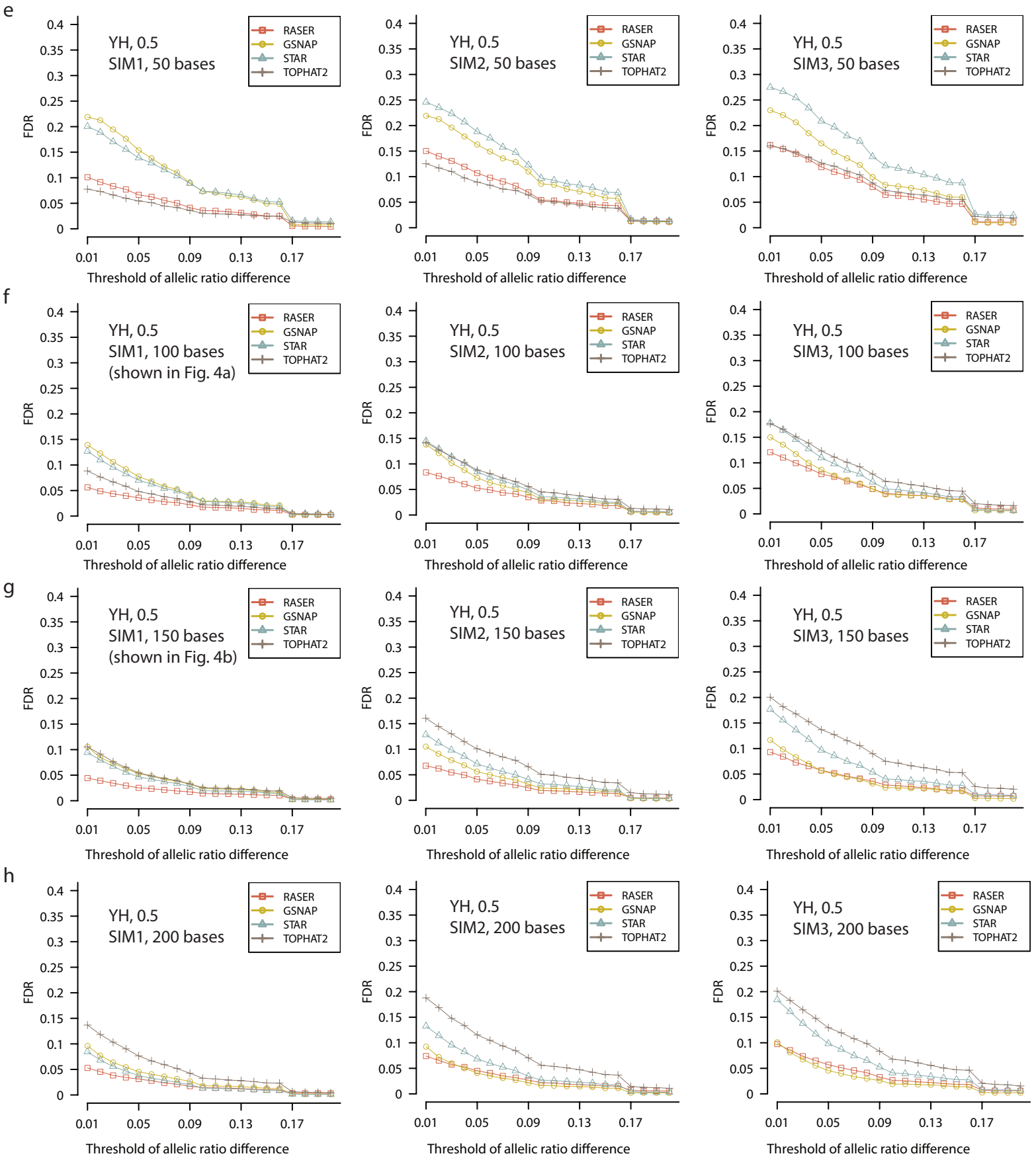


Figure S3. False discovery rate in the quantification of allelic ratios of SNPs expressed in RNA-Seq (continued). "obviously best" scheme; other aligners reported uniquely mapped reads. Each panel is labeled to show the corresponding SNPs (GM12878 or YH), simulated allelic ratio, simulation data set (SIM1, SIM2 or SIM3) and read length (50, 100, 150 or 200) used to generate the corresponding plot.

Figure S3

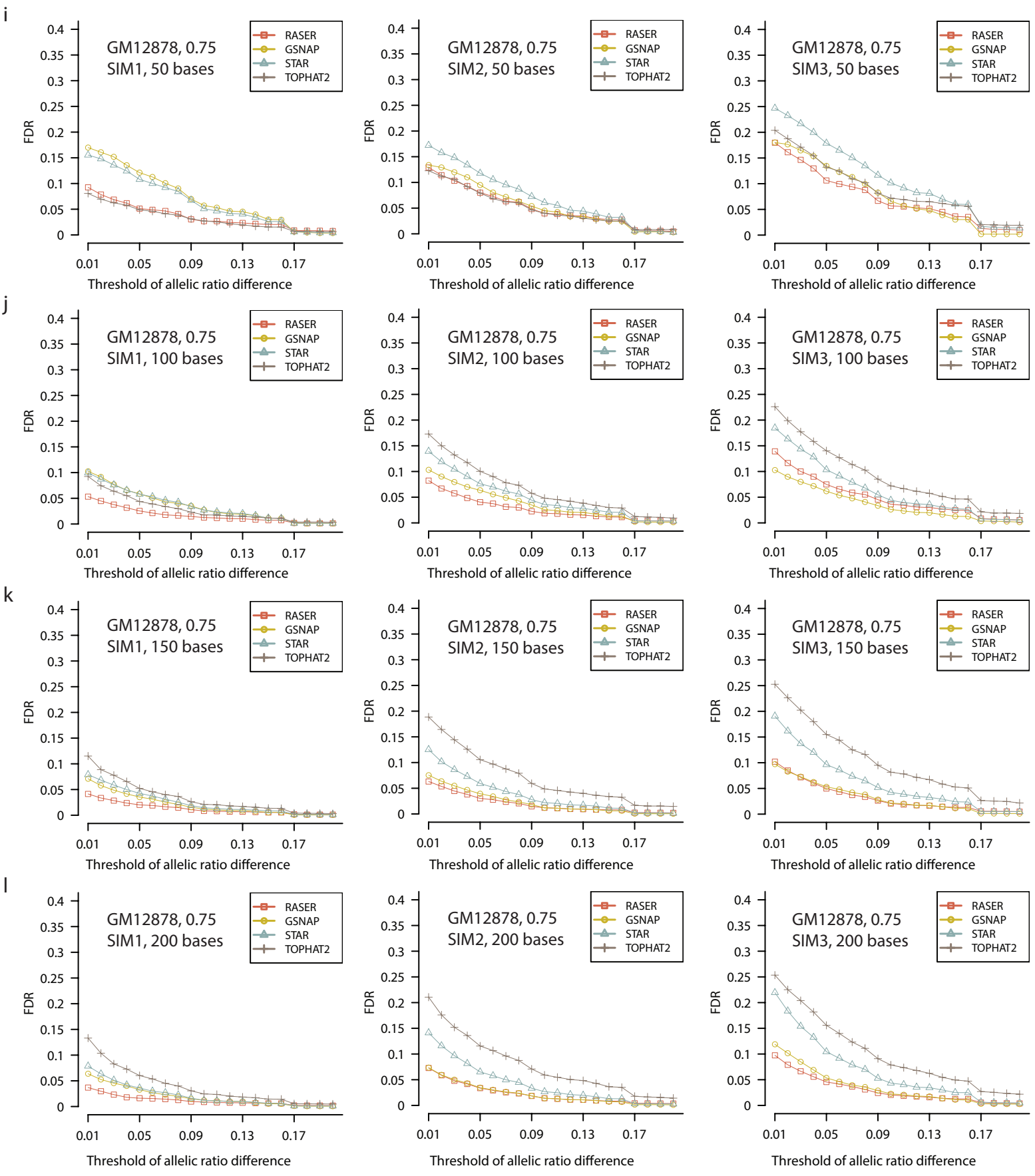


Figure S3. False discovery rate in the quantification of allelic ratios of SNPs expressed in RNA-Seq (continued).

Figure S3

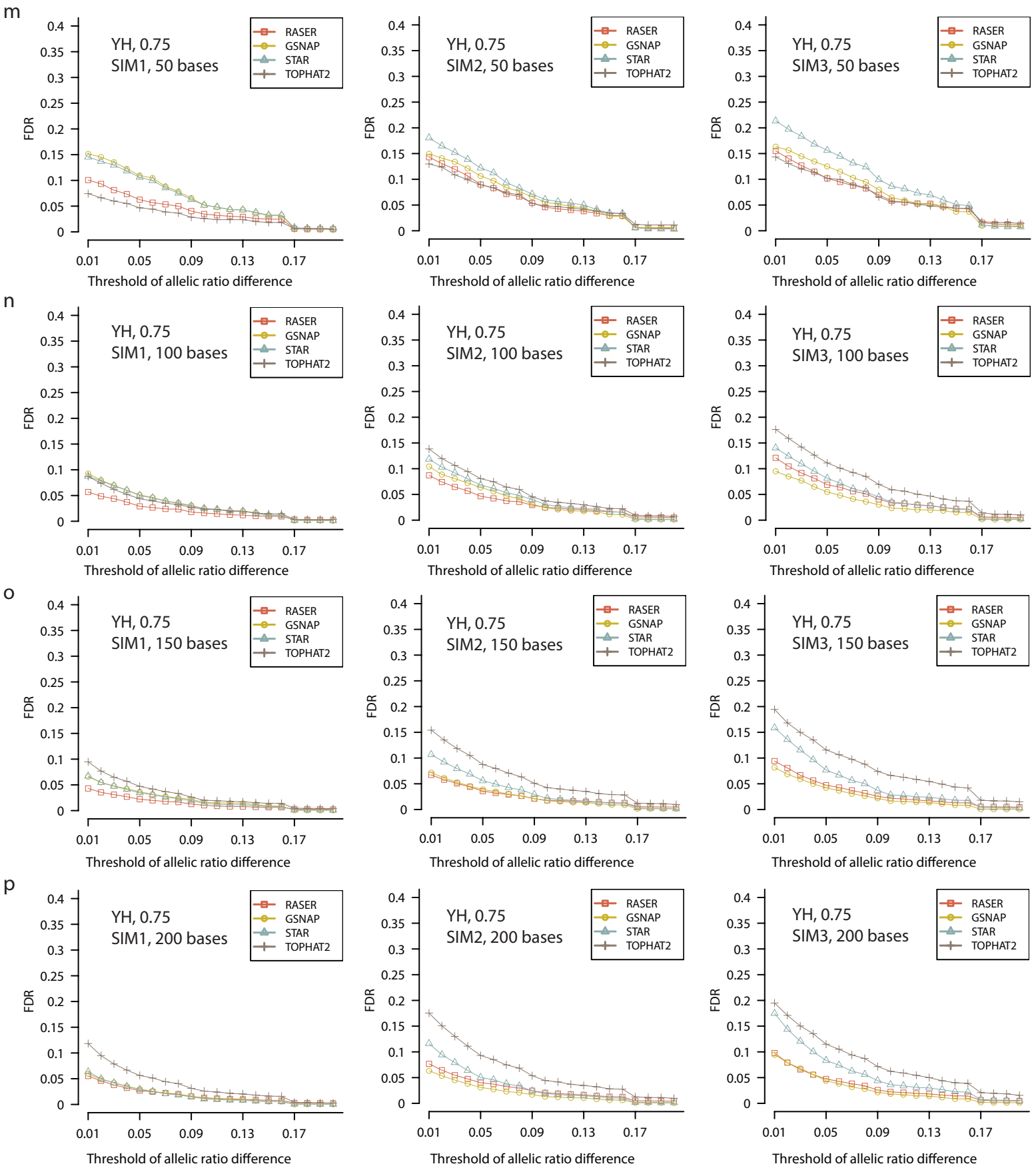


Figure S3. False discovery rate in the quantification of allelic ratios of SNPs expressed in RNA-Seq (continued).

Figure S3

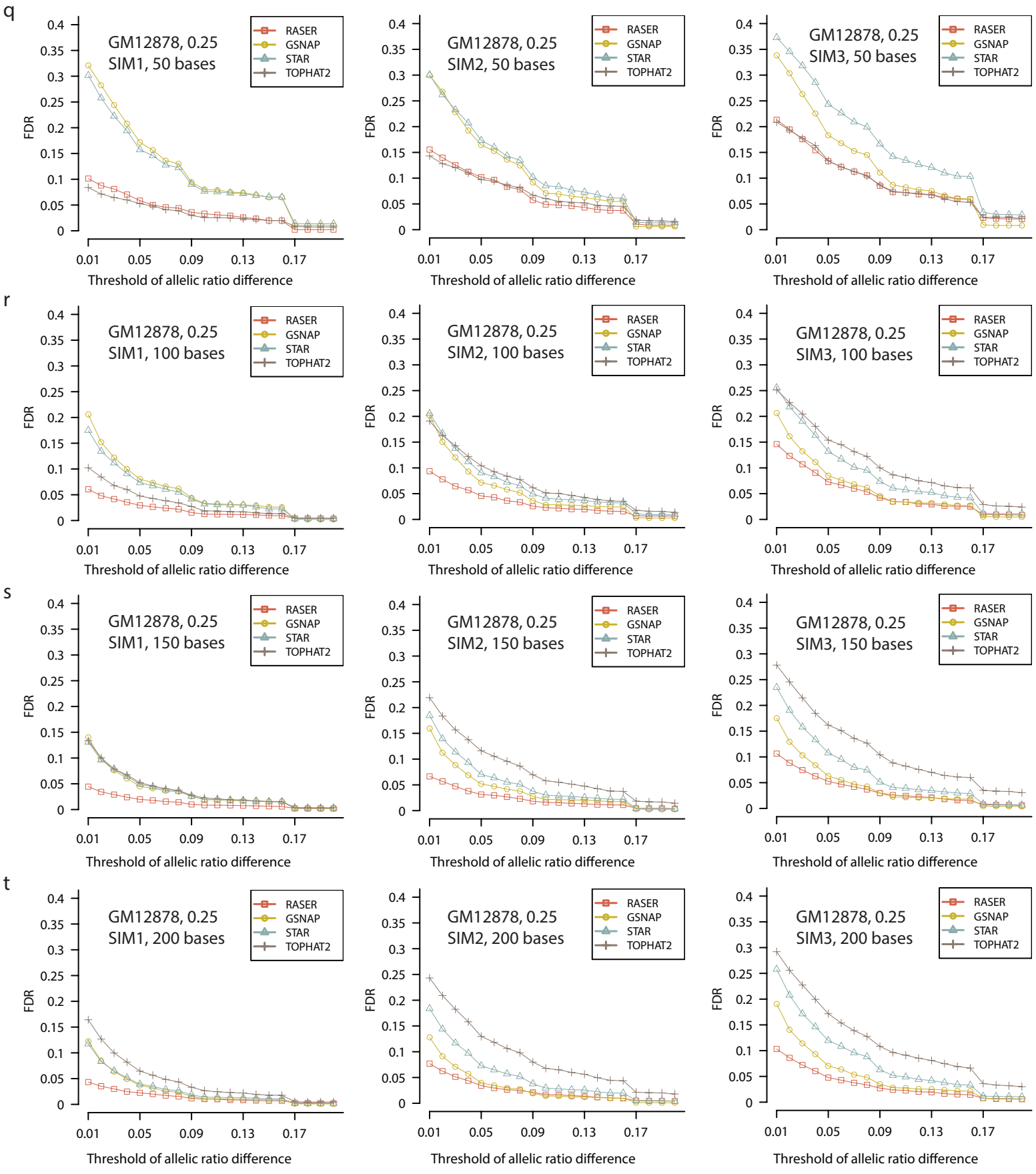


Figure S3. False discovery rate in the quantification of allelic ratios of SNPs expressed in RNA-Seq (continued).

Figure S3

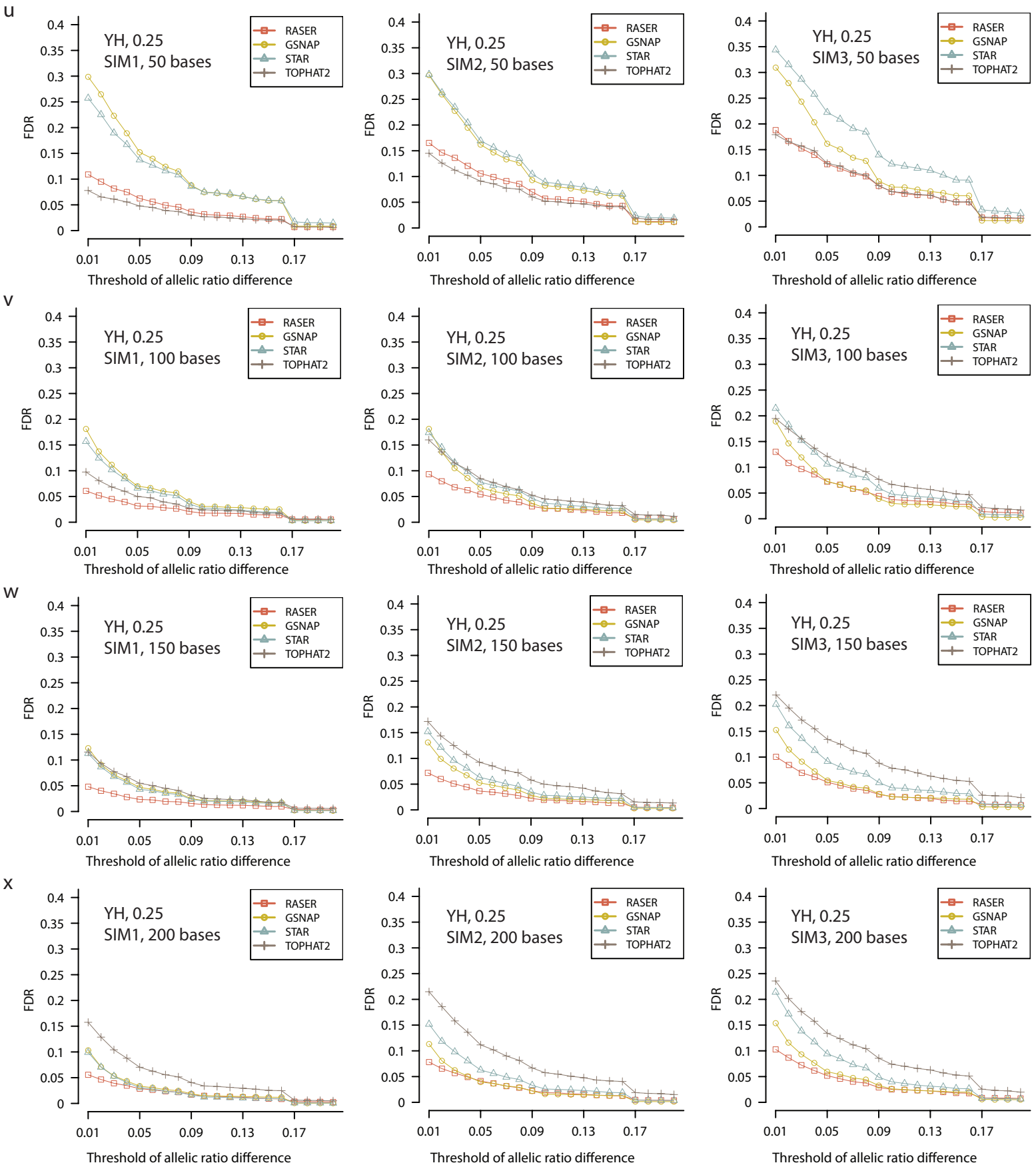


Figure S3. False discovery rate in the quantification of allelic ratios of SNPs expressed in RNA-Seq (continued).

Figure S4

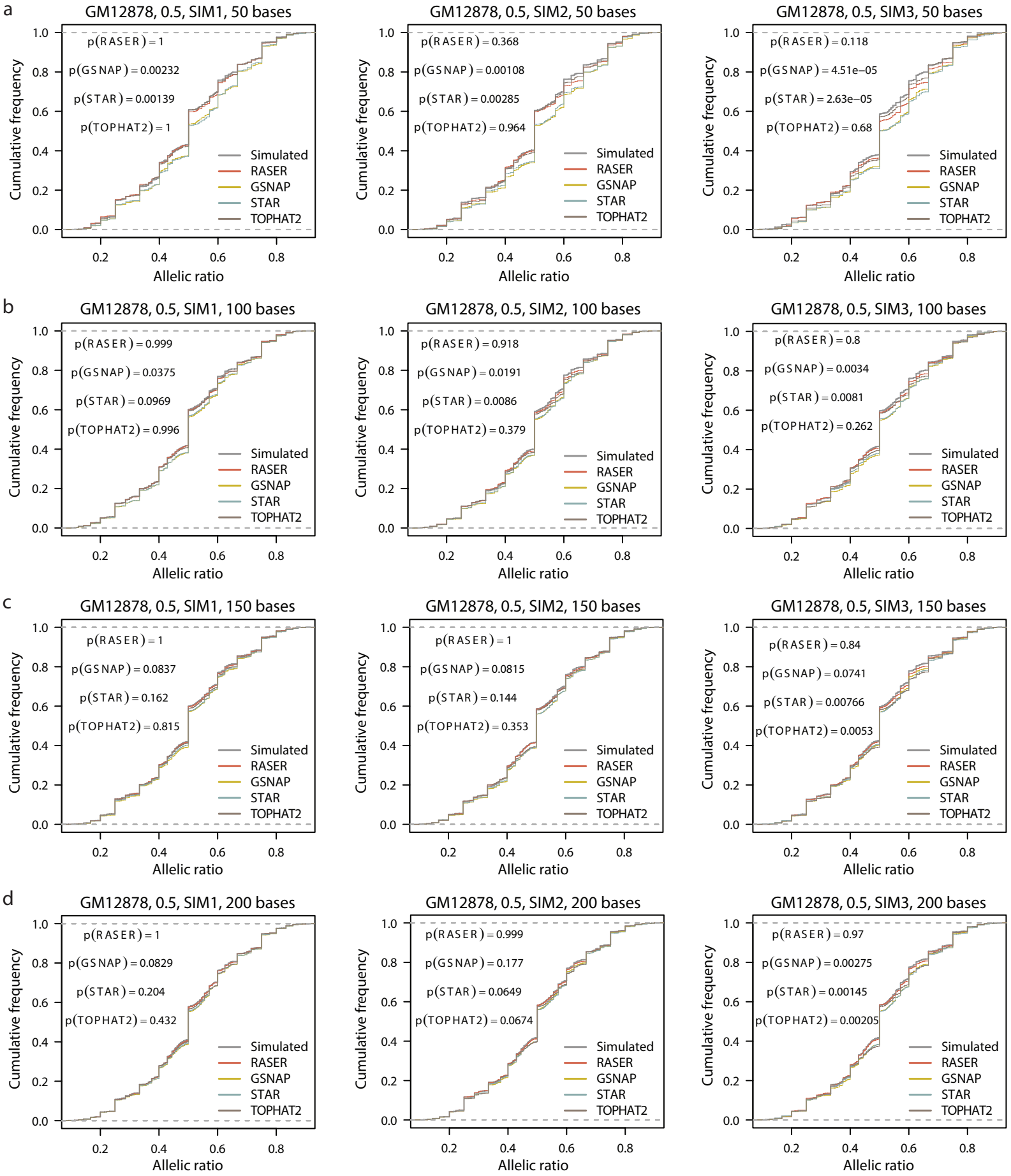


Figure S4. Distributions of simulated and observed allelic ratios based on various read alignment algorithms. SNPs in the GM12878 and YH samples which had corresponding whole genome sequencing data were implanted into the simulation data of 50, 100, 150 and 200 bases in length. SIM1, SIM2 and SIM3 data were used for the plots in the first, middle and last column, respectively. For each read that contains a known SNP, the probability for the SNP to have the reference or alternative allele was set to be 0.5, 0.75 and 0.25, similarly as in Figure S3. Each panel is labeled in the same way as in Figure S3. P-values were calculated using the Kolmogorov–Smirnov (KS) test between the simulated and predicted allelic ratios from each aligners. For all data sets, RASER was used with the "obviously best" mapping scheme; other aligners reported uniquely mapped reads.

Figure S4

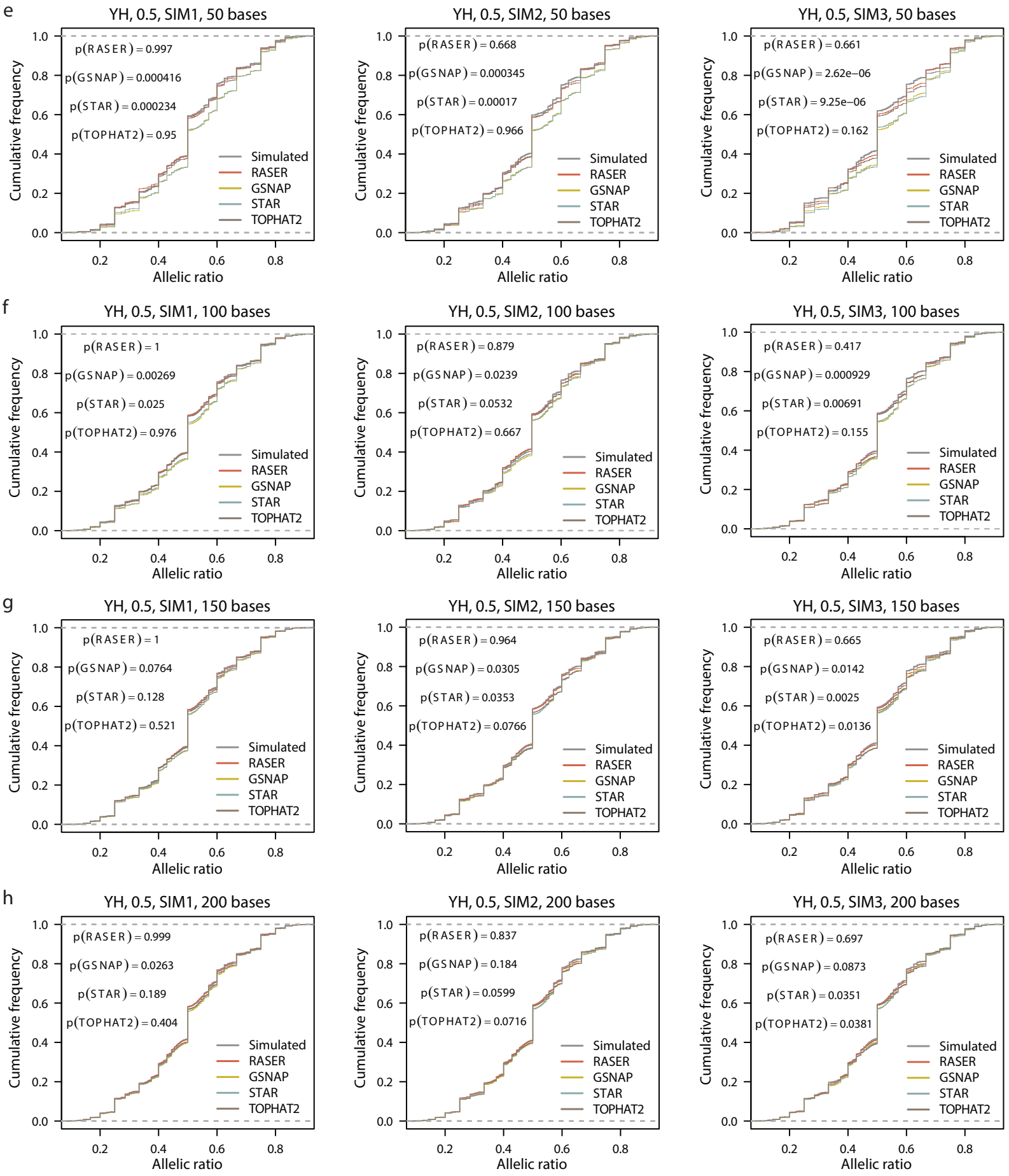


Figure S4. Distributions of simulated and observed allelic ratios based on various read alignment algorithms (continued).

Figure S4

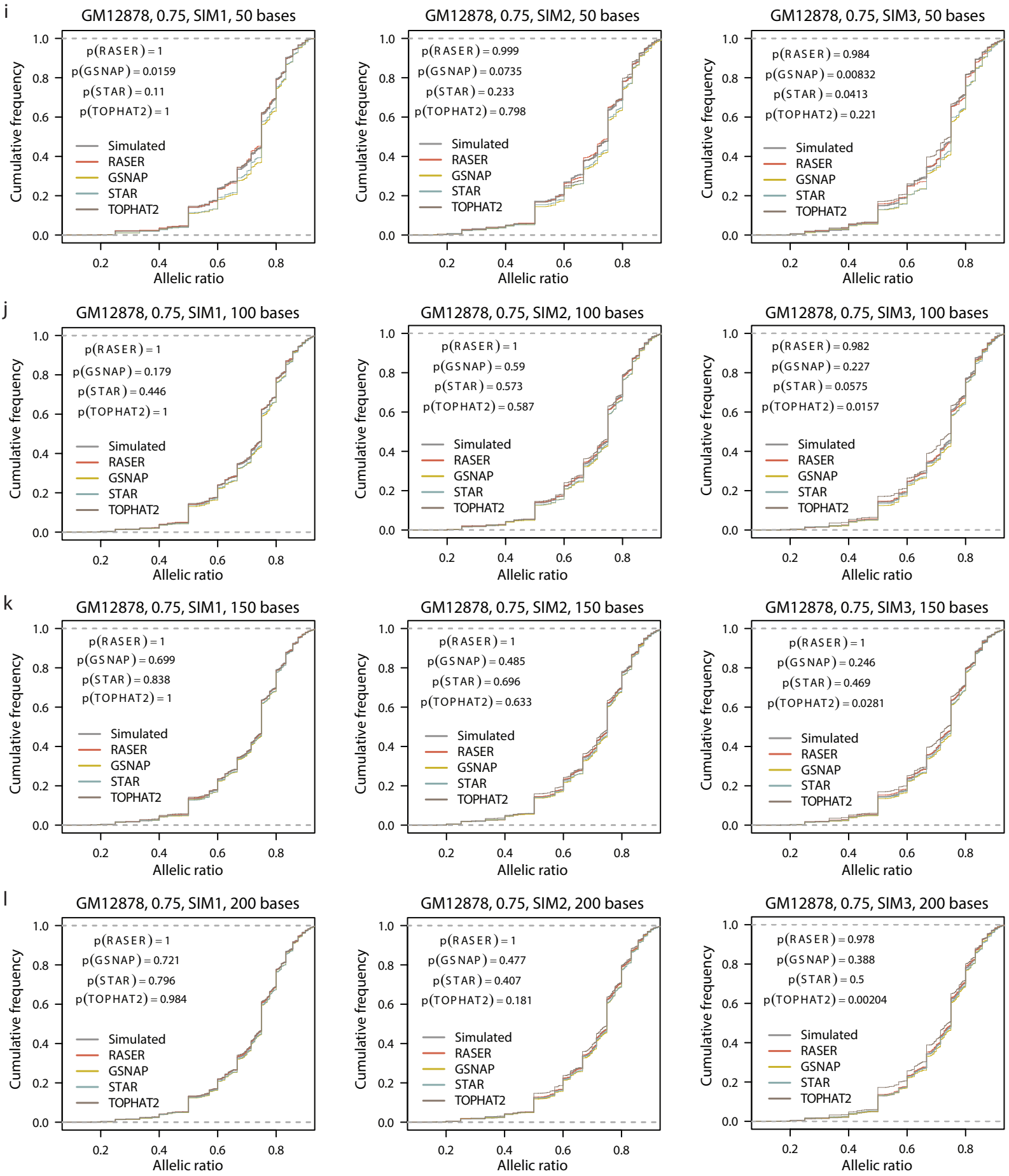


Figure S4. Distributions of simulated and observed allelic ratios based on various read alignment algorithms (continued).

Figure S4

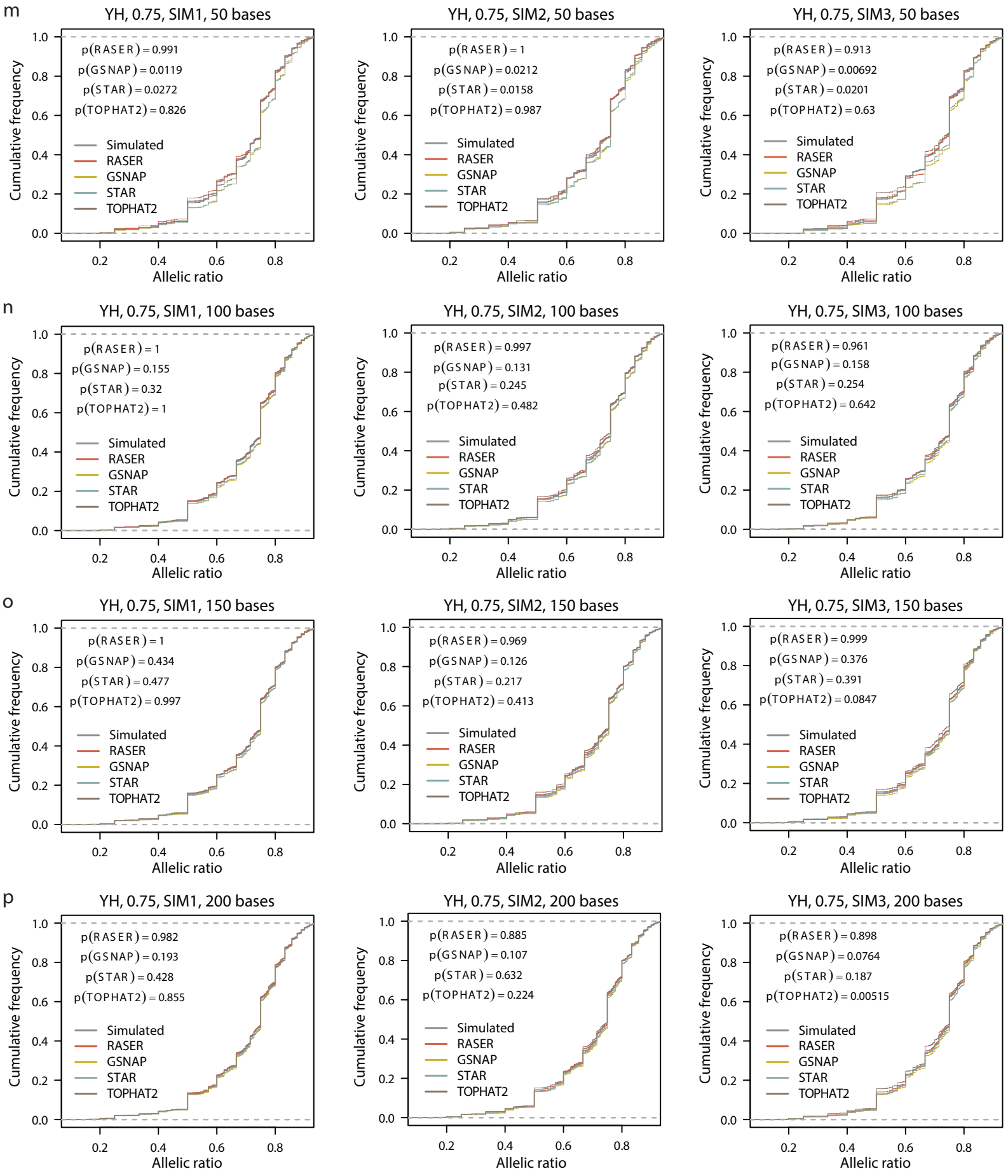


Figure S4. Distributions of simulated and observed allelic ratios based on various read alignment algorithms (continued).

Figure S4

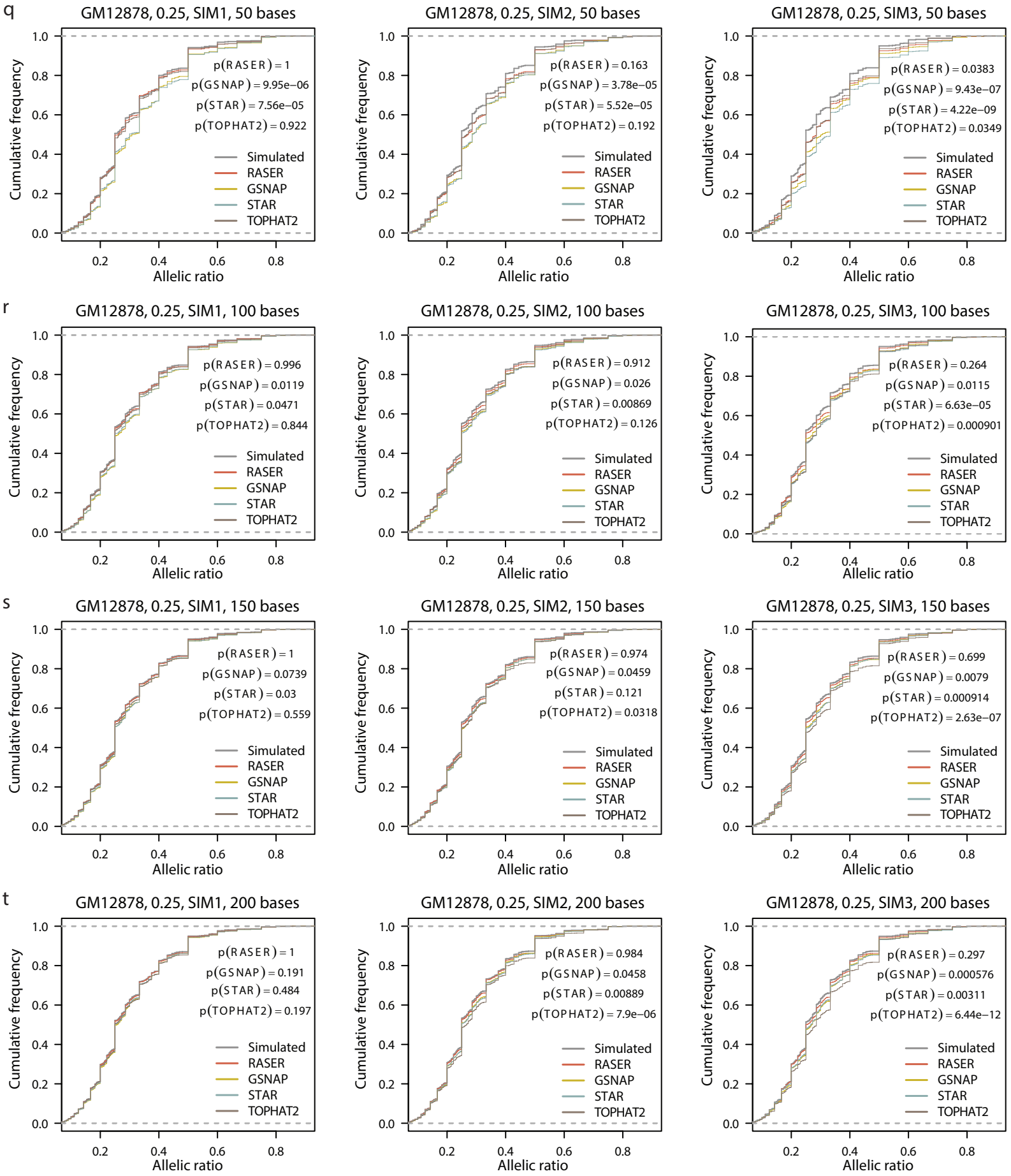


Figure S4. Distributions of simulated and observed allelic ratios based on various read alignment algorithms (continued).

Figure S4

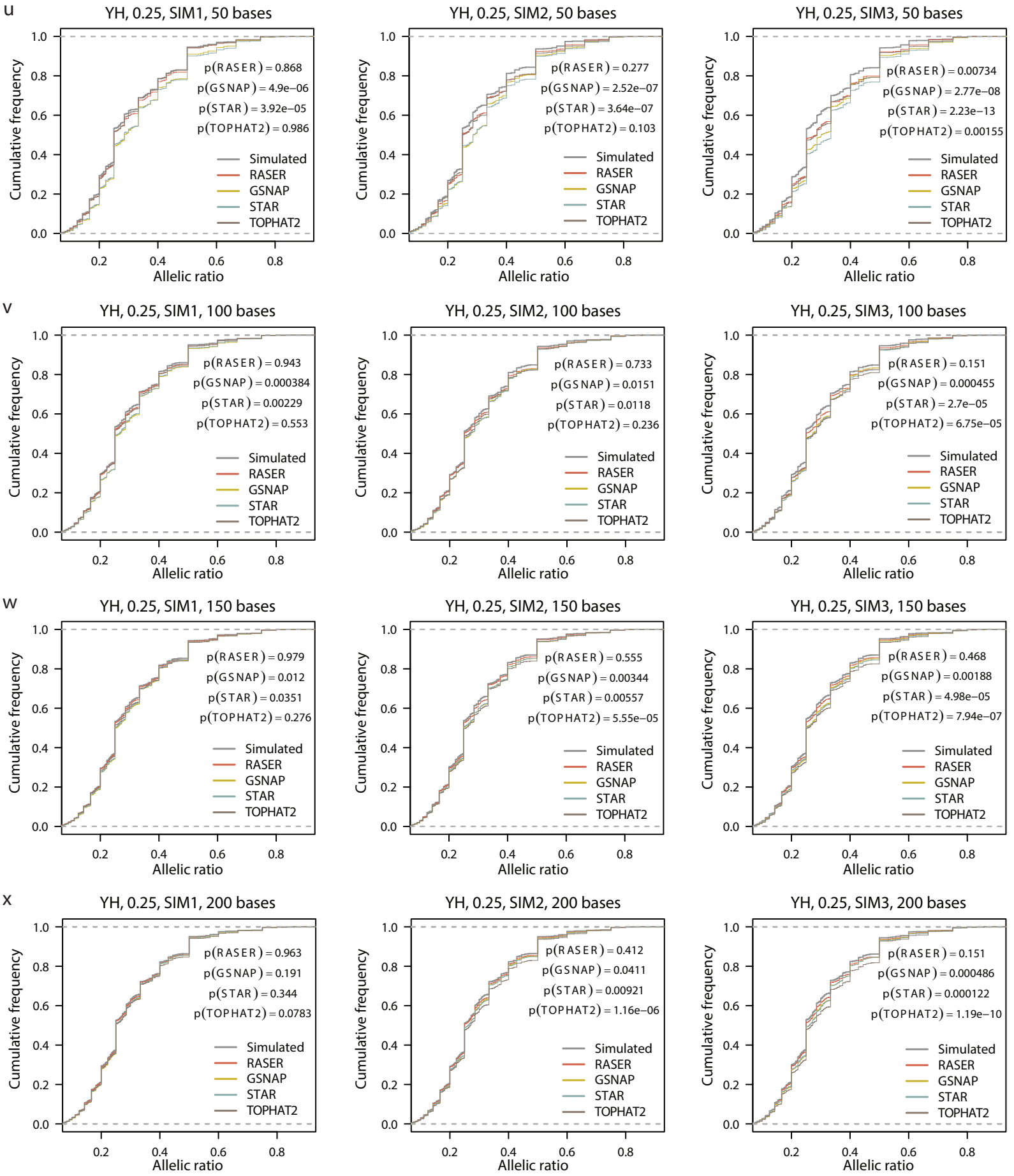
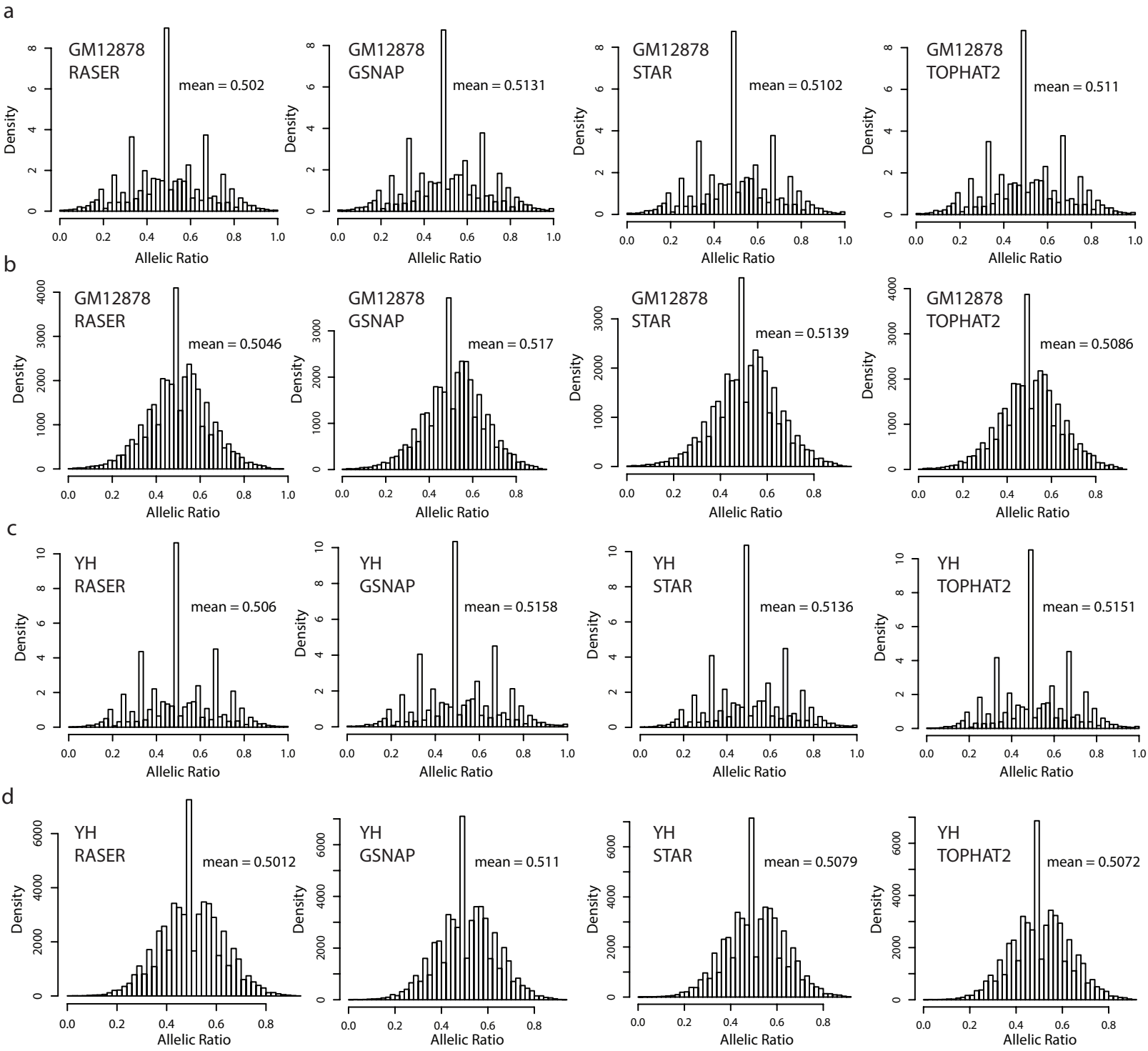


Figure S4. Distributions of simulated and observed allelic ratios based on various read alignment algorithms (continued).

Figure S5



e

(GM12878)	RASER	GSNAP	STAR	TOPHAT2
true positive	64845	62773	63208	61469
false positive	41499	50138	52119	38887
FDR	39.02%	44.40%	45.19%	38.75%

f

(YH)	RASER	GSNAP	STAR	TOPHAT2
true positive	94835	93779	94856	90013
false positive	32440	43353	46673	33243
FDR	25.49%	31.61%	32.98%	26.97%

Figure S5. Distribution of allelic ratios of heterozygous SNPs in the GM12878 or YH data sets calculated using reads mapped by different alignment algorithms. For (a) and (c), customized post-mapping filters were applied prior to calculation of allelic ratios, including removal of duplicated mappings, removal of reads where the mismatches (corresponding to SNPs) were (1) within 5nt of read ends or (2) with Sanger base quality less than 30, and removal of SNPs with less than 5 reads. We calculated allelic ratios of the remaining mismatches. For (b) and (d), we applied GATK (with default parameters) to the mapping results and calculated allelic ratios of detected SNPs. Allelic ratios were defined as (number of reads harboring the reference allele / total number of reads overlapping the SNP). (e) Performance of aligners in SNP calling based on GATK using GM12878 data. True positives were defined as correctly detected SNPs by GATK when compared to known GM12878; false positives were defined as SNPs which 1) do not overlap with known GM12878 SNPs, and 2) do not overlap with known human editing sites from the RADAR database (Ramaswami and Li, 2014) and 3) have more than 20 mapped reads. The false discovery rate (FDR) was calculated as #false positives / (#true positives + #false positives). (f) Similar as (e), for YH data. For all data sets, RASER was used with the "obviously best" mapping scheme; other aligners reported uniquely mapped reads.

Figure S6

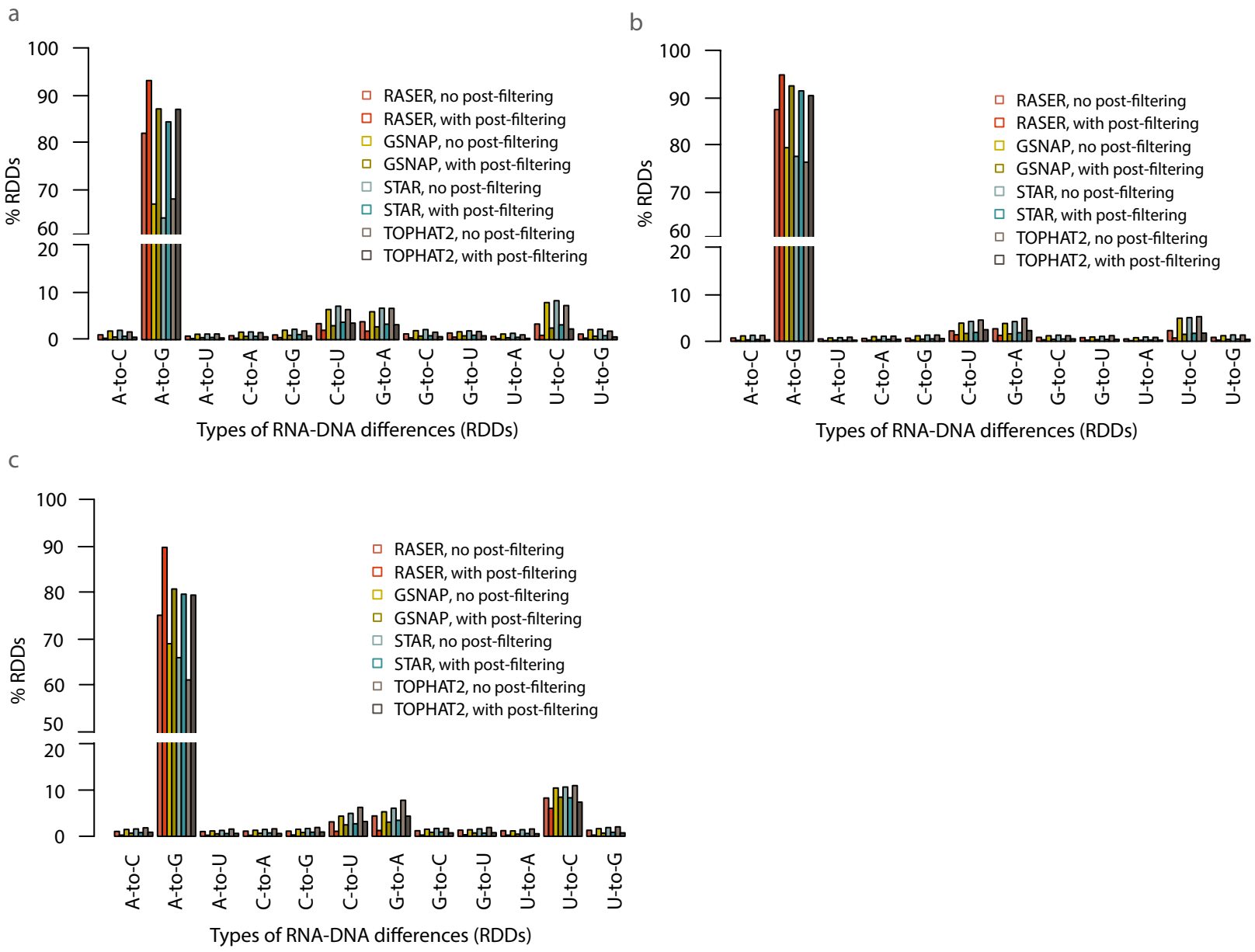


Figure S6. Identification of RNA editing sites in (a) K562, (b) GM12878 and (c) YH data sets. Y axis shows percentage of different types of RNA-DNA differences (RDDs) among all RDDs identified by each aligner, prior to and after post-mapping artifact filtering, respectively. The artifact filters remove RDDs sites that satisfy one of the following: 1) covered by reads with a strand bias, 2) with 100% editing, 3) with low-editing levels (<10% or <3 edited reads), 4) close to splice sites (i.e., ≤4 nt from spliced junctions), 5) within simple repeats (defined by Repeatmasker), 6) within homopolymer repeats (Repeatmasker), 7) overlapping known SNPs in public databases (dbSNP) (see Lee et al 2013 for details of these filters). For all data sets, RASER was used with the "obviously best" mapping scheme; other aligners reported uniquely mapped reads.

Figure S7

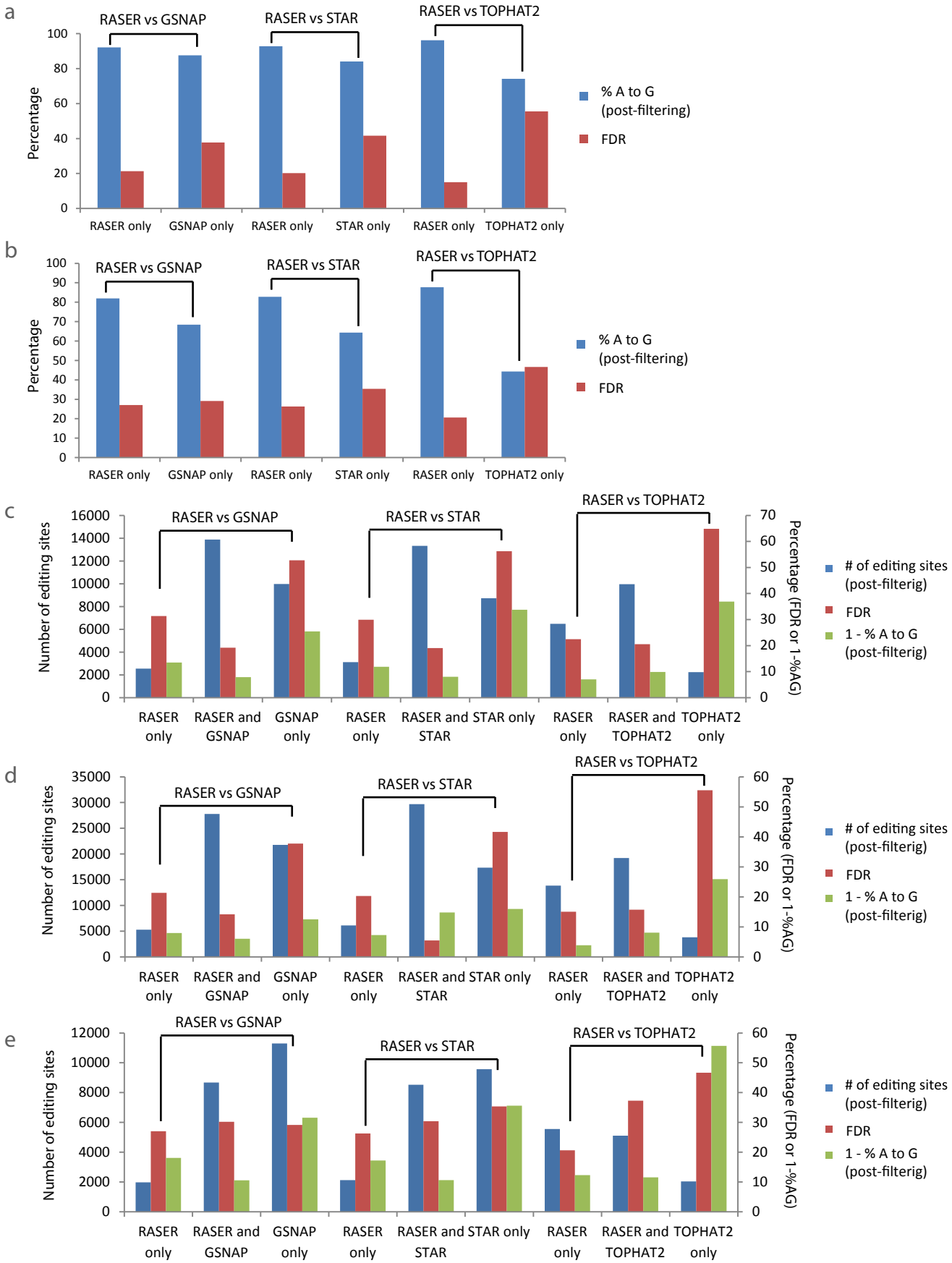


Figure S7 . Performance comparison of aligners in identifying RNA editing sites. (a) Percentage of A-to-G editing sites (blue) and FDR values (red) in analyzing GM12878 data. FDR is defined as the percentage of originally predicted editing sites that were removed by the filters which remove RDDs sites as described in Figure S6. RASER was compared to each of the other 3 aligners individually. Editing sites that were only identified by RASER (RASER only) or only by the other aligner are shown. (b) Similar as (a), but using YH data. (c) Number of editing sites (blue), FDR (red, as defined in (a)) and percentage of non-A-to-G editing sites (green, possible false positives) are shown. K562 data were analyzed. Similarly as in (a), RASER was compared to each of the other aligners. Editing sites were grouped into: "RASER only" (those that were identified only by RASER), "RASER and GSNAP" (those that were identified by both aligners) and "GSNAP only" (those there were identified only by GSNP), similarly for STAR and TOPHAT2. (d) Similar as (c), but for GM12878 data. (e) Similar as (c), but for YH data. For all data sets, RASER was used with the "obviously best" mapping scheme; other aligners reported uniquely mapped reads.

Figure S8

<READ Pair>

```
@READ1
TGCTGTTCATAAGTTTATTATCTATATGTGAAAAAATCAAGAAAATTGCTGGGTTTAGCTCTCAGCAGCCCGCCCC
+
IIIIIHIIHHIIIIIHIIIIHHIIHIIIIIIIIIHIIIIIGIIIIHIIIHIIIIHIIIIIDGIIHHGIIHHHHIIIG

@READ2
GCGTAACTCCAGACATGATGGAGGAGATGTATAAGAAAGCTCATGCTGCTATACGGGAGAATCCAGTCTATGAAAA
+
IIIIIIIIIIIIIFIIIIIIHIIHIFIIIGIIIIIHDIIIIIIIIGIGIHIBIDGIFIDIIIGHIHIIIEEBFEH
```

<Algorithm>

1. Get reverse complement sequences of READ1 and READ2 (➔ READ1_RC, READ2_RC)
2. For READ1, READ1_RC, READ2, READ2_RC
   1) Apply sliding window and retrieve each window's locations within genome and/or transcriptome using the index
   2) Extend mapping without allowing insertions/deletions (making fragments)
   3) Merge fragments
3. For merged fragments from (READ1 and READ2_RC) and (READ2 and (READ1_RC), get merged fragments pair which are within insert size
4. Fine mapping of merged fragment pair with Smith-Waterman algorithm
5. Report merged fragments pair

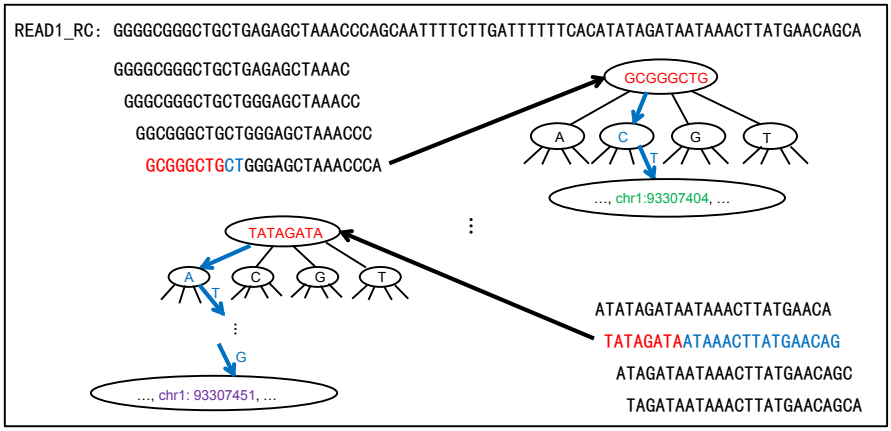
1. Get reverse complement sequences of READ1 and READ2 (➔READ1_RC, READ2_RC)

```
READ1:     TGCTGTTCATAAGTTTATTATCTATATGTGAAAAAATCAAGAAAATTGCTGGGTTTAGCTCTCAGCAGCCCGCCCC
READ1_RC:  GGGGCGGGCTGCTGAGAGCTAAACCCAGCAATTTTCTTGATTTTTTCACATATAGATAATAAACTTATGAACAGCA

READ2:     GCGTAACTCCAGACATGATGGAGGAGATGTATAAGAAAGCTCATGCTGCTATACGGGAGAATCCAGTCTATGAAAA
READ2_RC:  TTTTCATAGACTGGATTCTCCCGTATAGCAGCATGAGCTTTCTTATACATCTCCTCCATCATGTCTGGAGTTACGC
```

2. For READ1, READ1_RC, READ2 and READ2_RC,

1) Apply sliding window of length s+16 (s = 8), and retrieve each window's locations within genome and/or transcriptome using the index



2) Extend mapping without allowing insertions/deletions for all windows (complete search) ➔ fragments generated for all windows

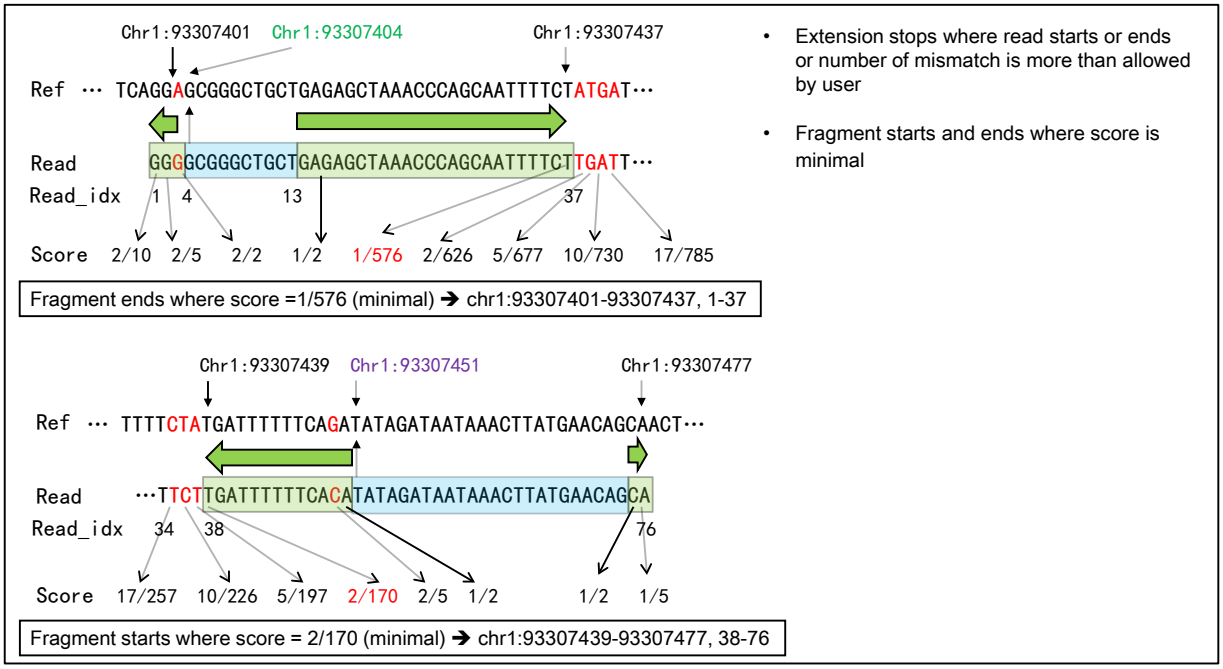
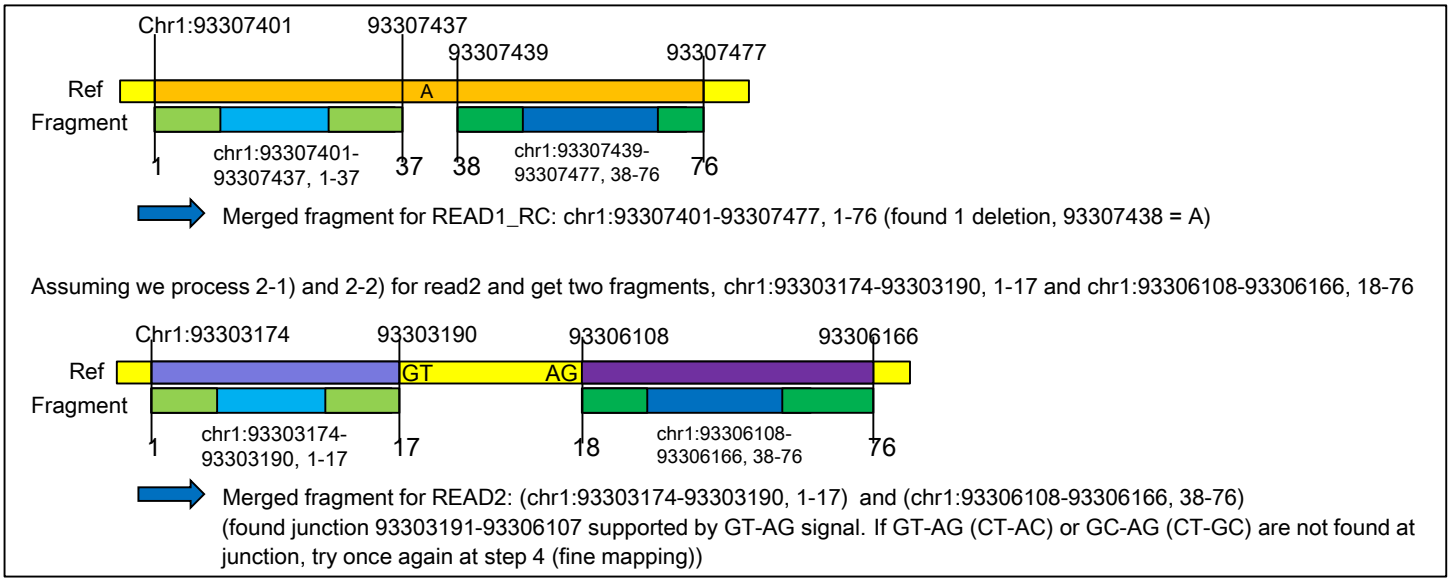


Figure S8. Flow of the entire alignment process using example 76 bp paired end reads.

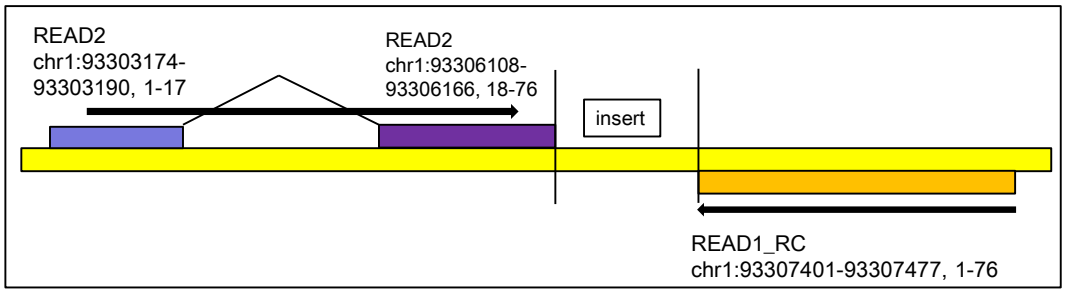
Figure S8

2. For READ1, READ1_RC, READ2, READ2_RC,

3) Merge fragments from all windows



3. For merged fragments from (READ1 and READ2_RC) and (READ2 and (READ1_RC), get merged fragments pair which are within insert size



4. Fine mapping of merged fragment pair with Smith-Waterman algorithm – minimizing mapping score (by mismatches and indels) and finding splicing signals (GT-AG (CT-AC) or GC-AG (CT-GC)) at splicing junctions, if they were not found at step 2-3) (splicing junctions are reported even if splicing signals are not found).

```
<READ1_RC>
     93307401                                                                    93307477
..TCAGGAGCGGGCTGCTGAGAGCTAAACCCAGCAATTTTCTATGATTTTTTCAGATATAGATAATAAACTTATGAACAGCAACT..
     GGGGCGGGCTGCTGAGAGCTAAACCCAGCAATTTTCT-TGATTTTTTCACATATAGATAATAAACTTATGAACAGCA
     1                                                                           76

<READ2>
     93303174        93303190   93306108                                         93306166
..ACAGCGTAACTCCAGACATGGT......AGATGGAGGAGATGTATAAGAAAGCTCATGCTGCTATACGAGAGAATCCAGTCTATGAAAAGAA..
     GCGTAACTCCAGACATG          ATGGAGGAGATGTATAAGAAAGCTCATGCTGCTATACGGGAGAATCCAGTCTATGAAAA
     1               17         18                                               76
```

5. Report merged fragments pair

```
READ1    83    chr1    93307401    37M1D39M    =    93303174    -4304
GGGGCGGGCTGCTGAGAGCTAAACCCAGCAATTTTCTTGATTTTTTCACATATAGATAATAAACTTATGAACAGCA
GIIIHHHHIIGHHIIGDIIIIIHIIIIIHIIIHIIIIGIIIIIHIIIIIIIIIHIIHHIIIIIHIIIIIHHIIHIIIII
*   MD:Z:2A34^C11G27    NH:i:1    HI:i:1    NM:i:*    AS:i:0.03947

READ2    163    chr1    93303174    17M2917N59M    =    93307401    4304
GCGTAACTCCAGACATGATGGAGGAGATGTATAAGAAAGCTCATGCTGCTATACGGGAGAATCCAGTCTATGAAAA
IIIIIIIIIIIIIIIFIIIIIIHIIHIFIIIGIIIIIHDIIIIIIIIIGIGIHIBIDGIFIDIIIGHIHIIIEEBFEH
*   MD:Z:55A20    NH:i:1    HI:i:1    NM:i:*    AS:i:0.01315
```

Figure S8. Flow of the entire alignment process using example 76 bp paired end reads (continued).

Table S1. Mapping results for simulated and actual RNA-Seq data sets.

		RASER (obviously best)	GSNAP (unique)	STAR (unique)	TOPHAT2 (unique)
50bps	SIM1 (1M read pairs)	943,552	998,449	983,562	857,462
	SIM2 (1M read pairs)	898,789	994,506	972,622	762,305
	SIM3 (1M read pairs)	817,034	984,233	940,661	650,775
100bps	SIM1 (1M read pairs)	984,581	996,540	993,619	929,975
	SIM2 (1M read pairs)	961,401	995,977	989,543	815,562
	SIM3 (1M read pairs)	909,735	993,279	973,519	680,089
150bps	SIM1 (1M read pairs)	991,290	998,711	996,795	931,644
	SIM2 (1M read pairs)	970,215	998,844	992,713	779,501
	SIM3 (1M read pairs)	914,822	993,461	971,627	624,705
200bps	SIM1 (1M read pairs)	990,101	999,563	997,208	915,554
	SIM2 (1M read pairs)	970,153	996,031	987,130	731,583
	SIM3 (1M read pairs)	939,190	997,567	969,593	569,369
GM12878 (#raw read pairs = 223,216,889)		133,625,444	168,617,704	173,524,596	135,232,645
K562 (#raw read pairs = 213,271,407)		124,073,580	159,540,180	164,359,963	125,115,501
YH (#raw read pairs = 161,803,471)		102,365,610	132,786,247	133,830,768	100,221,747