Supplementary Information

Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions

Yutaka Saito[1,2], Junko Tsuji[3], Toutai Mituyama[1,2,*]

[1]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.
[2]Japan Science and Technology Agency, CREST, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan.
[3]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, Massachusetts 01655, USA.

* To whom correspondence should be addressed:
mituyama-toutai@aist.go.jp

Supplementary Note 1

Details for mC calling by other methods

We compared accuracy of mC calling between Bisulfighter and ten other tools. They included Bismark and BS_Seeker, which both use Bowtie in the read mapping procedure, and another Bowtie-based method used in Lister *et al*. 2011 (denoted by *Lister*). Also included were BatMeth, BRAT, BSMAP, BSmooth, MethylCoder, RMAP, and Novoalign. After mapping reads, mC levels were estimated by scripts enclosed in the software packages, or the same procedure as in Bisulfighter if such scripts were not available. We used the recommended combination of options for tuning each tool. These options were obtained from original papers, manuals in software packages, or personal communications with software developers. For BRAT, BSMAP, and Novoalign, since multiple combinations of options were recommended, we conducted separate experiments to determine the best settings (Supplementary Figures 13-15). The optimized options and further details are described in Supplementary Note 2.

Details for DMR detection by other methods

Bisulfighter was compared with other methods for DMR detection found in the previous studies (Hansen *et al*. 2012; Lister *et al*. 2011). Due to limited availability or applicability of software packages, these methods were re-implemented in the present study if needed. We implemented a method based on the Fisher's exact test (denoted by *Fisher*) (Lister *et al*. 2011), which determines individual DMCs by the statistical test, then chains neighbor DMCs using the distance parameter of 300 bp. The method controls the number of output DMRs by changing the threshold for *P*-values in the statistical tests. We also implemented a method based on smoothing techniques used in the BSmooth package (denoted by *Smoothing*) (Hansen *et al*. 2012). Since the original version of BSmooth requires biological replicates for DMR detection, we implemented a modified version applicable to samples without replicates. Specifically, the modified method performs smoothing of mC levels, detects individual DMCs based on the difference in smoothed mC levels, and then chains neighbor DMCs within 300 bp. The method controls the number of output DMRs by changing the difference threshold for smoothed mC levels. We used the distance parameter of 300 bp as recommended in the original paper describing BSmooth (Hansen *et al*. 2012).

The BSmooth package has its own module for mC calling. However, we found that mC calling by BSmooth resulted in poor performance (Figure 2). To focus our evaluation on accuracy of DMR detection, alignment results of LAST were used as the common input for Bisulfighter, Smoothing, and Fisher.

Supplementary Note 2

BatMeth

BatMeth 1.03 was executed with the following commands:

```
   batmeth -g reference.fa -i reads.fastq -n 3 -o filePrefix -O 1 -p 2
   split mappingOut.txt reference.fa 3 y filePrefix.0 filePrefix.2
```

After mapping reads, we counted mCs for each context, and calculated mC levels by dividing the mC counts by the number of reads mapped at the column.

Bismark

Bismark v0.7.6 was executed with the following commands:

```
   bismark_genome_preparation --verbose wkdir && cd wkdir
   trimReadEnd -q 3 reads.fastq > trimmed.fastq
   bismark -n 2 -l 50 wkdir trimmed.fastq
   methylation_extractor -s --comprehensive mappingOut.sam
```

trimReadEnd is an in-house script to trim the 3'-end of reads to just before the first base with phred scores below threshold. In the above commands, we used the quality threshold of 3.

BRAT

BRAT 1.2.4 was executed with the following commands:

```
   trim -s reads.fastq -P filePrefix -q $q -m 2
   brat -r refFileNames.txt -s filePrefix_reads1.txt -bs -m $m -f $f ¥
      -o mappingOut.txt
   acgt-count -r refFileNames.txt -P mappingOut.txt -s mcResultPrefix -B
```

We tested three combinations of options, (q, m, f) = (3, 0, 36), (3, 1, 36), (0, 4, 64), and determined (3, 1, 36) as the best setting (Supplementary Figure 13).

BS_Seeker

BS_Seeker was executed with the following commands:

```
   python BS_Seeker.py -d referenceDir/ -i reads.fastq -t N -e 50 -m 3 ¥
      -p bowtie-0.12.8/ -o mappingOut.txt
```

After mapping reads, we counted mCs for each context, and calculated mC levels by dividing the mC counts by the number of reads mapped at the column.

BSMAP

BSMAP 2.72 was executed with the following commands:

```
   bsmap -a reads.fastq -d refence.fa -o mappingOut.txt -v 10 -s 16 -w 2
```

After mapping reads, we tested two options for mC calling which care about SNPs or not:

```
   methratio.py -o mcResult.txt -d reference.fa -i 'correct' mappingOut.txt
   methratio.py -o mcResult.txt -d reference.fa mappingOut.txt
```

The option aware of SNPs was determined as the best setting (Supplementary Figure 14).

BSmooth

BSmooth 0.7.0 was executed with the following commands:

```
   perl bs_merman_align.pl --bsc ¥
      --merman=bsmooth-align-0.7.0/merman/merman ¥
      --output=dirPrefix -- reference.fa -- reads.fastq
   perl bsev_sort.pl --ev=dirPrefix/ --out=dirPrefix_sort/
   perl bsev_tabulate.pl --c=mcResult -- dirPrefix_sort -- reference.fa
```

Lister

We followed the whole procedure described in the literature (Lister *et al*. 2011) except the adaptor-trimming step which is not necessary for our simulated data.

MethylCoder

MethylCoder 5b849ac was executed with the following commands:

```
   methylcoder --gsnap gmap-2012-07-20/bin --outdir mappingOutDir ¥
      --extra-args '--quiet-if-excessive --npaths1' --mismatches=2 ¥
      --reference reference.fa reads.fastq
```

After mapping reads, we counted mCs for each context, and calculated mC levels by dividing the mC counts by the number of reads mapped at the column.

RMAP

RMAP v2.05 was executed with the following commands:

```
   rmapbs -v -Q -B -F refFileNames.txt -o mappingOut.txt reads.fastq
```

After mapping reads, we counted mCs for each context, and calculated mC levels by dividing the mC counts by the number of reads mapped at the column.

Novoalin

Novoalign V2.08.02 was executed with the following commands:

```
   novoalign -b 2 -c 1 --Q2Off -d reference.fa -o SAM -t $t ¥
      -f reads.fastq > mappingOut.txt
```

We tested three options, t = 60, 90 or the default value, and determined the default value as the best setting (Supplementary Figure 15). After mapping reads, we counted mCs for each mC context using samtools and scripts in the Novoalign package as follows:

```
grep -e “^@¥|ZB:Z:CT” mappingOut.txt | samtools view -T chrom -ubS - | ¥
   samtools sort - filePrefix.ct
grep -e “^@¥|ZB:Z:GA” mappingOut.txt | samtools view -T chrom -ubS - | ¥
   samtools sort - filePrefix.ga
samtools mpileup -BC 0 ¥
   -f reference.fa filePrefix.ct.bam filePrefix.ga.bam ¥
   > filePrefix.mpileup
novomethyl filePrefix.mpileup > mcResult.txt
```

Supplementary table legends

Supplementary Table 1. Summary of the datasets. (a) Simulated data. Mean error rates were calculated from Phred scores of reads. In the actual simulation, errors were introduced in a position-specific manner by DNemulator. **(b)** Real data. Any of the real datasets was produced from multiple sequencing experiments, and thus may consist of both single- and paired-end reads with various lengths. Source: identifier in the Sequence Read Archive. WGBS: whole-genome bisulfite sequencing. RRBS: reduced representation bisulfite sequencing.

Supplementary Table 2. Computational cost of each tool. Computation time was shown for 3M simulated reads, evaluated on the Intel Xeon 2.53 GHz CPU.

Supplementary figure legends

Supplementary Figure 1. Tuning Bisulfighter options in mC calling. Experiments similar to Supplementary Figures 5-7 were conducted. bf: Bisulfighter. ct: equal treatment of Cs and Ts. prob: weighting by alignment probability. qual: weighting by read quality.

Supplementary Figure 2. Distinct distance distributions among neighbor DMCs observed in real data. We conducted mC calling by Bisulfighter, and determined high confidence DMCs whose change in mC levels was more than 0.2 (UP), less than -0.2 (DOWN), or neither (NoCh). **(a)** Carcinogenesis dataset. **(b)** Adipogenesis dataset. **(c)** Hematopoiesis dataset. **(d)** Fibroblast development dataset. **(e)** Aging dataset. **(f)** Social rank dataset. **(g)** Neuronal development dataset. **(h)** Tet1 mutation dataset. **(i)** Leukemia subtype dataset. **(j)** IDH1 mutation dataset. See Supplementary Table 1b for the details of datasets. Analysis was focused on DMCs with a sufficient number of aligned reads (more than 20 for **a** and **b**; more than 10 for the others). **(k)** Statistical significance of the differences between distance distributions. *P*-values were calculated by the Kolmogorov-Smirnov test.

Supplementary Figure 3. State transition diagrams for HMMs in Bisulfighter. (a) Naïve model with one pair of states per direction. **(b)** Dual model with two pairs of states per direction.

Supplementary Figure 4. Estimation of mC levels at selected CpG sites above coverage threshold. Distributions of errors between estimated and true mC levels are shown as box plots (25th-75th percentile). CpG sites were selected by coverage threshold of **(a)** 3X, **(b)** 5X, and **(c)** 10X. **(d)** The number of captured CpG sites for given coverage threshold.

Supplementary Figure 5. Benchmark for mC calling in binary classification of mCs. For each context, cytosines were called as mCs if non-zero mC levels were estimated. Trade-off between the true positive rates and the number of false positives is shown for varying sequencing depth. Dataset A: low quality reads. Dataset B: high quality reads.

Supplementary Figure 6. Binary classification of mCs at varying sequencing depths. For each context, cytosines were called as mCs if non-zero mC levels were estimated. The true positive rate (TPR) and the false discovery rate (FDR) are shown for each sequencing depth. In FDR figures, the right panel is a closeup of the left panel. Dataset A: low quality reads. Dataset B: high quality reads.

Supplementary Figure 7. Benchmark for mC calling in estimation of mC levels. Distributions of errors between estimated and true mC levels are shown as box plots (25th-75th percentile). Dataset A: low quality reads. Dataset B: high quality reads.

Supplementary Figure 8. Benchmark for DMR detection at varying sequencing depth. For each DMR length, true positives with 50% reciprocal overlap are shown.

Supplementary Figure 9. Accuracy in determining individual DMCs. Accuracy is evaluated by the area under the receiver operating characteristic curve (AUC) in which DMCs are detected by varying thresholds for posterior probabilities (Bisulfighter), difference in smoothed mC levels (Smoothing), and *P*-values in statistical tests (Fisher).

Supplementary Figure 10. Benchmark for mC calling and DMR detection with the GC-rich chromosome 19. Experiments similar to **(a)** Figure 2b, **(b)** Figure 2c, and **(c)** Figure 3a were conducted using the chromosome 19, instead of the chromosome X, for simulating bisulfite-converted reads.

Supplementary Figure 11. Agreement between DMRs and DEGs determined by various thresholds of expression fold change. The dashed vertical line indicates 5-fold expression change used in Figure 3c.

Supplementary Figure 12. Agreement between DMRs and DSSs for TSS-proximal and TSS-distal regions at various resolutions. Proximal and distal regions are separated by means of the 5 kbp distance from the nearest TSS.

Supplementary Figure 13. Tuning BRAT options in mC calling. Experiments similar to Supplementary Figures 5-7 were conducted. See Supplementary Note 2 for details.

Supplementary Figure 14. Tuning BSMAP options in mC calling. Experiments similar to Supplementary Figures 5-7 were conducted. See Supplementary Note 2 for details.

Supplementary Figure 15. Tuning Novoalign options in mC calling. Experiments similar to Supplementary Figures 5-7 were conducted. See Supplementary Note 2 for details.

a Simulated data

Name	Error source	Mean error rate (%)	Read	Depth (M reads)	Protocol
Dataset A (low quality)	SRR019072	1.11	Single, 87 bp	1, 3, 5, 7, 10, 20, 50	WGBS
Dataset B (high quality)	SRR094461	0.41	Single, 85 bp	1, 3, 5, 7, 10, 20, 50	WGBS

b Real data

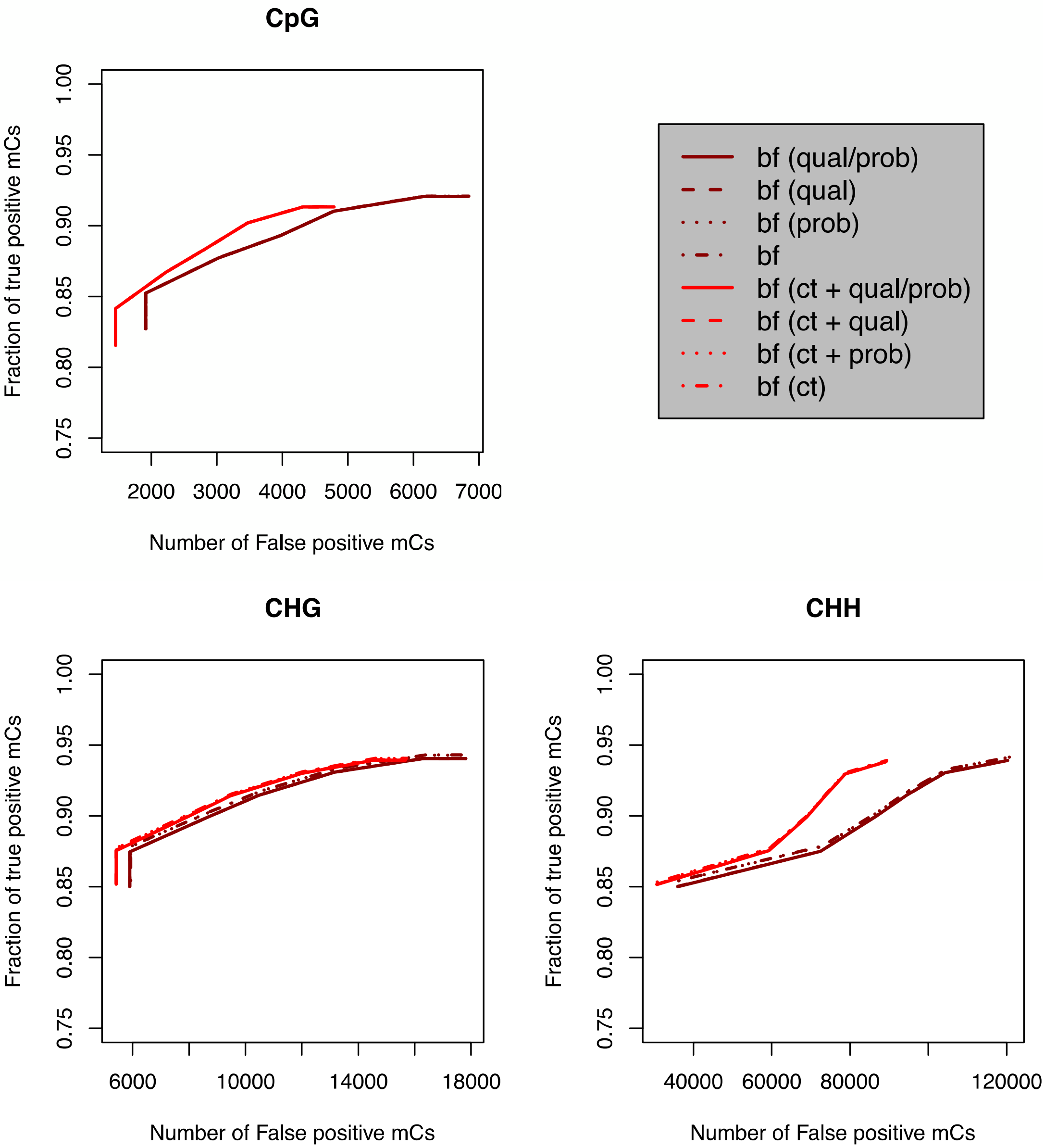
Name	Sample 1	Sample 2	Analysis in this study	Speceis	Read	Protocol	Source	Reference
Carcinogenesis	Breast cancer	Normal breast	Gene expression, Distance distribution	Human	Single	WGBS	SRP006728	Hon et al ., 2012
Adipogenesis	Mature fat cell	Adipose-derived stem cell	Gene expression, Distance distribution	Human	Paired	WGBS	SRP003529	Lister et al ., 2011
Hematopoiesis	Mature B cell	Hematopoietic stem cell	DNase I hypersensitivity, Distance distribution	Human	Both	WGBS	SRP008144	Hodges et al ., 2011
Fibroblast development	Foreskin fibroblast	Embryonic stem cell	DNase I hypersensitivity, Distance distribution	Human	Both	WGBS	SRP001720	Laurent et al ., 2010
Aging	Centenarian	Newborn	Distance distribution	Human	Paired	WGBS	SRP007820	Heyn et al ., 2012
Social rank	High-rank individual	Low-rank individual	Distance distribution	Rhesus	Single	WGBS	SRP009594	Tung et al ., 2012
Neuronal development	Neuronal progenitor cell	Embryonic stem cell	Distance distribution	Mouse	Both	WGBS	SRP007354	Stadler et al ., 2011
Tet1 mutation	Tet1-null primordial germ cell	Wildtype primordial germ cell	Distance distribution	Mouse	Both	WGBS	SRP016883	Yamaguchi et al ., 2012
Leukemia subtype	GMP-derived accute myeloid leukemia	HSC-derived accute myeloid leukemia	Distance distribution	Mouse	Single	RRBS	SRP017807	Krivtsov et al ., 2013
IDH1 mutation	IDH1(R132H) hematopoietic stem cell	Wildtype hematopoietic stem cell	Distance distribution	Mouse	Single	RRBS	SRP013732	Sasaki et al ., 2012

Supplementary Table 2

Tool	Time (m)
BatMeth	92
Bismark	138
BRAT	284
BS_Seeker	61
BSMAP	144
BSmooth	3051
Lister	126
MethylCoder	8271
RMAP	427
Novoalign	329
Bisulfighter	57

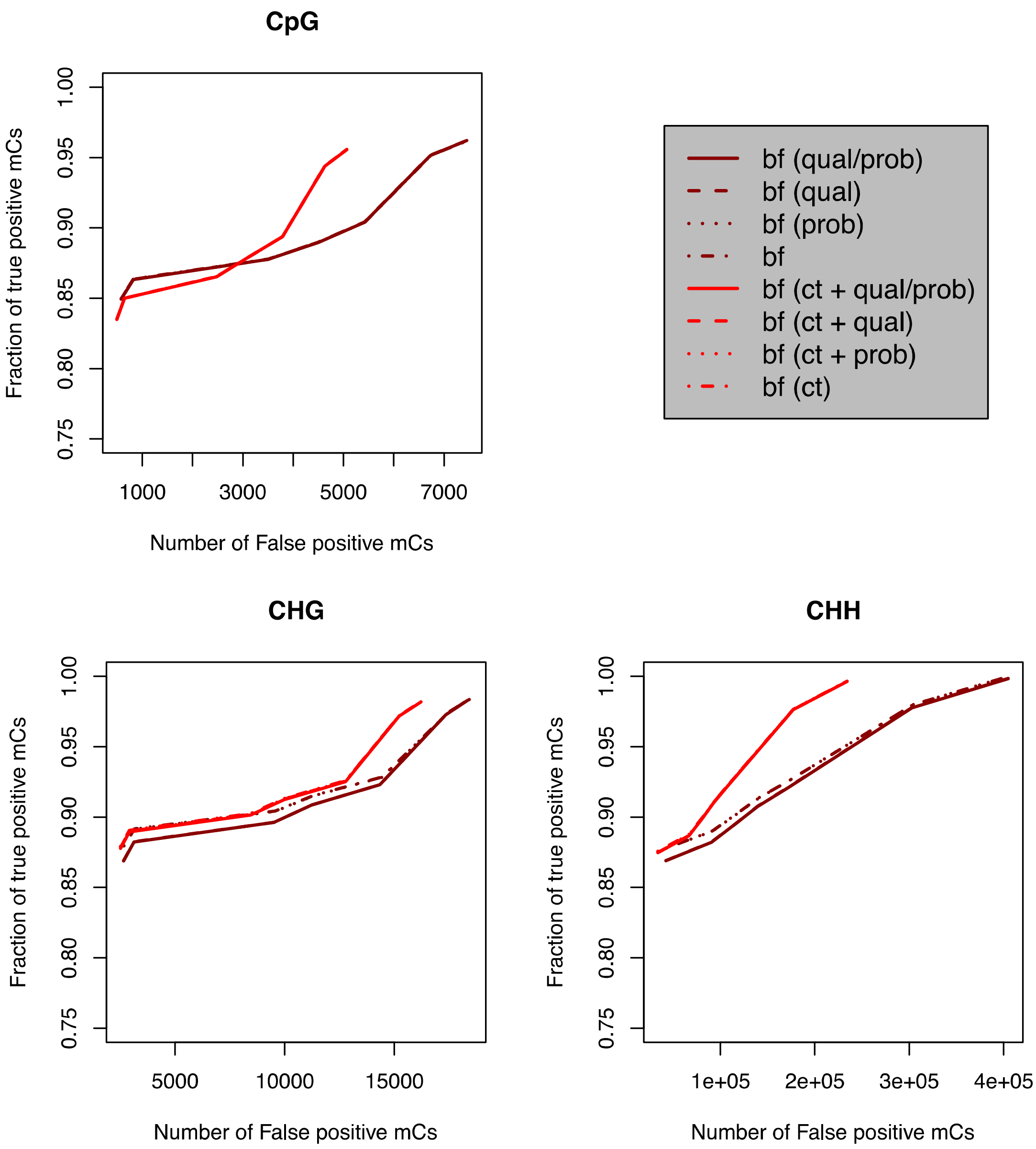
Supplementary Figure 1

Dataset A

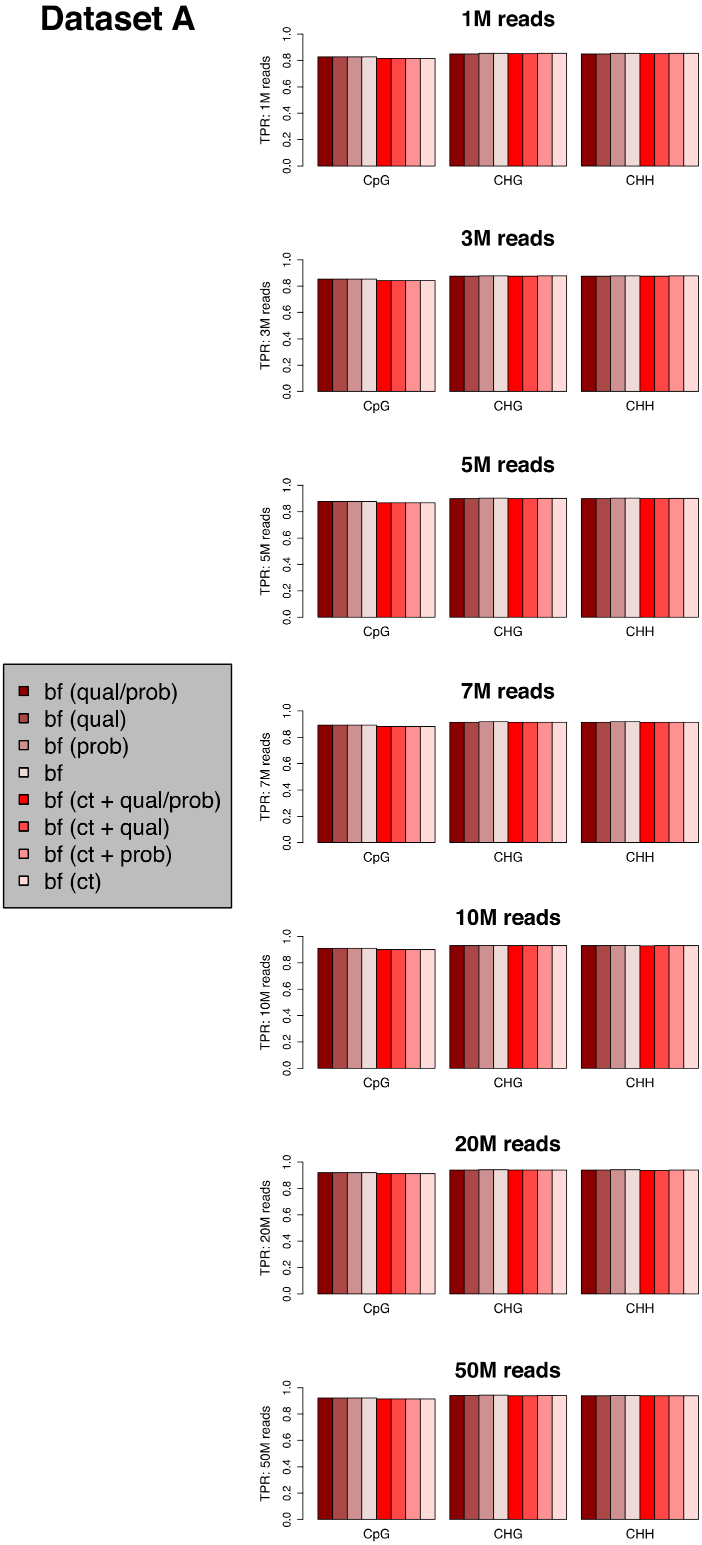


Supplementary Figure 1 (continued)

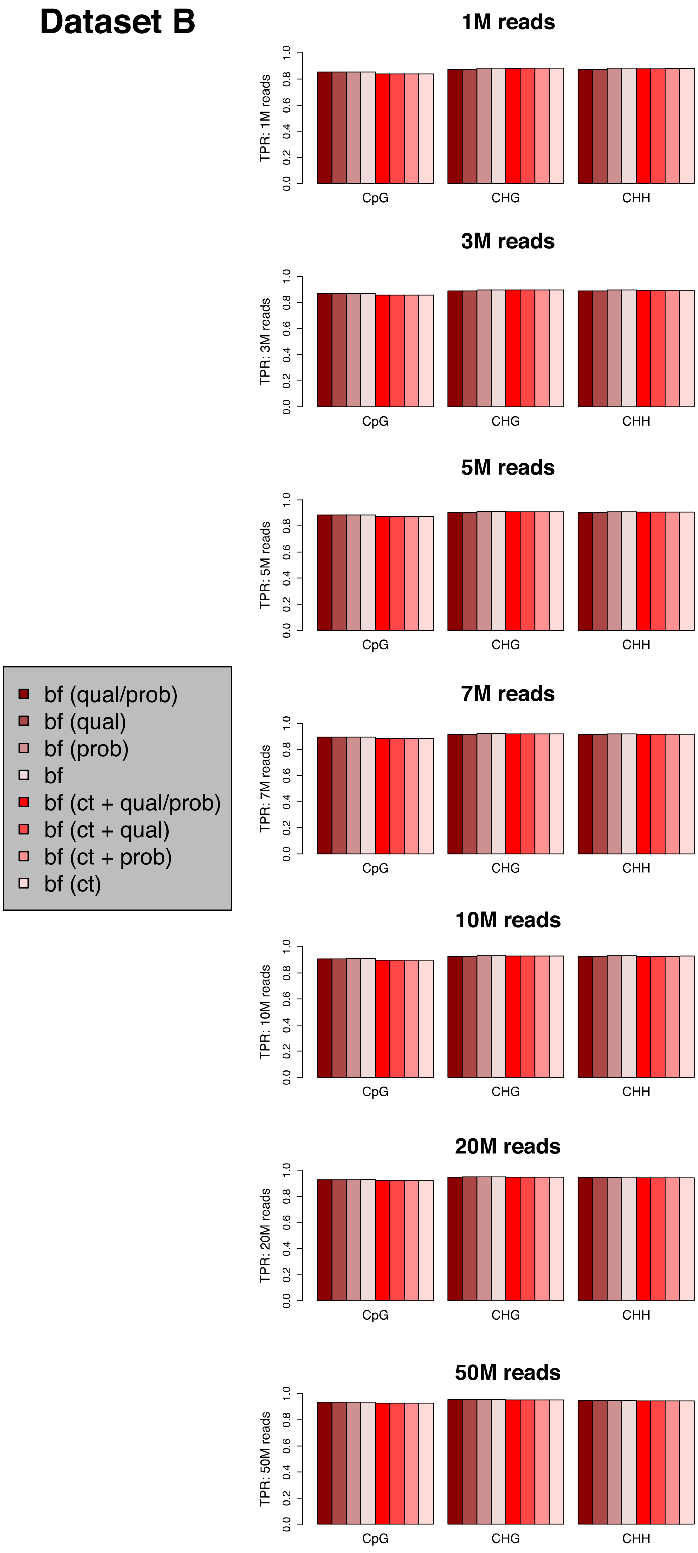
Dataset B



Supplementary Figure 1 (continued)

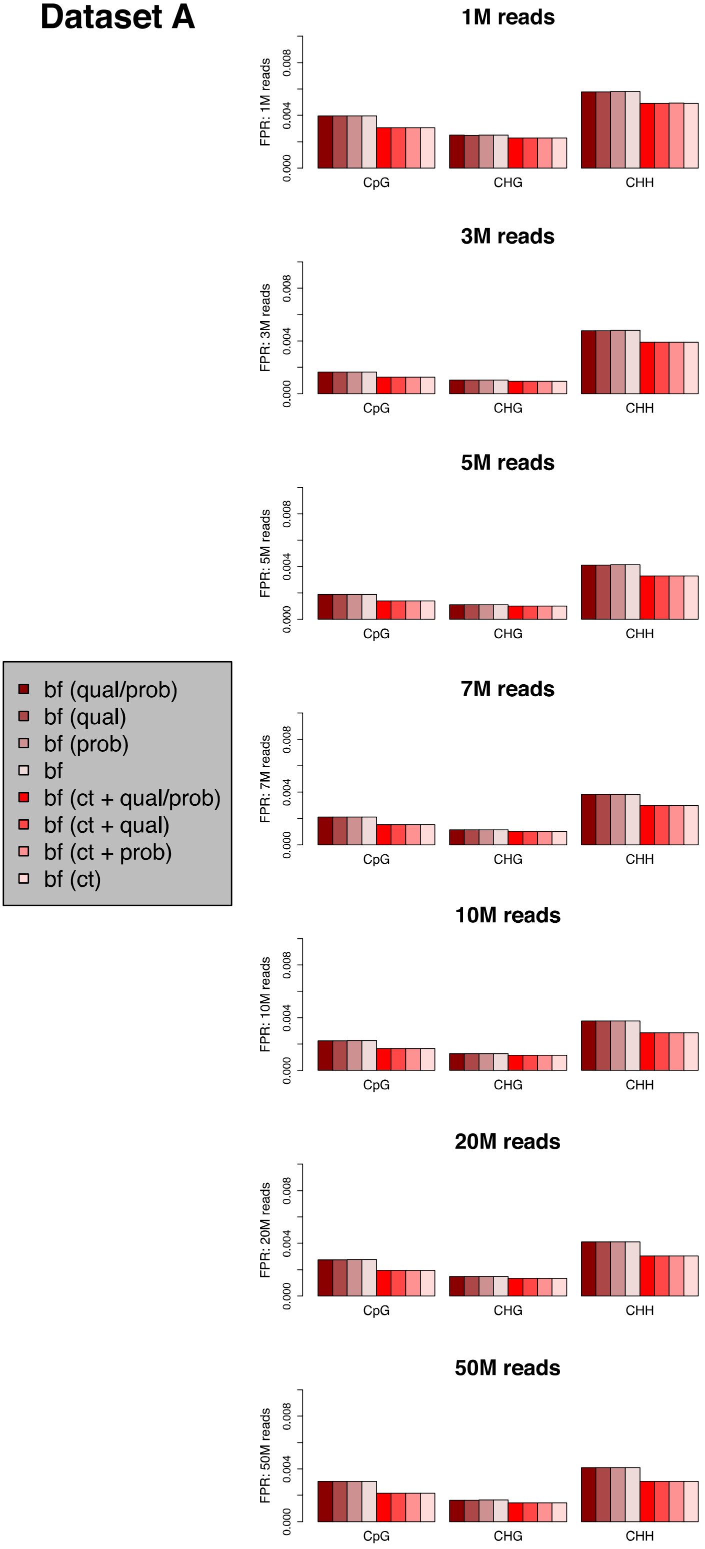


Supplementary Figure 1 (continued)



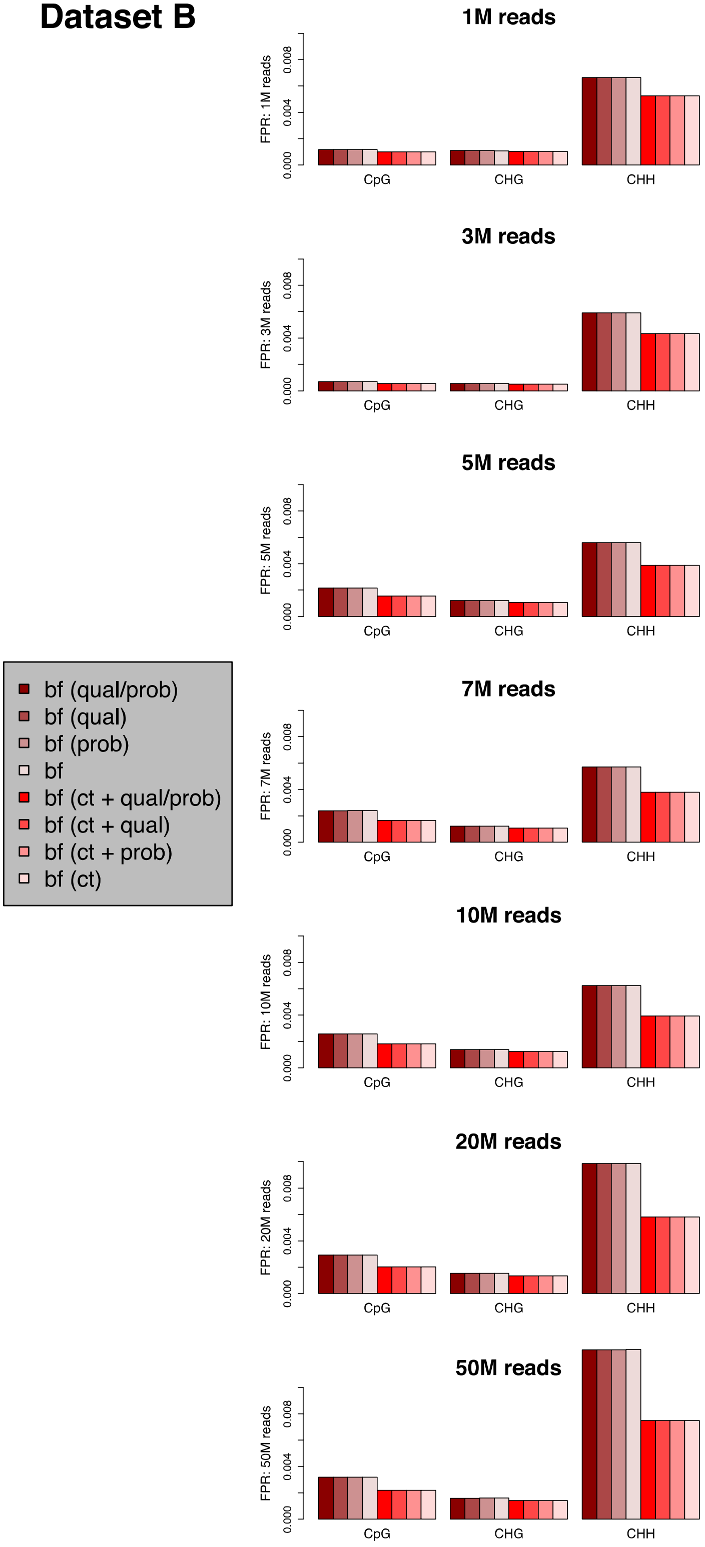
Supplementary Figure 1 (continued)

Dataset A



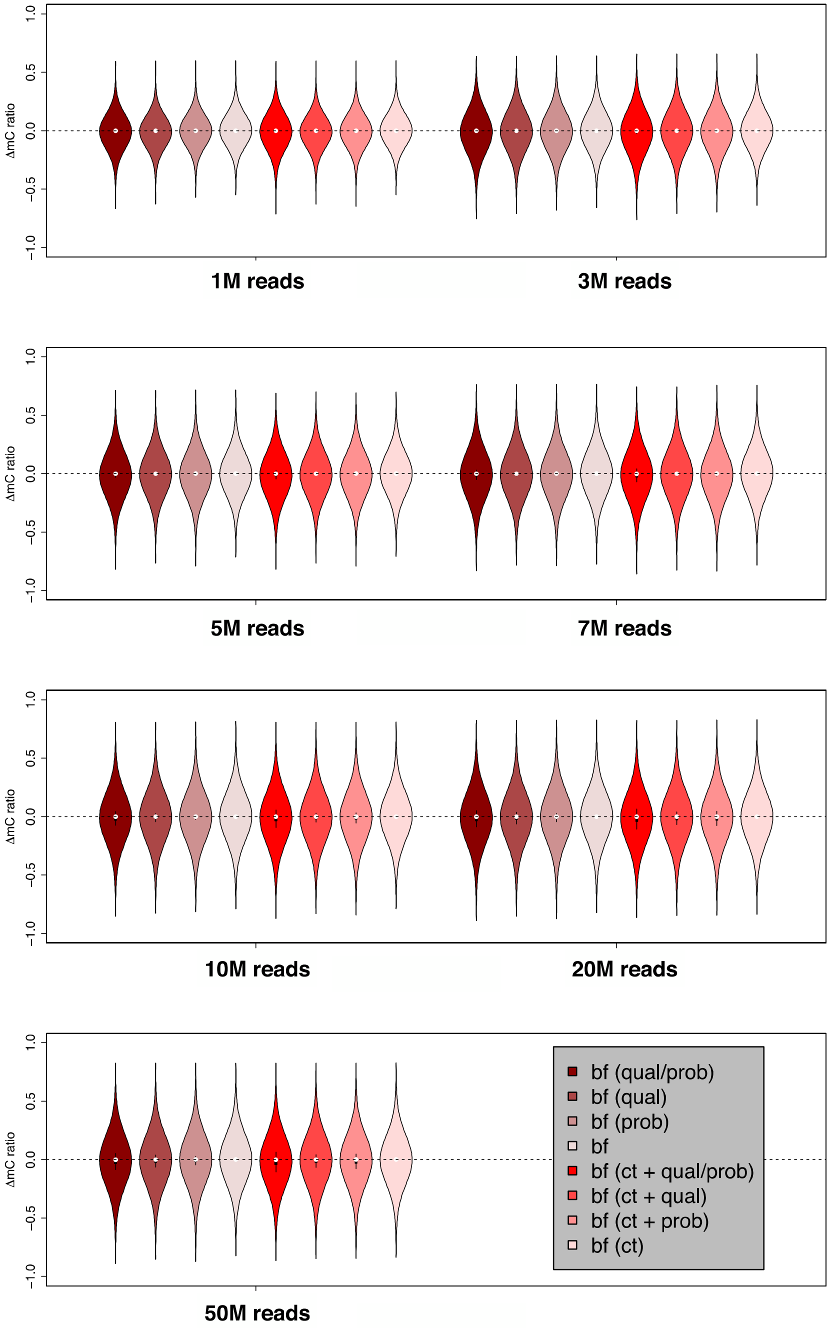
Supplementary Figure 1 (continued)

Dataset B



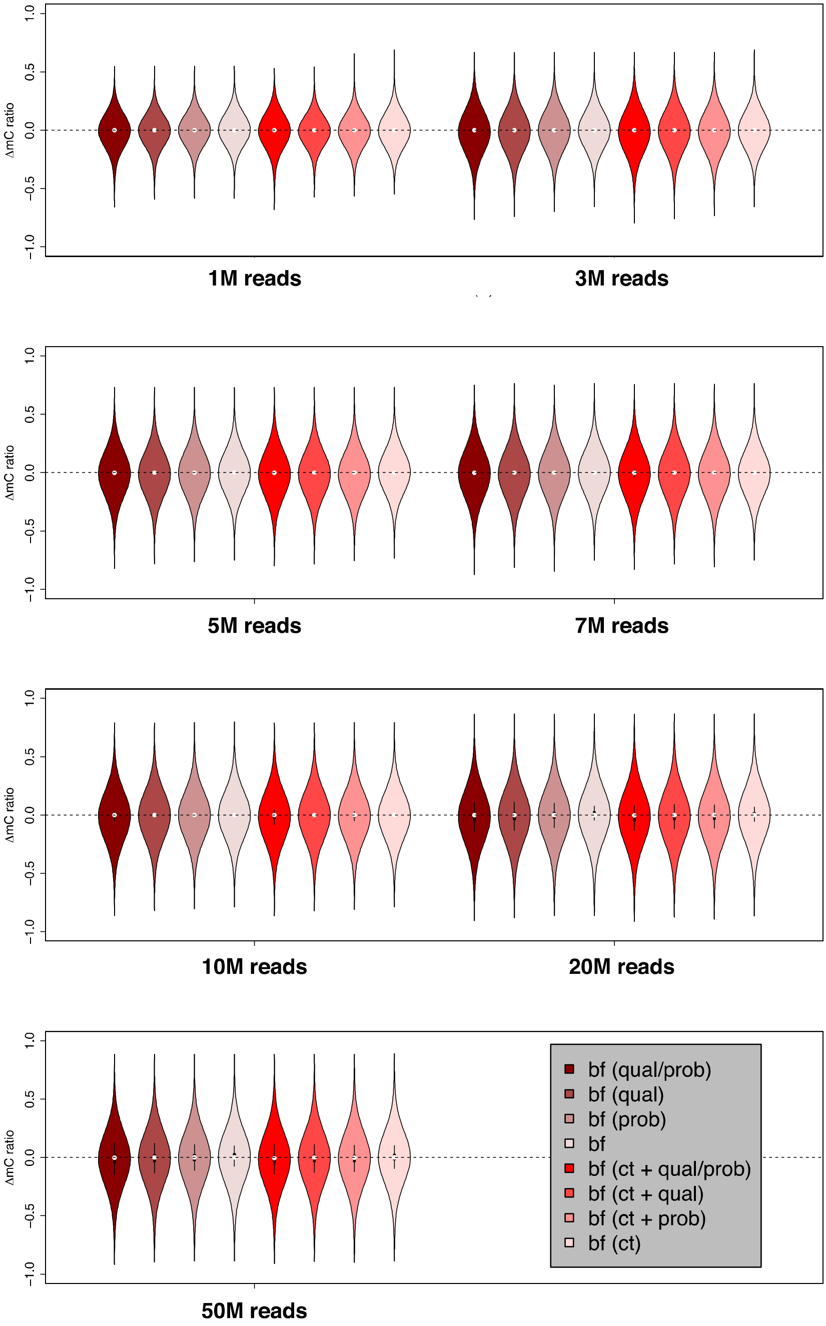
Supplementary Figure 1 (continued)

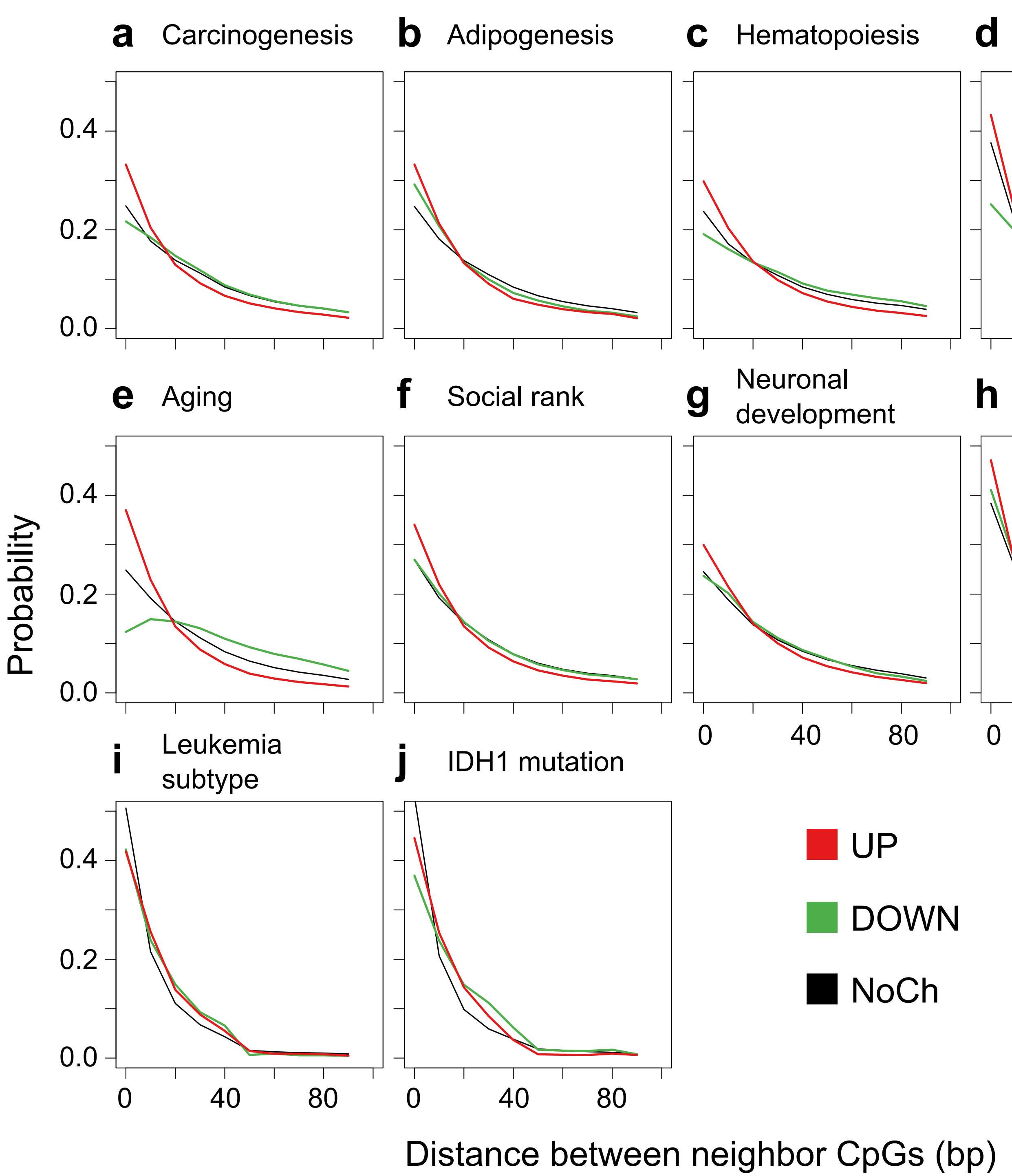
Dataset A



Supplementary Figure 1 (continued)

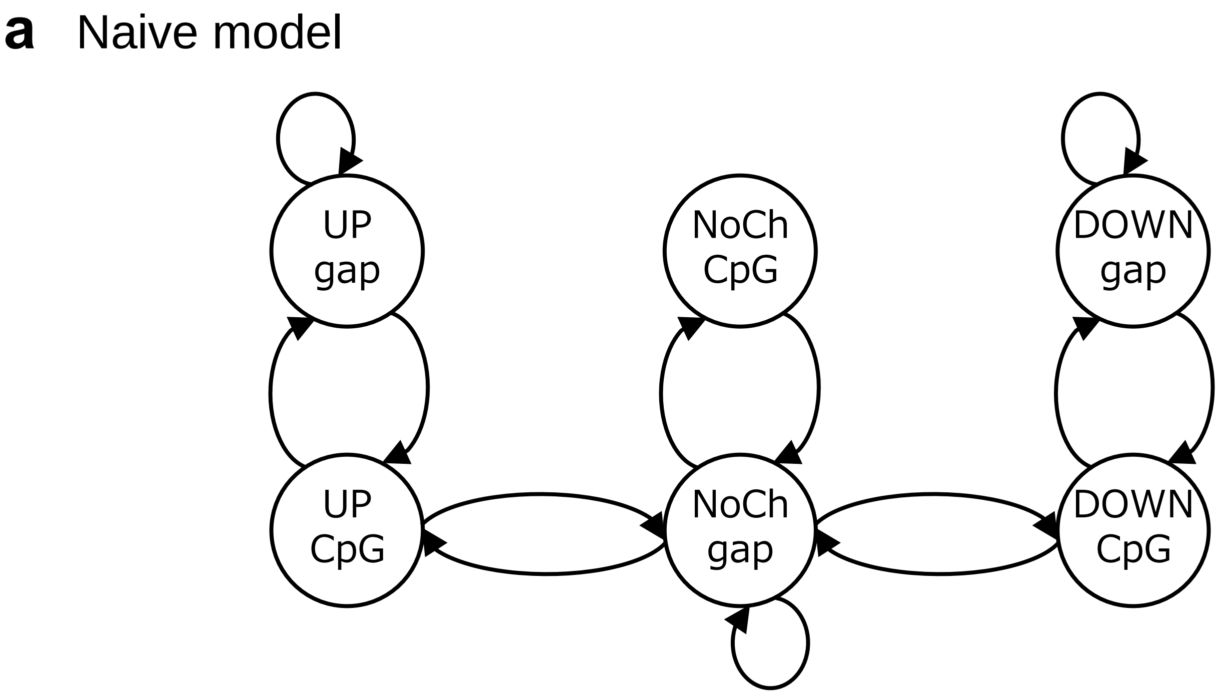
Dataset B

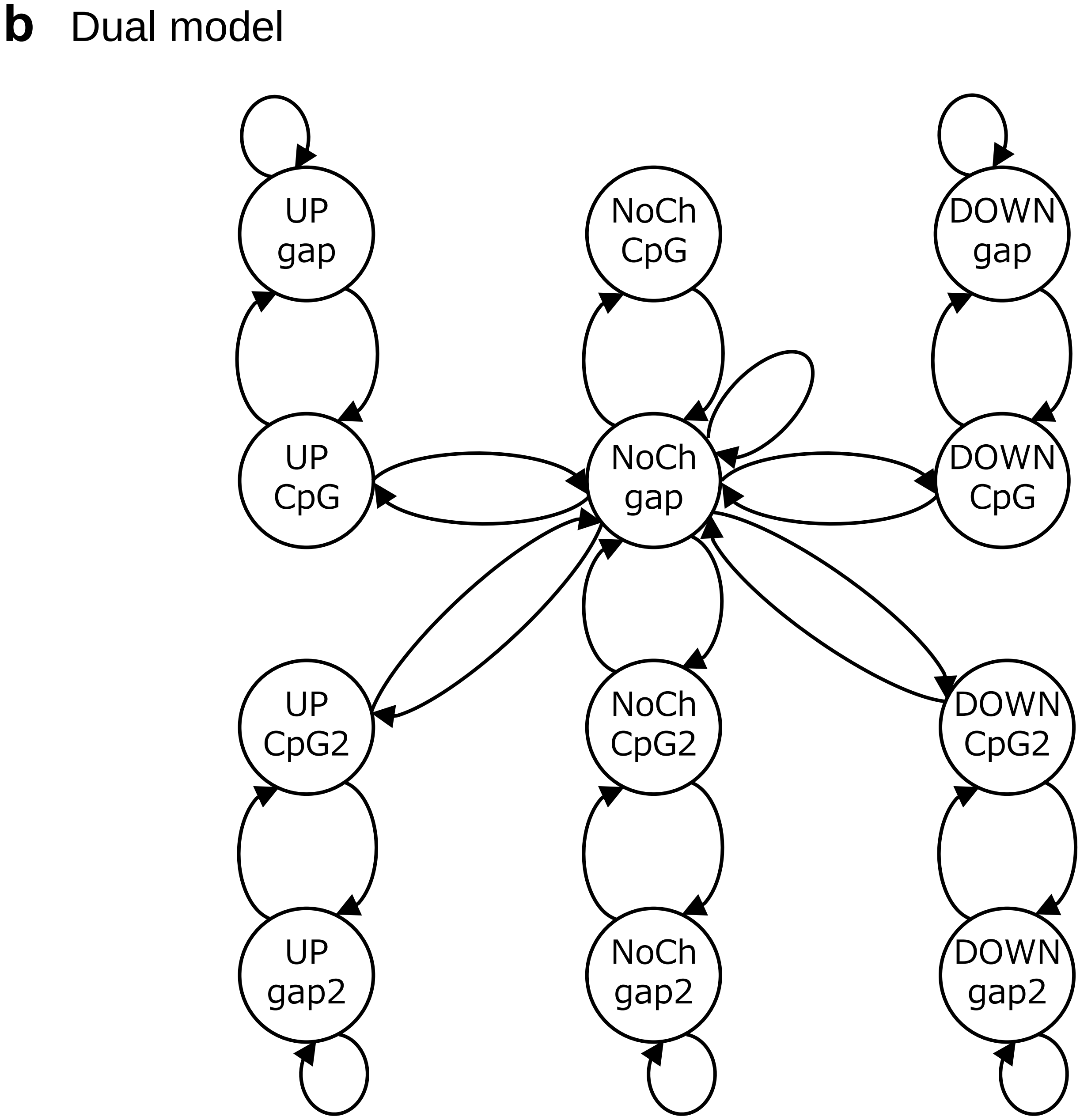


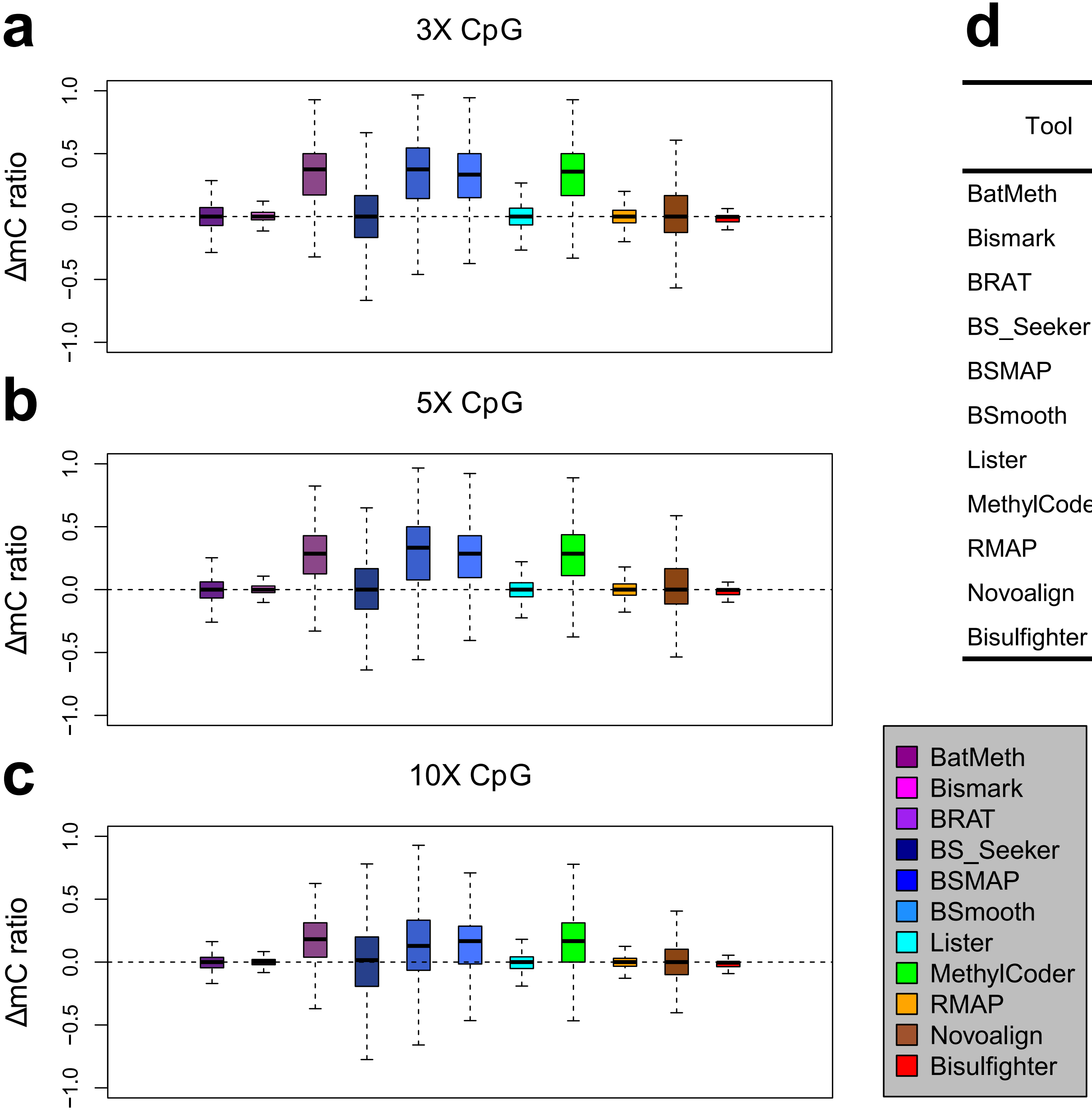


k Statistical significance

Dataset	*P*-value		
	UP NoCh	DOWN NoCh	UP DOWN
Carcinogenesis	<1e-15	<1e-15	<1e-15
Adipogenesis	<1e-15	<1e-15	<1e-15
Hematopoiesis	<1e-15	<1e-15	<1e-15
Fibroblast development	<1e-15	<1e-15	<1e-15
Aging	<1e-15	<1e-15	<1e-15
Social rank	<1e-15	6e-15	<1e-15
Neuronal development	<1e-15	7e-8	<1e-15
Tet1 mutation	<1e-15	<1e-15	<1e-15
Leukemia subtype	<1e-15	<1e-15	0.5
IDH1 mutation	<1e-15	<1e-15	3e-13


a
Naive model
UP gap
NoCh CpG
DOWN gap
UP CpG
NoCh gap
DOWN CpG


b
Dual model
UP gap
NoCh CpG
DOWN gap
UP CpG
NoCh gap
DOWN CpG
UP CpG2
NoCh CpG2
DOWN CpG2
UP gap2
NoCh gap2
DOWN gap2

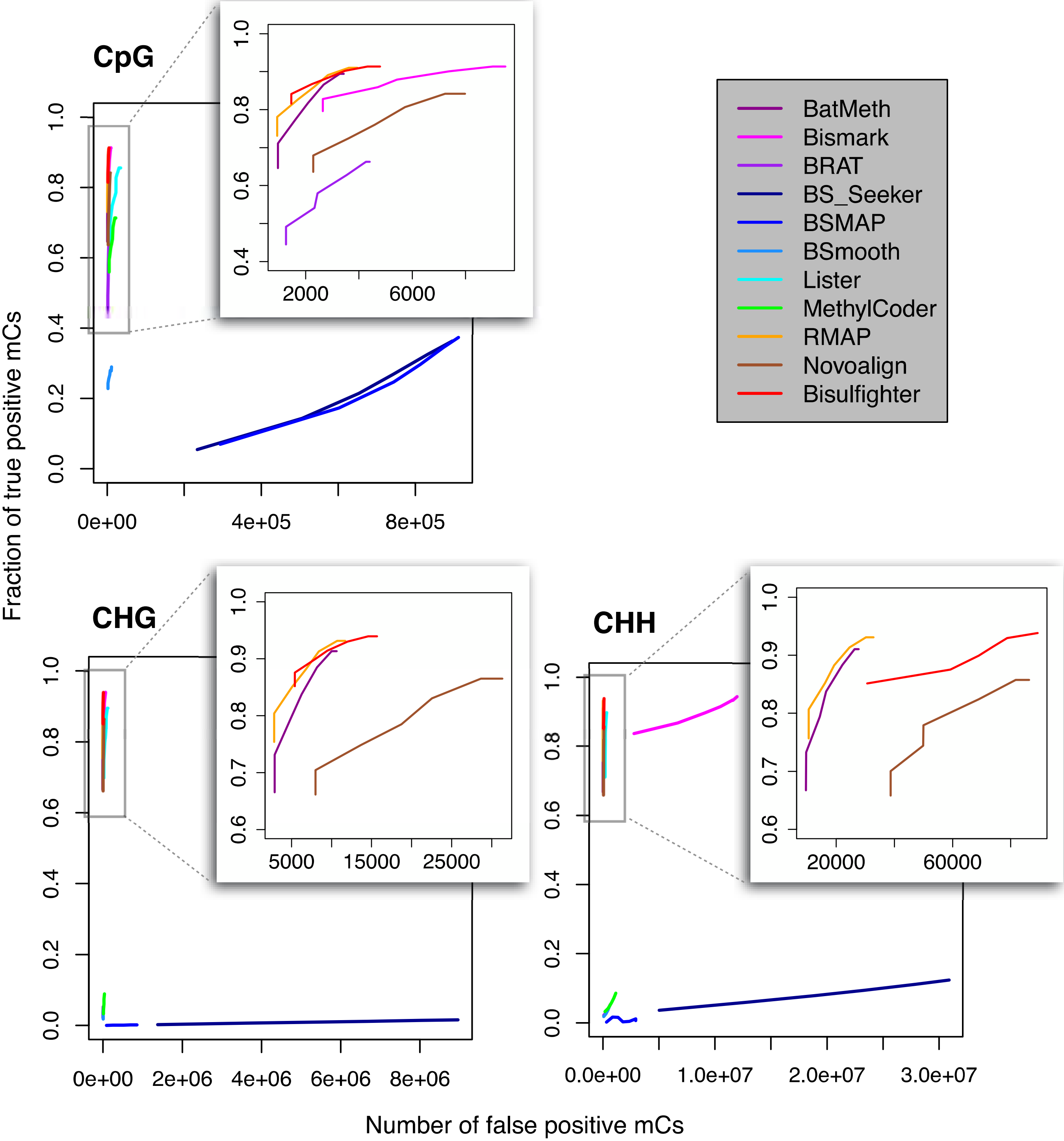


d

Tool	Number of captured CpGs		
	3X	5X	10X
BatMeth	2012940	1350373	114417
Bismark	2129912	1684724	294281
BRAT	913548	259577	1043
BS_Seeker	744921	352094	10561
BSMAP	598847	221746	6387
BSmooth	493924	192207	1761
Lister	1920353	1370262	205598
MethylCoder	1181911	461315	6085
RMAP	2102626	1577052	216124
Novoalign	1871710	1224222	140024
Bisulfighter	2130895	1700999	314585

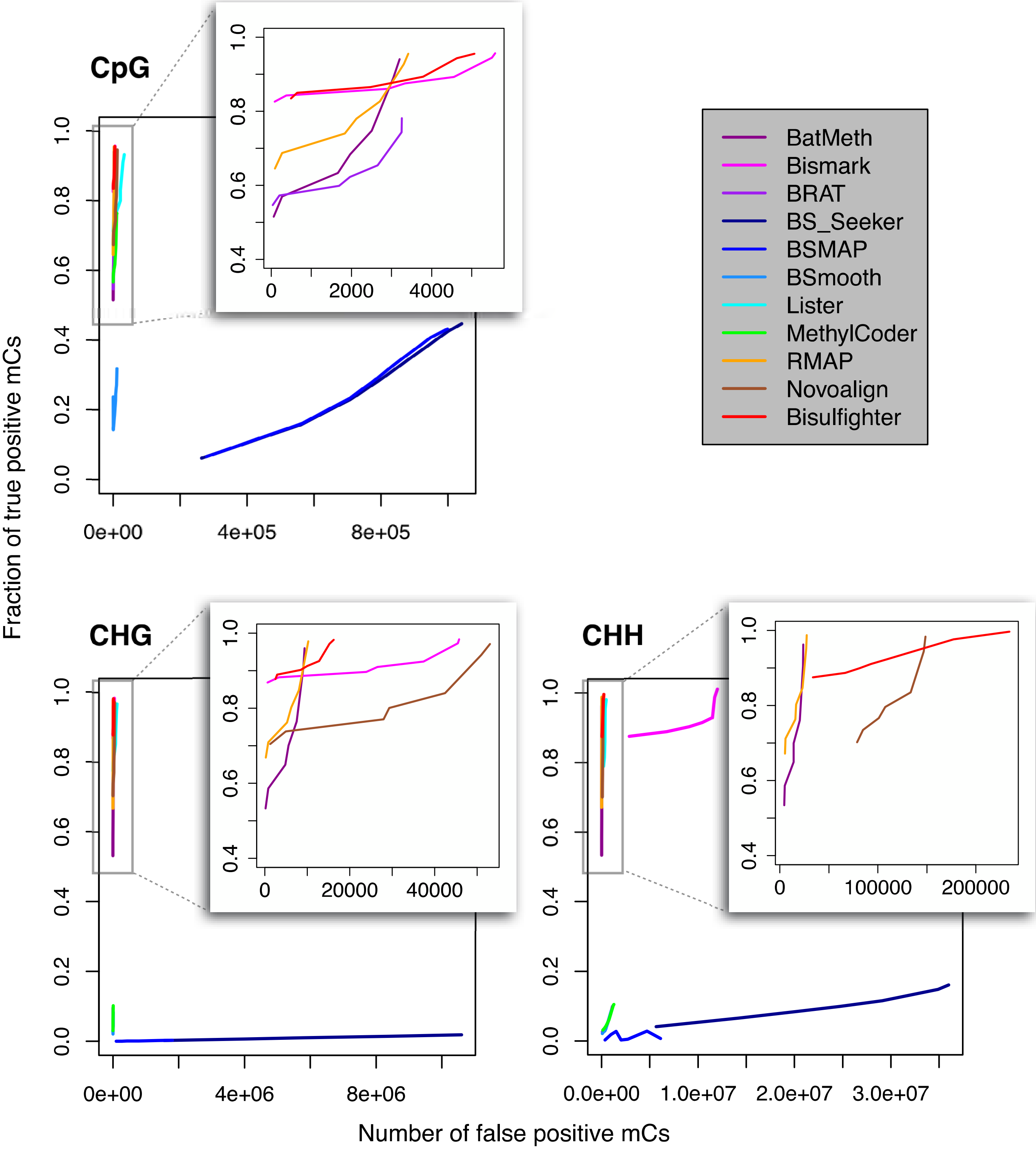
Supplementary Figure 5

Dataset A



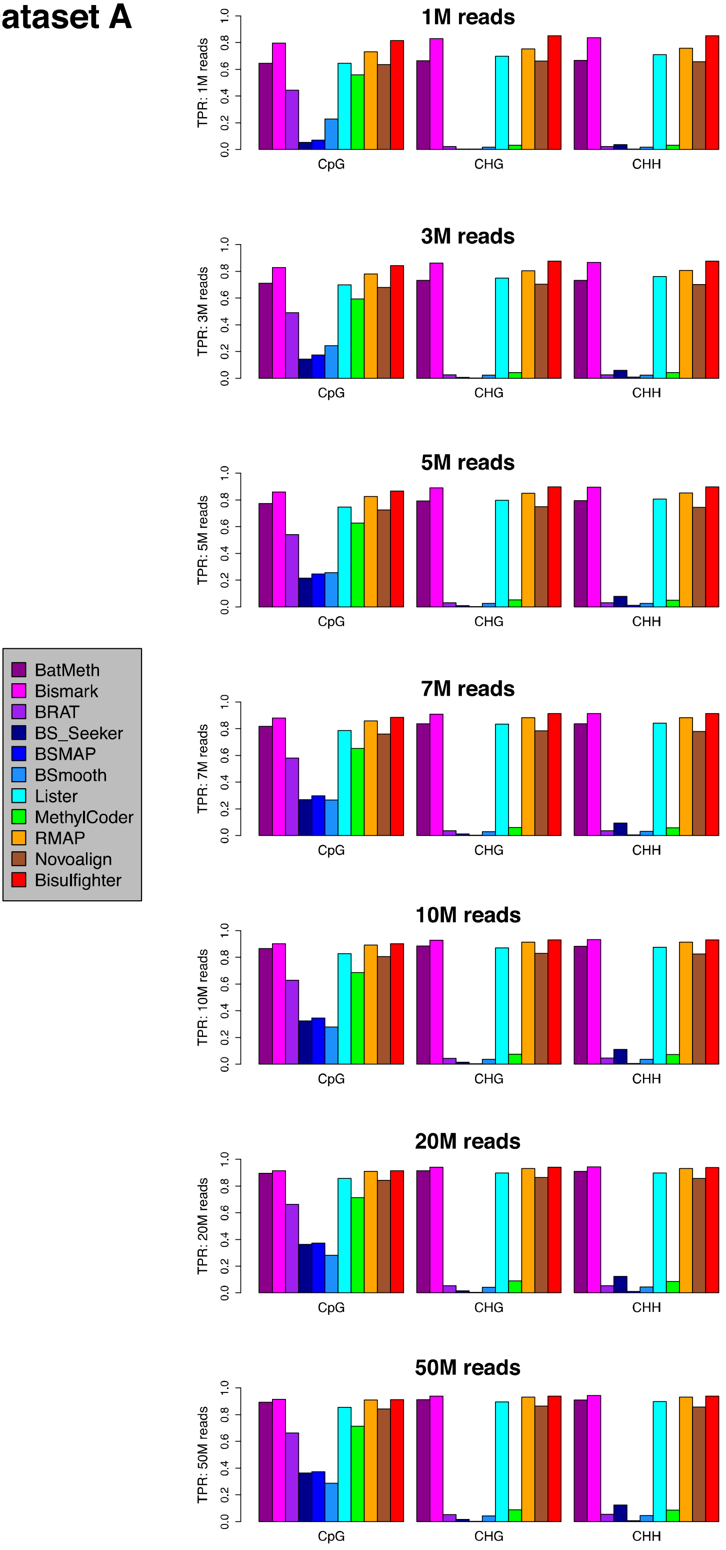
Supplementary Figure 5 (continued)

Dataset B



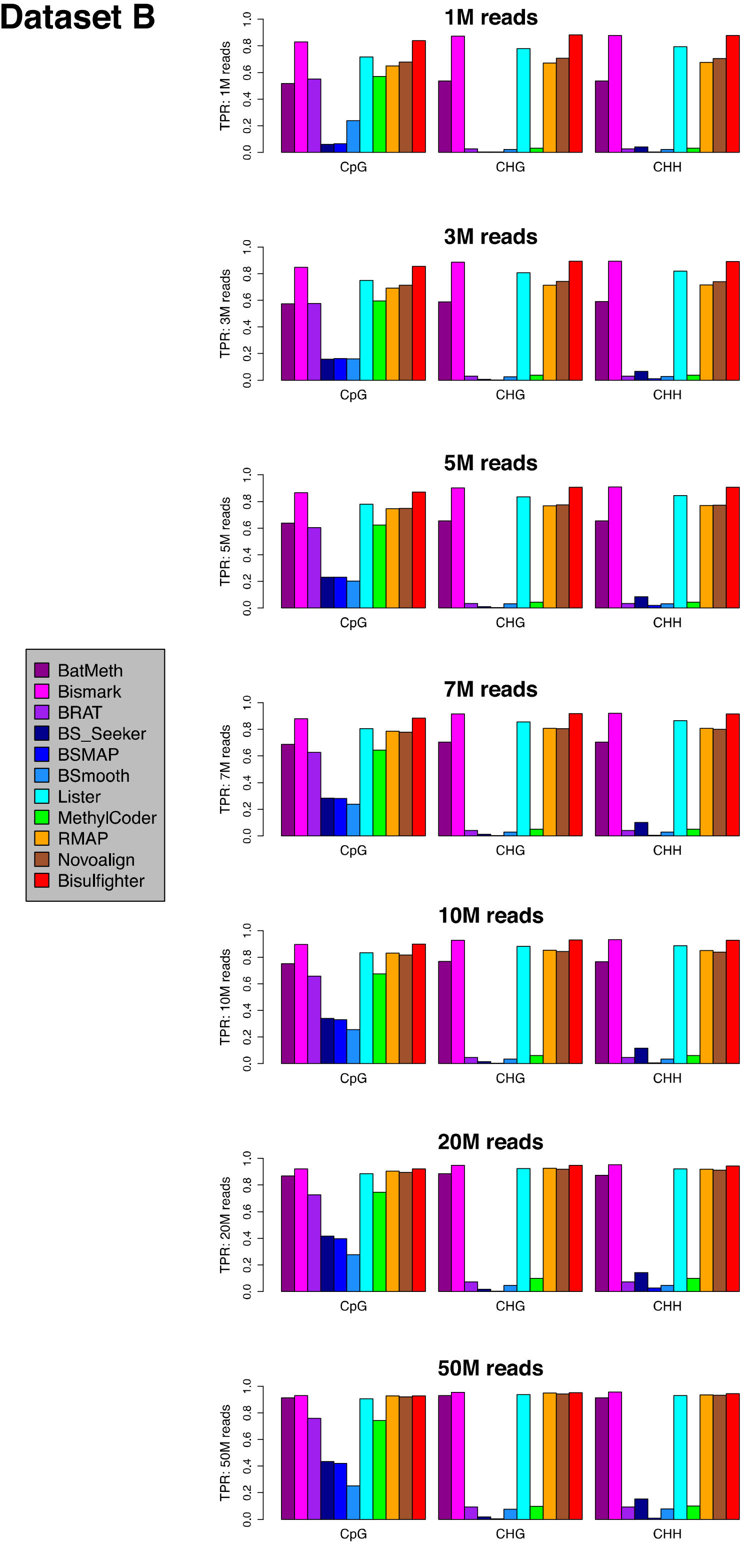
Supplementary Figure 6

Dataset A



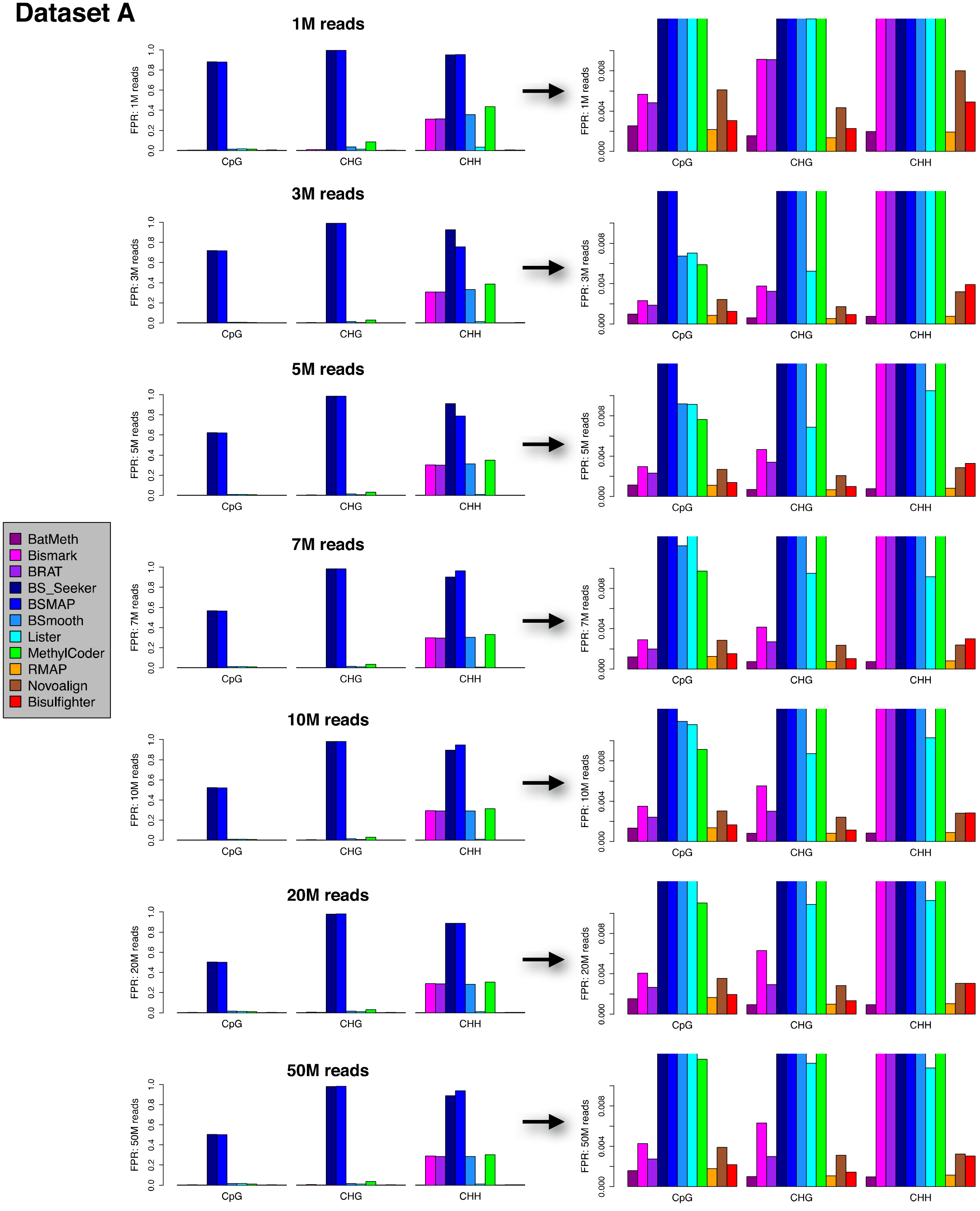
Supplementary Figure 6 (continued)

Dataset B



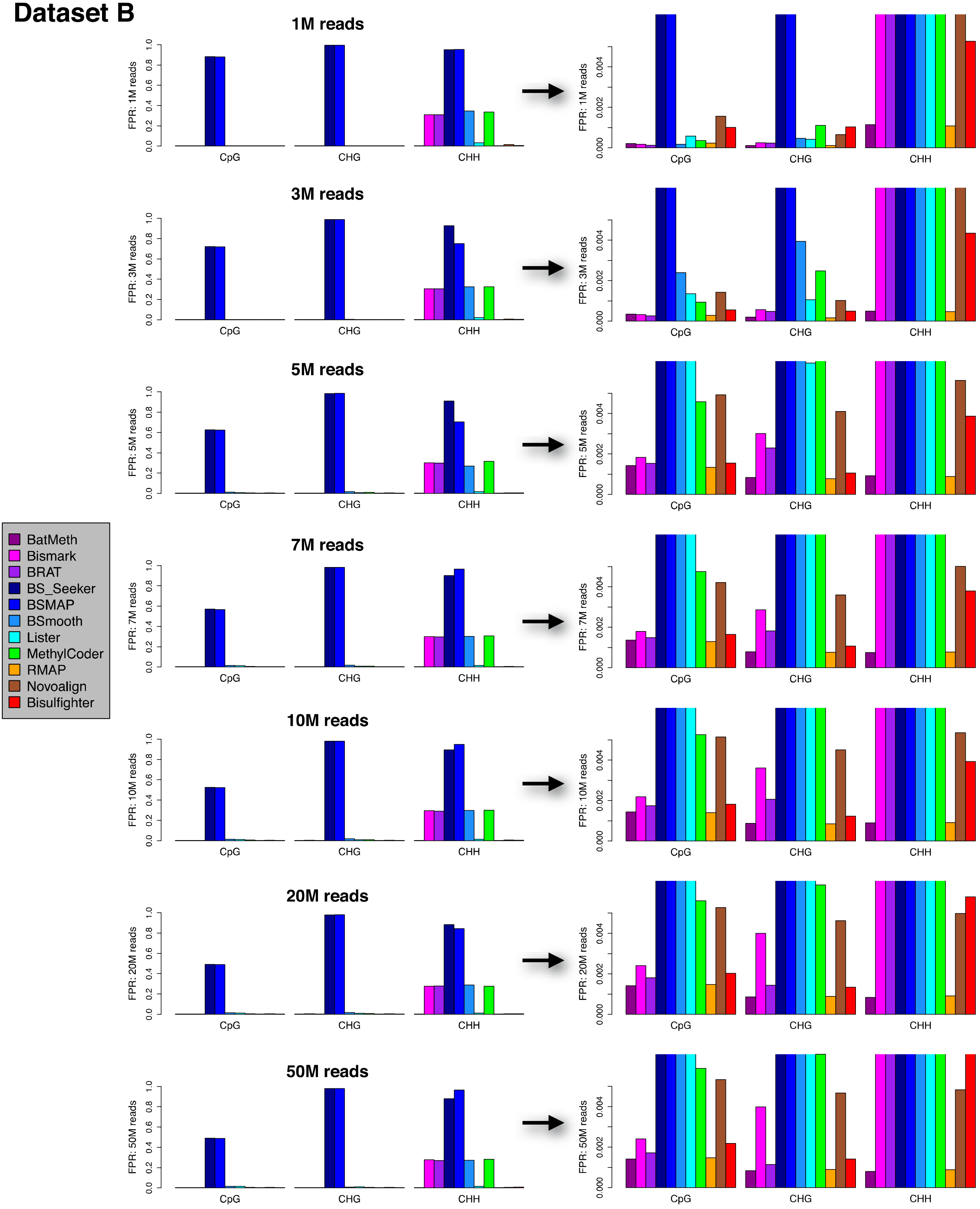
Supplementary Figure 6 (continued)

Dataset A



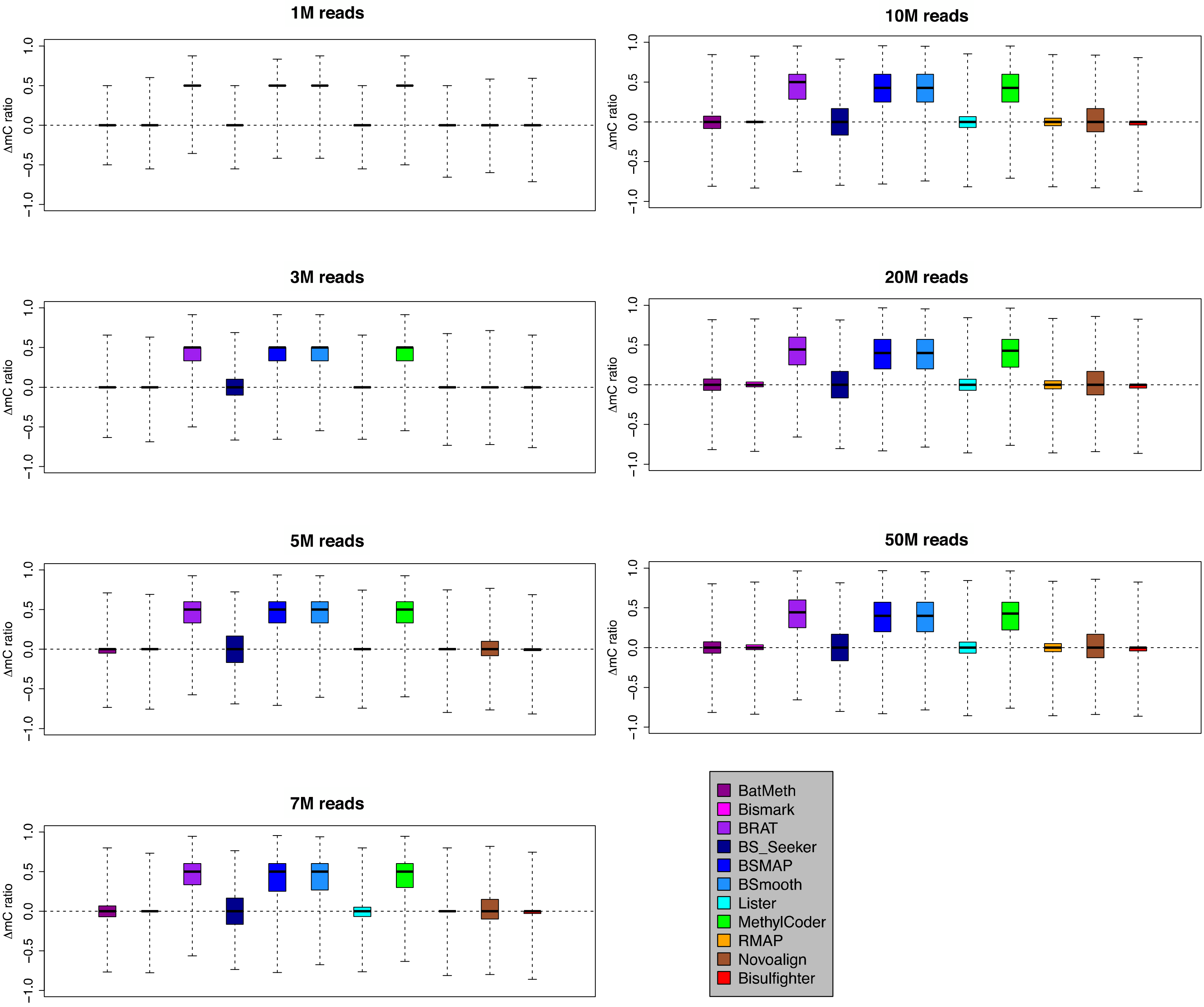
Supplementary Figure 6 (continued)

Dataset B



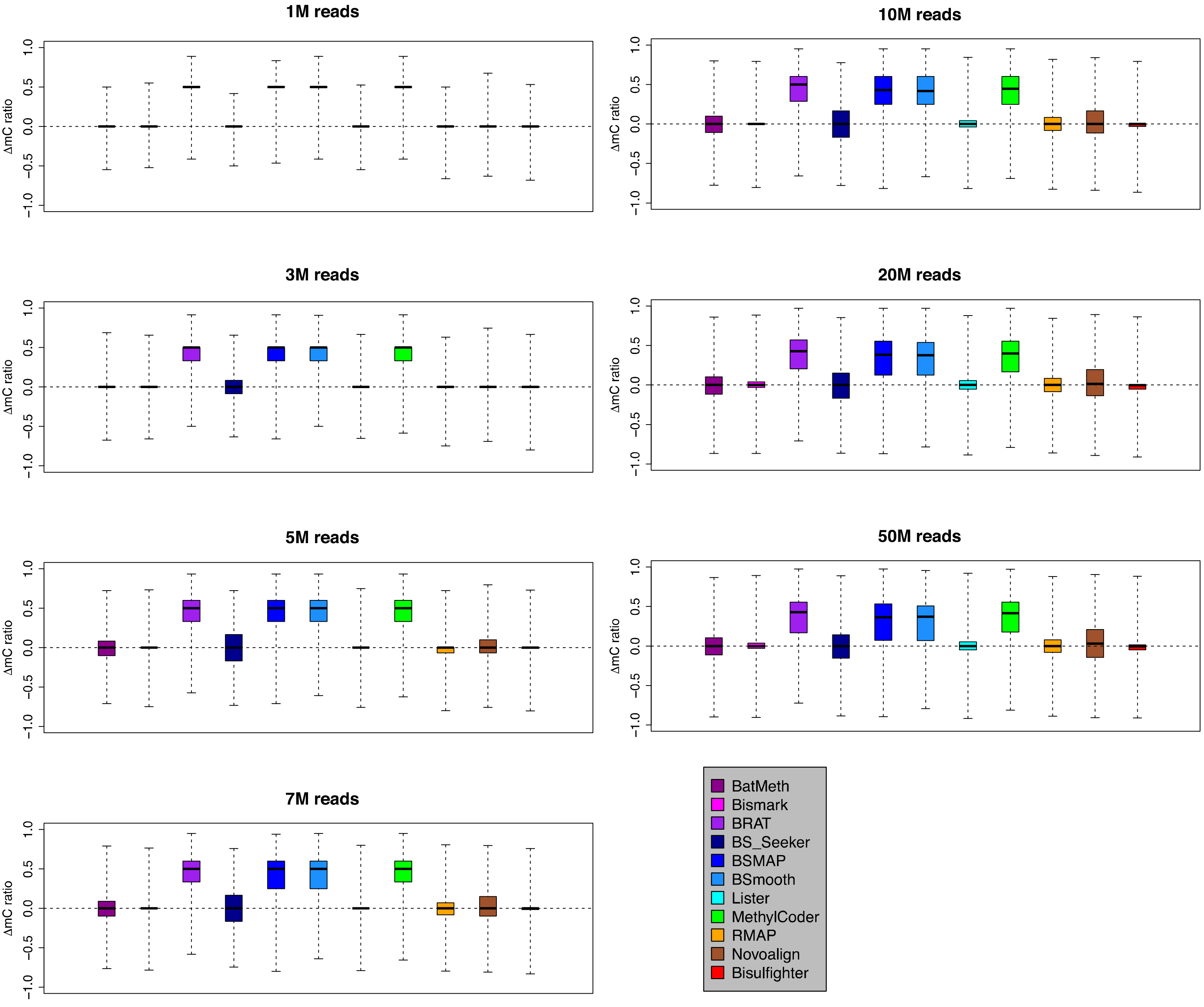
Supplementary Figure 7

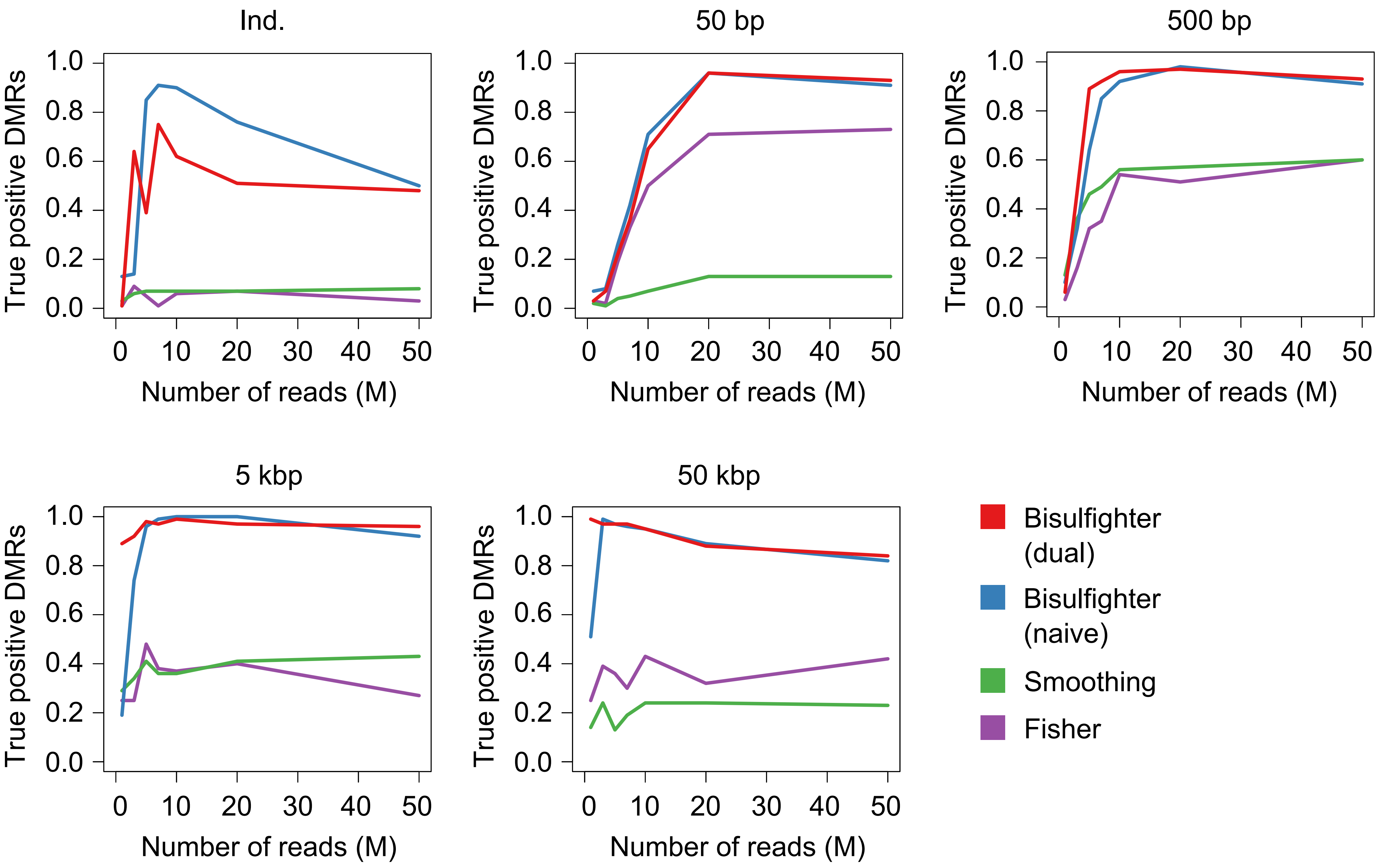
Dataset A

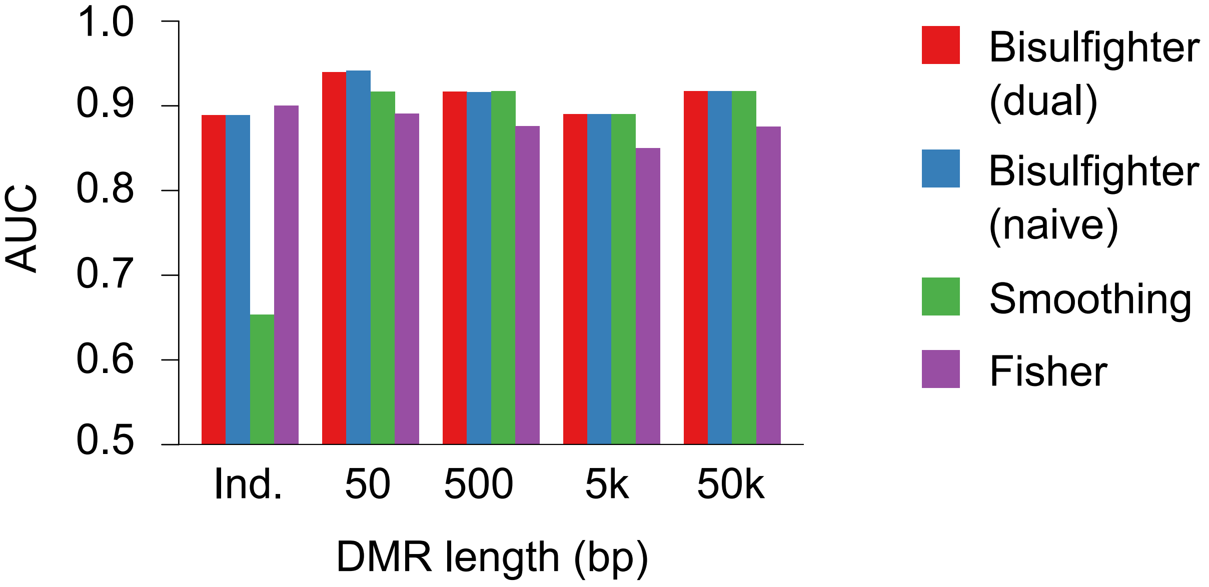


Supplementary Figure 7 (continued)

Dataset B

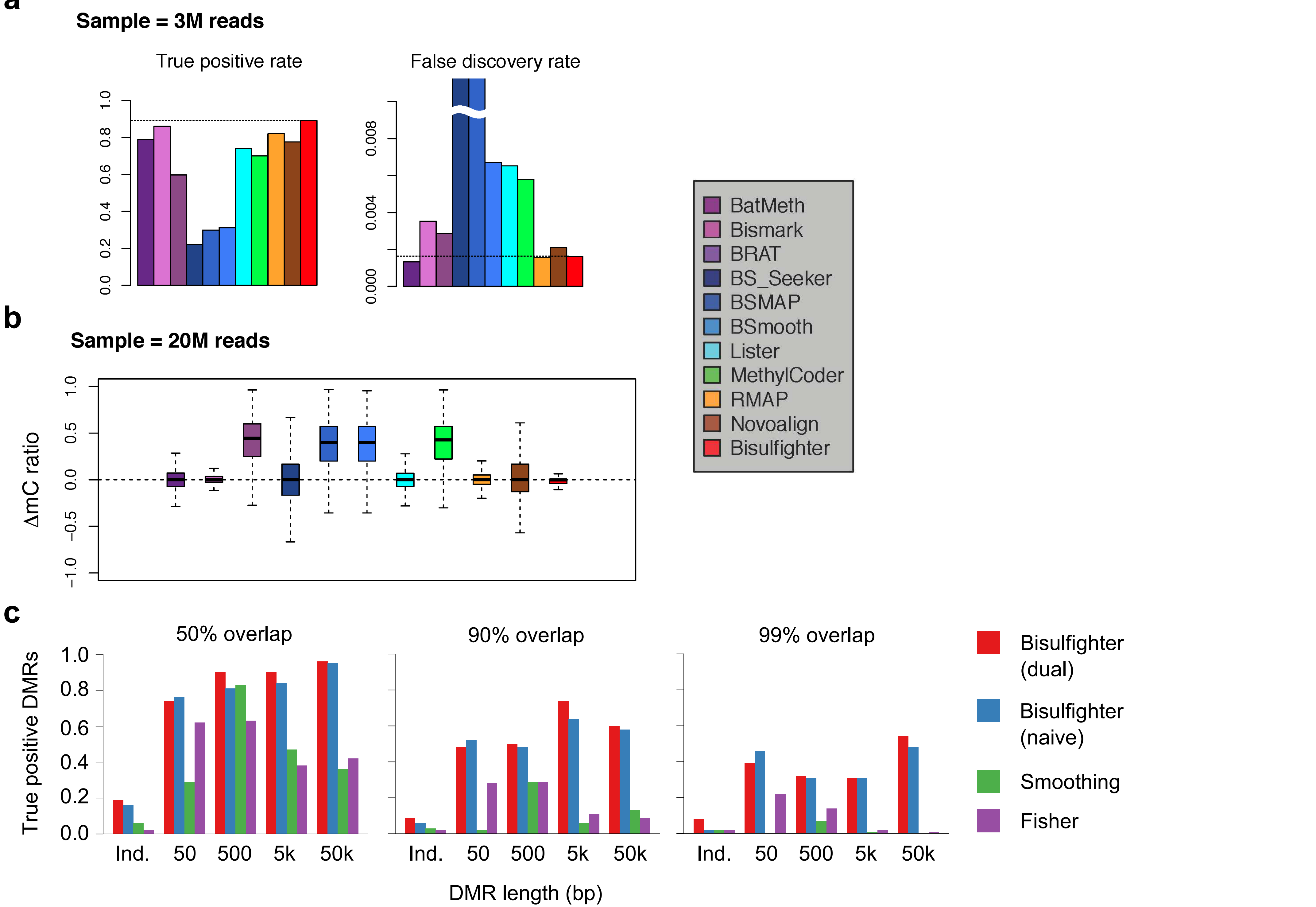



Ind.
50 bp
500 bp
5 kbp
50 kbp
True positive DMRs
Number of reads (M)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
🟥 Bisulfighter (dual)
🟦 Bisulfighter (naive)
🟩 Smoothing
🟪 Fisher

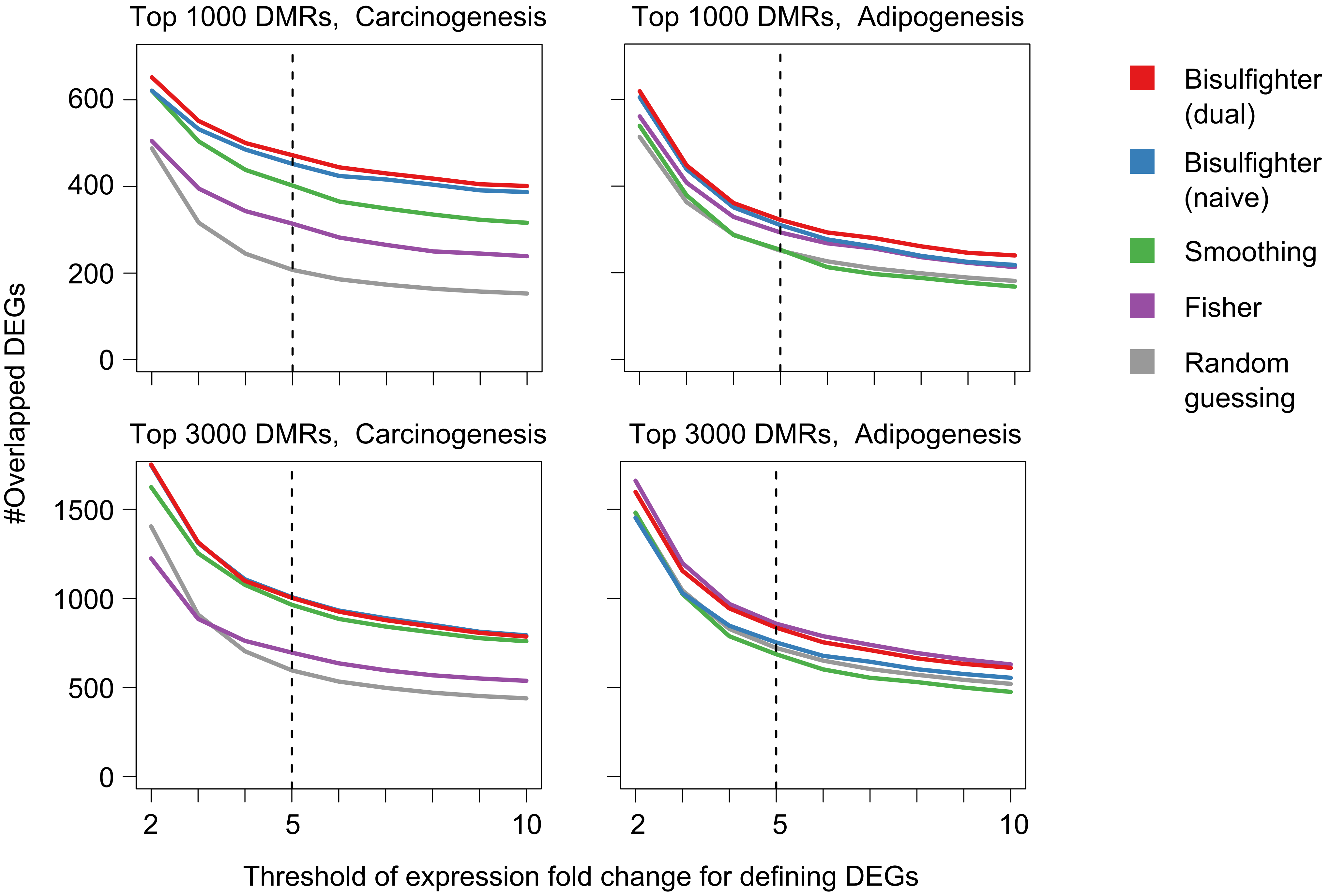


1.0
0.9
0.8
0.7
0.6
0.5
AUC
Ind.
50
500
5k
50k
DMR length (bp)
🟥 Bisulfighter (dual)
🟦 Bisulfighter (naive)
🟩 Smoothing
🟪 Fisher

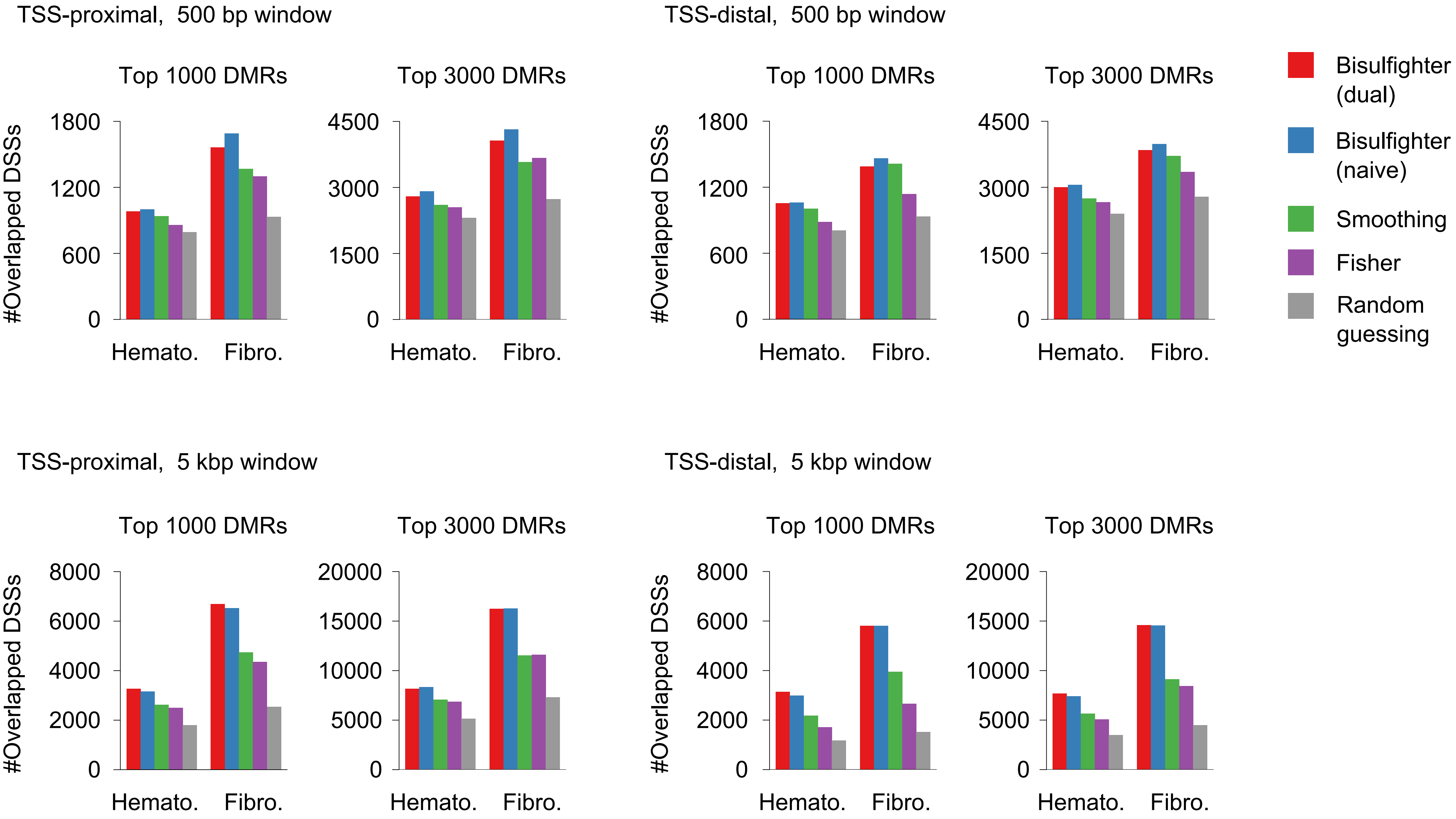
Supplementary Figure 10



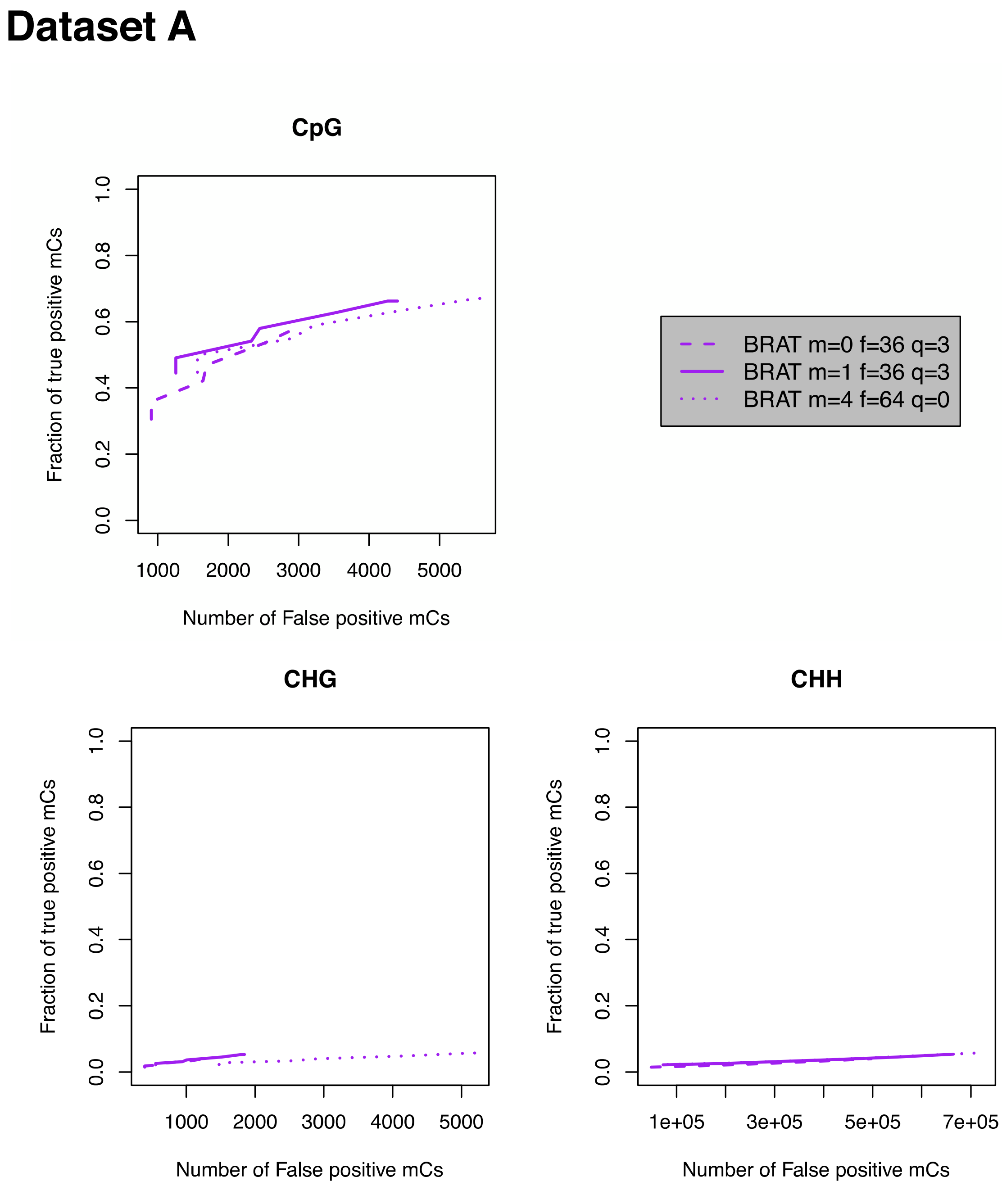
Supplementary Figure 11



Supplementary Figure 12

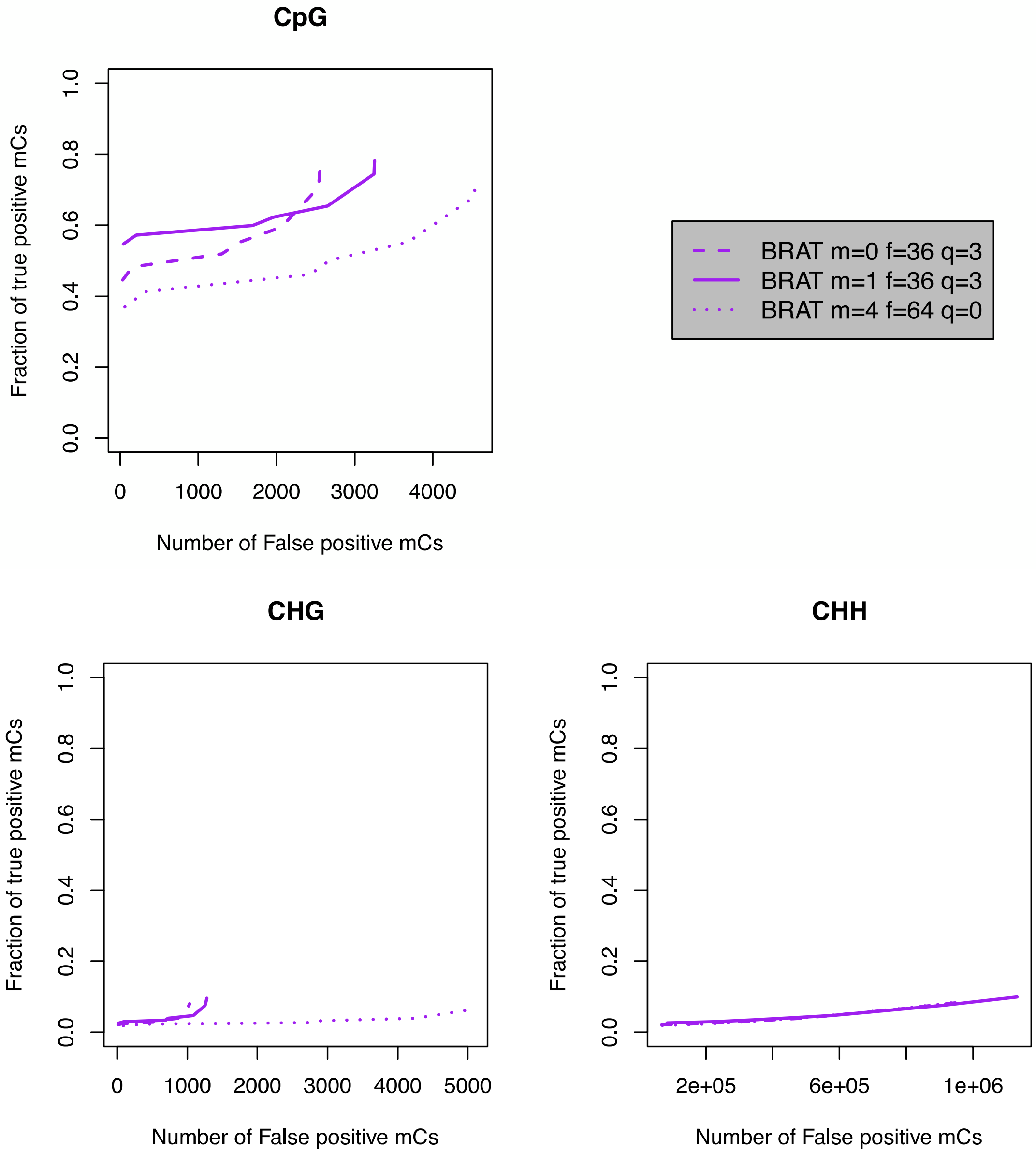


Supplementary Figure 13



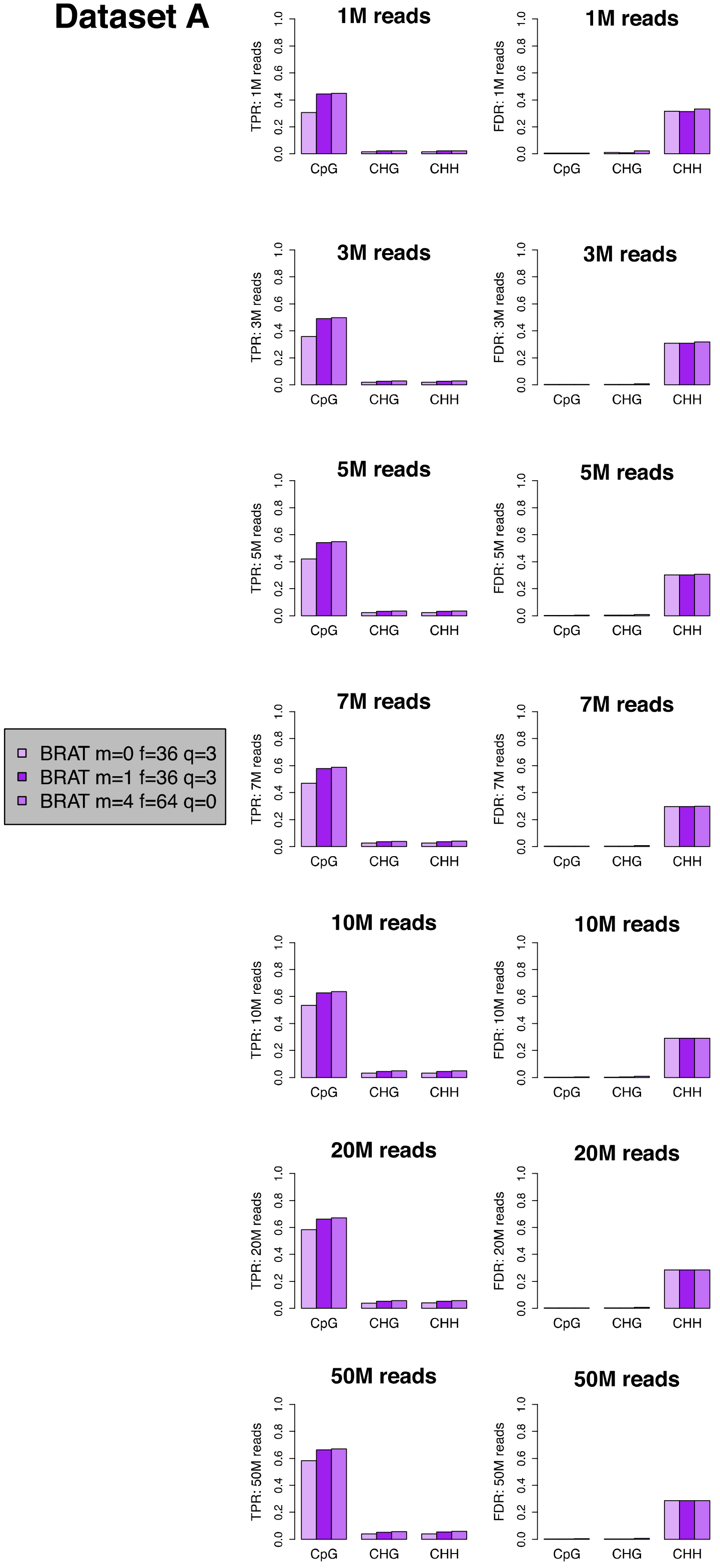
Supplementary Figure 13 (continued)

Dataset B



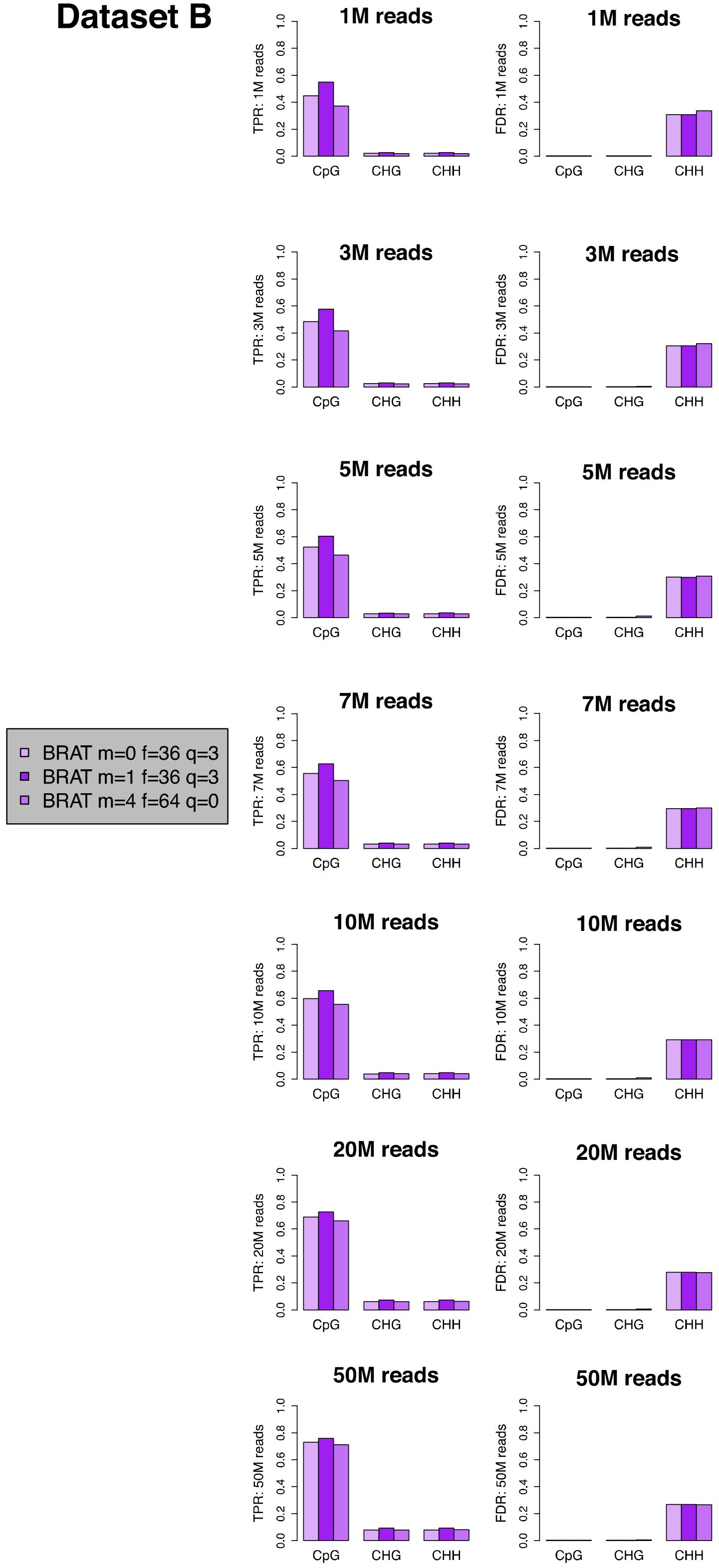
Supplementary Figure 13 (continued)

Dataset A

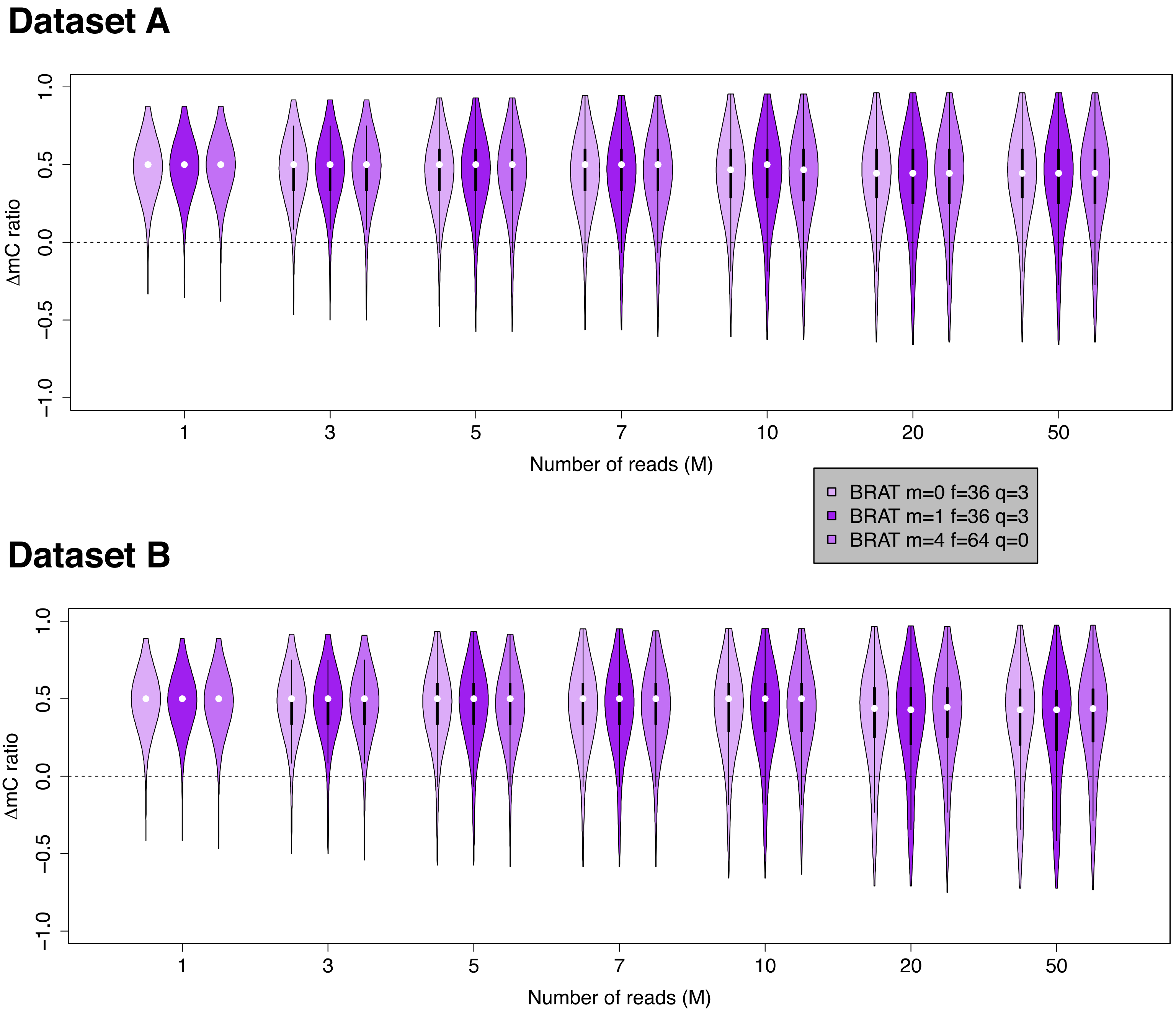


Supplementary Figure 13 (continued)

Dataset B



Supplementary Figure 13 (continued)

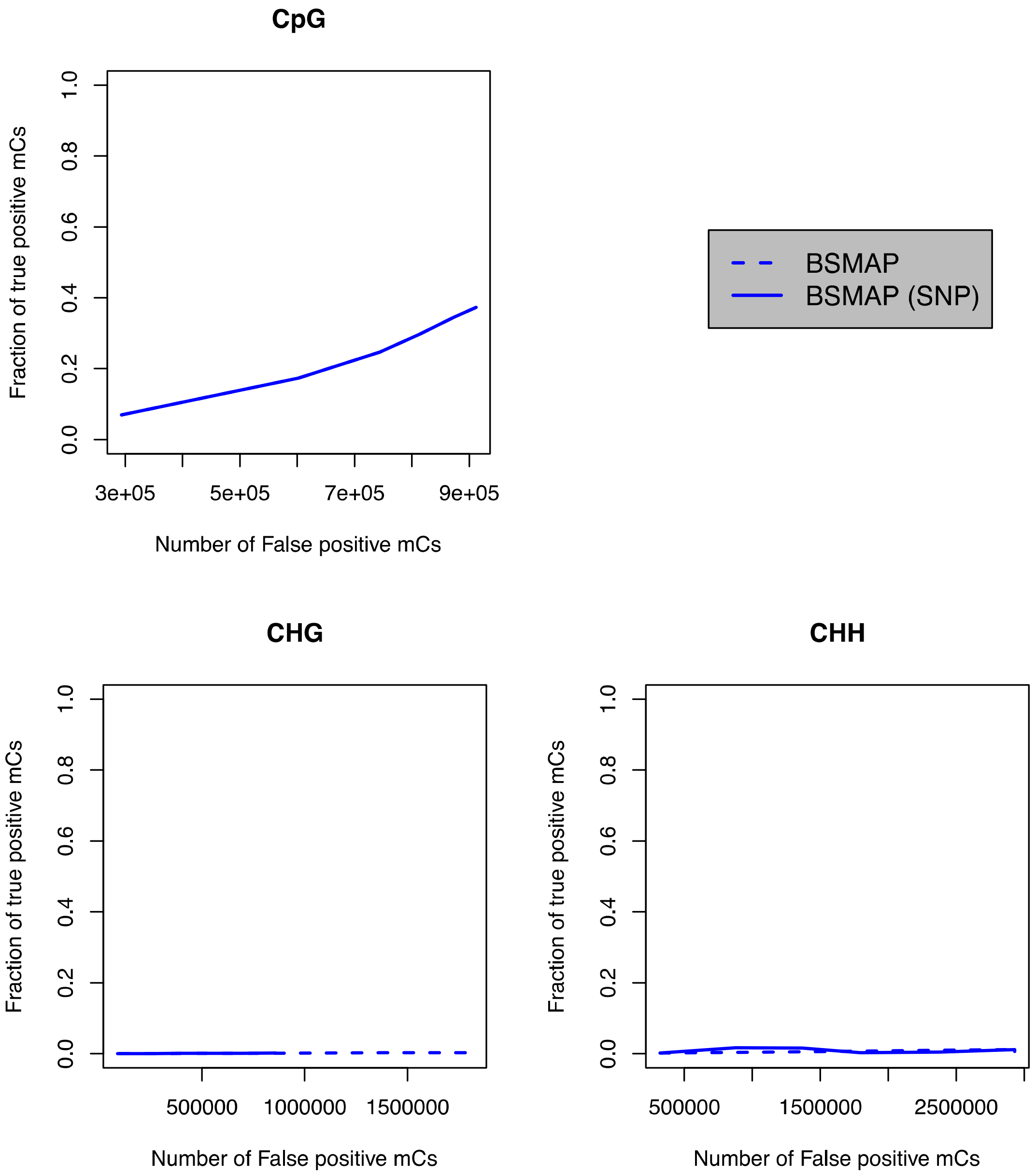


Dataset A

Dataset B

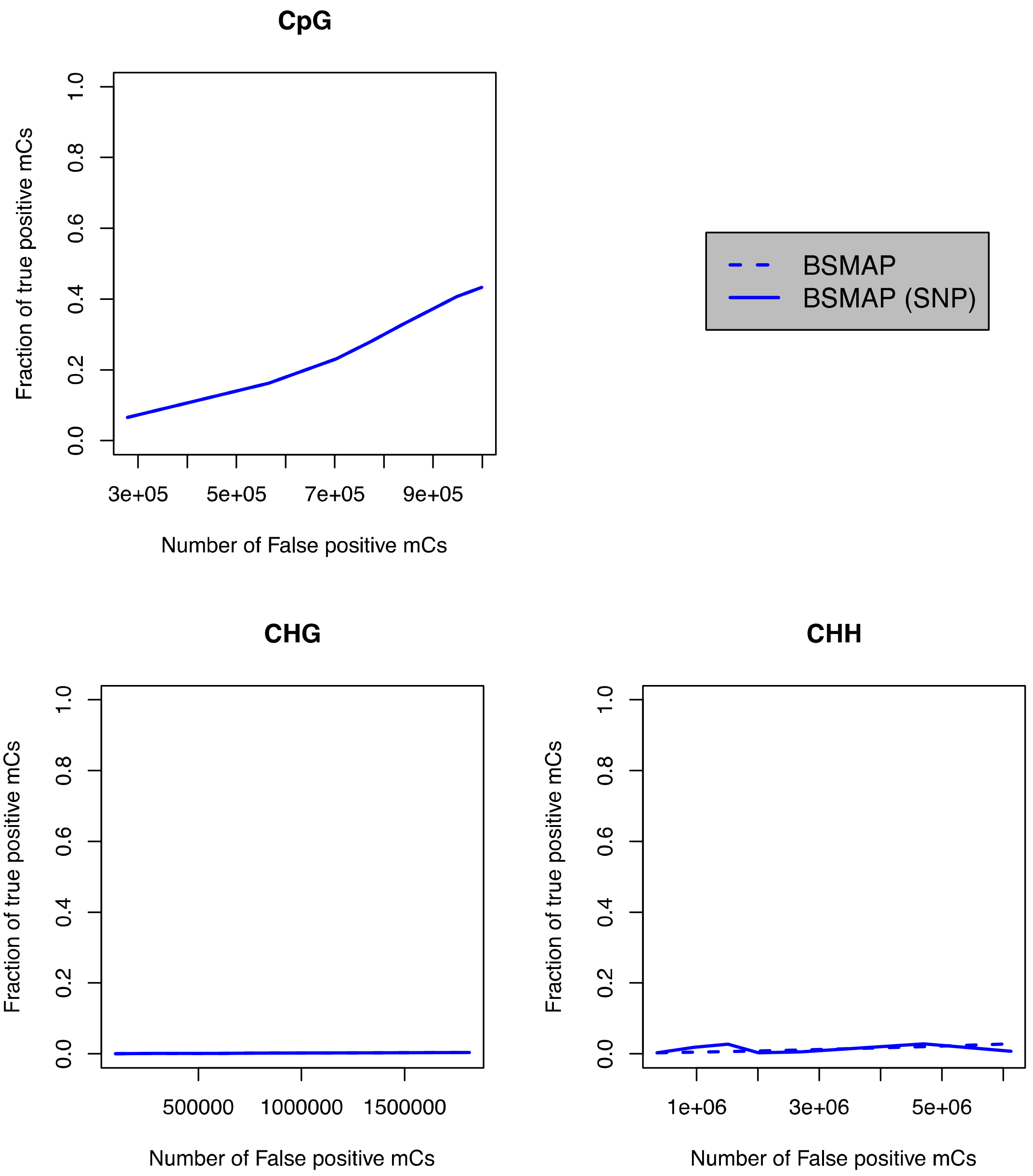
Supplementary Figure 14

Dataset A

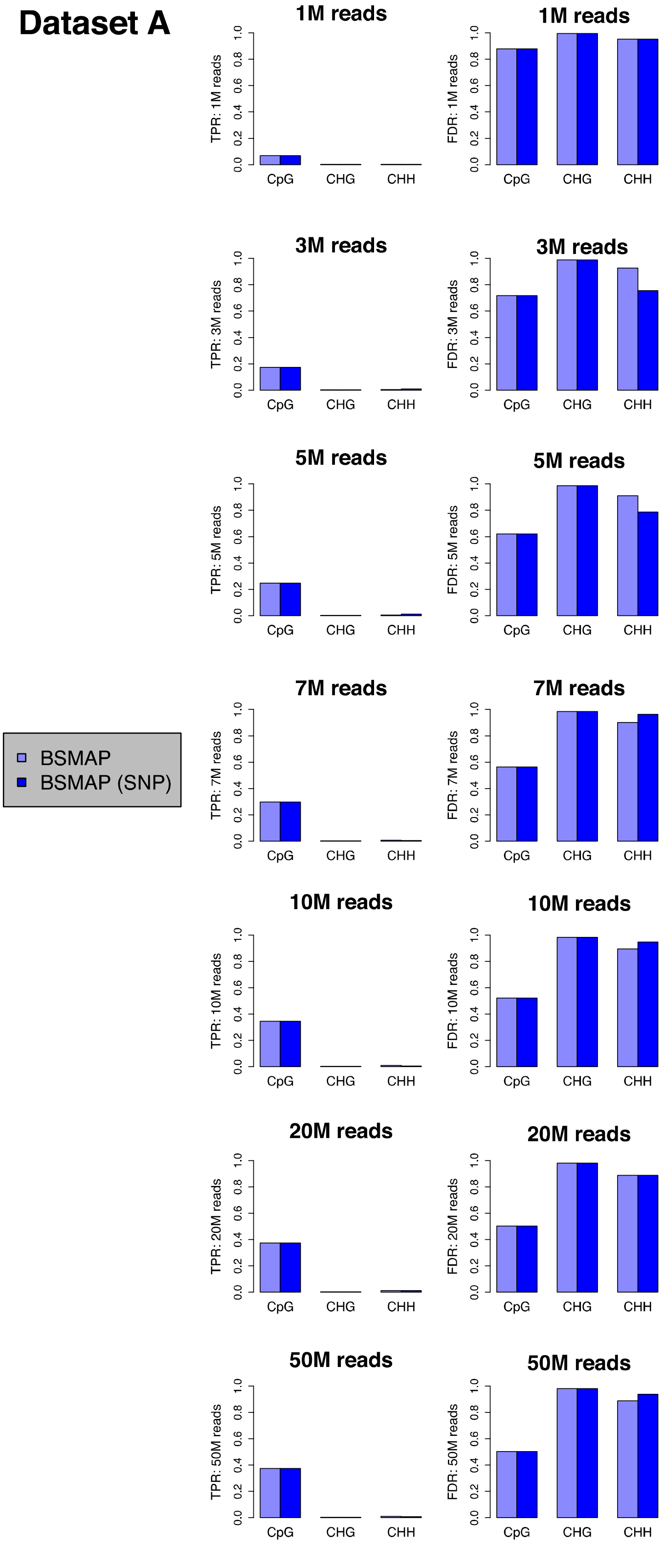


Supplementary Figure 14 (continued)

Dataset B

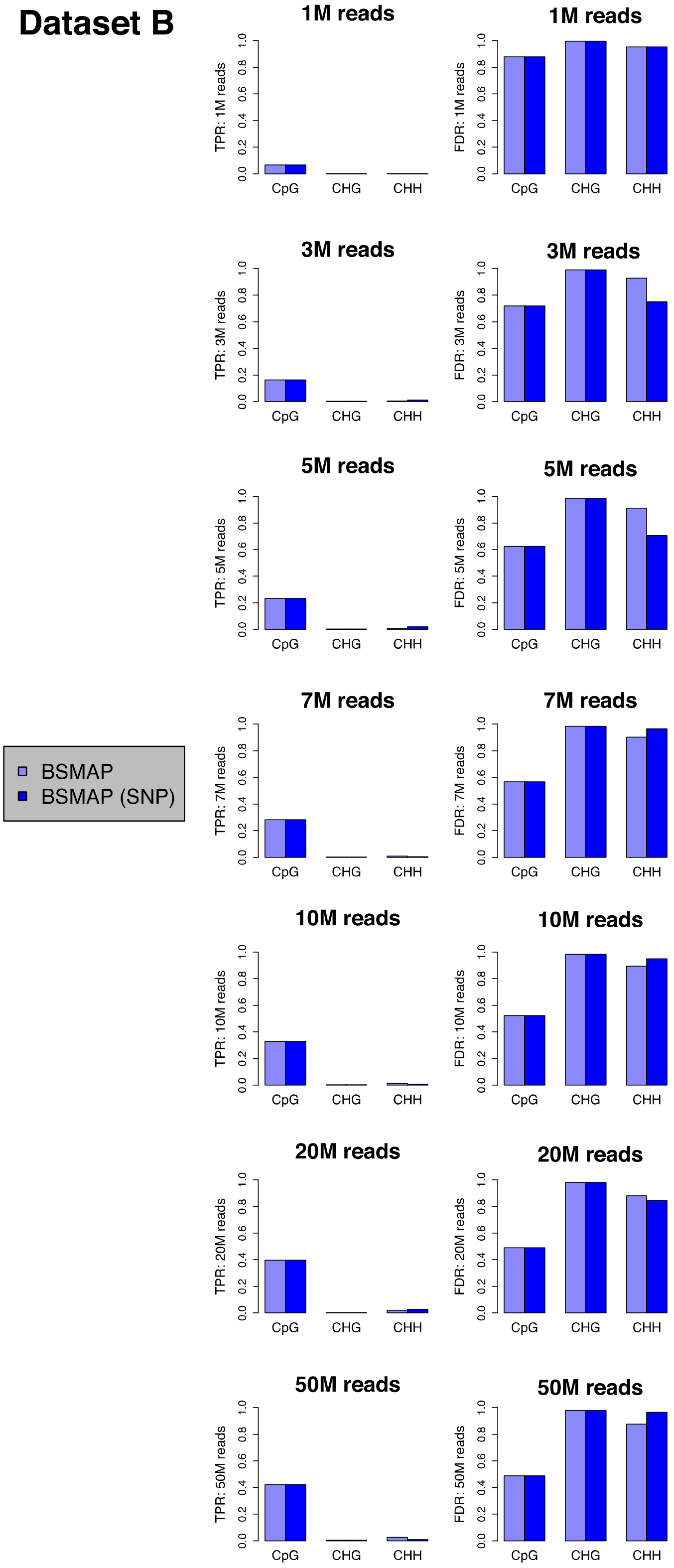


Supplementary Figure 14 (continued)

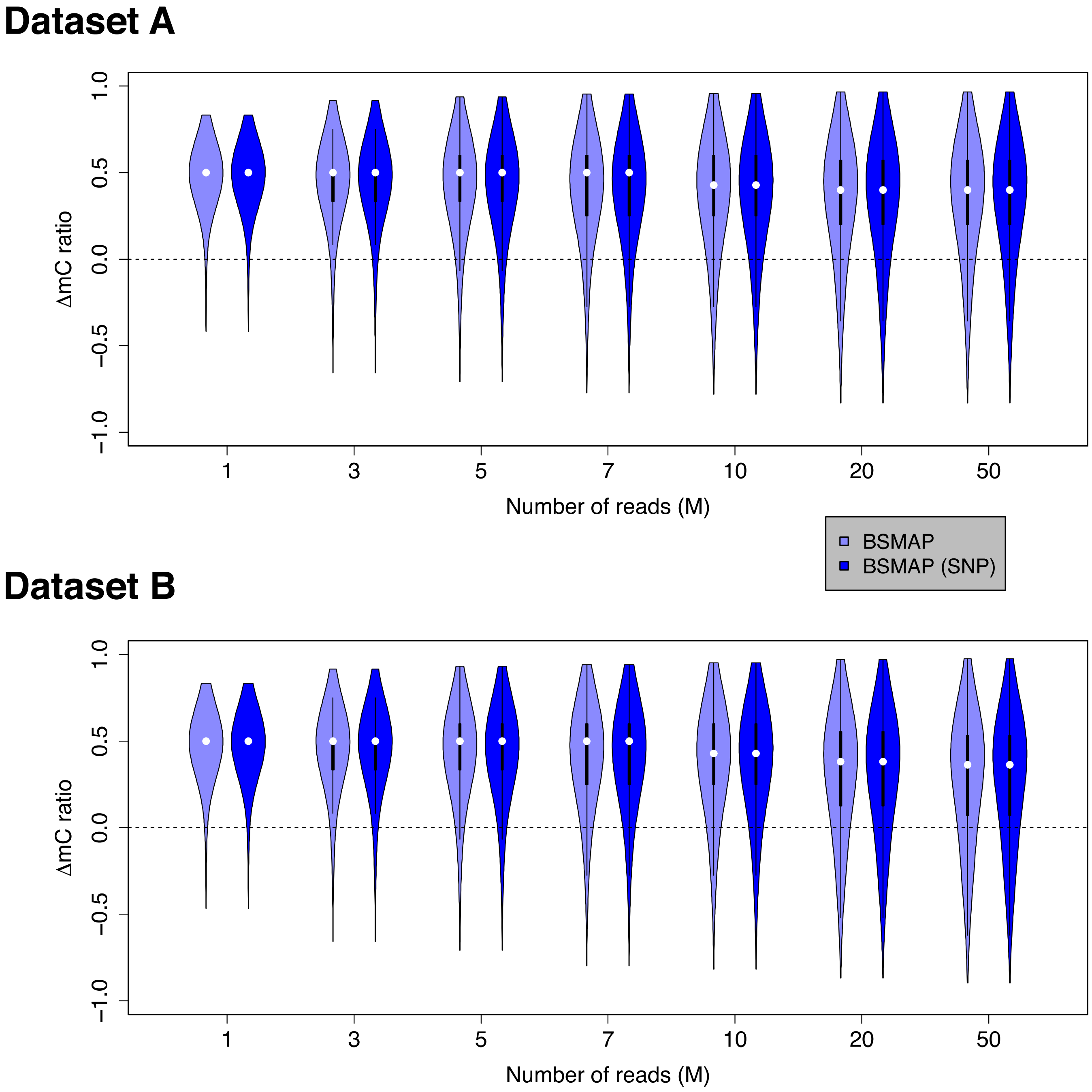


Supplementary Figure 14 (continued)

Dataset B

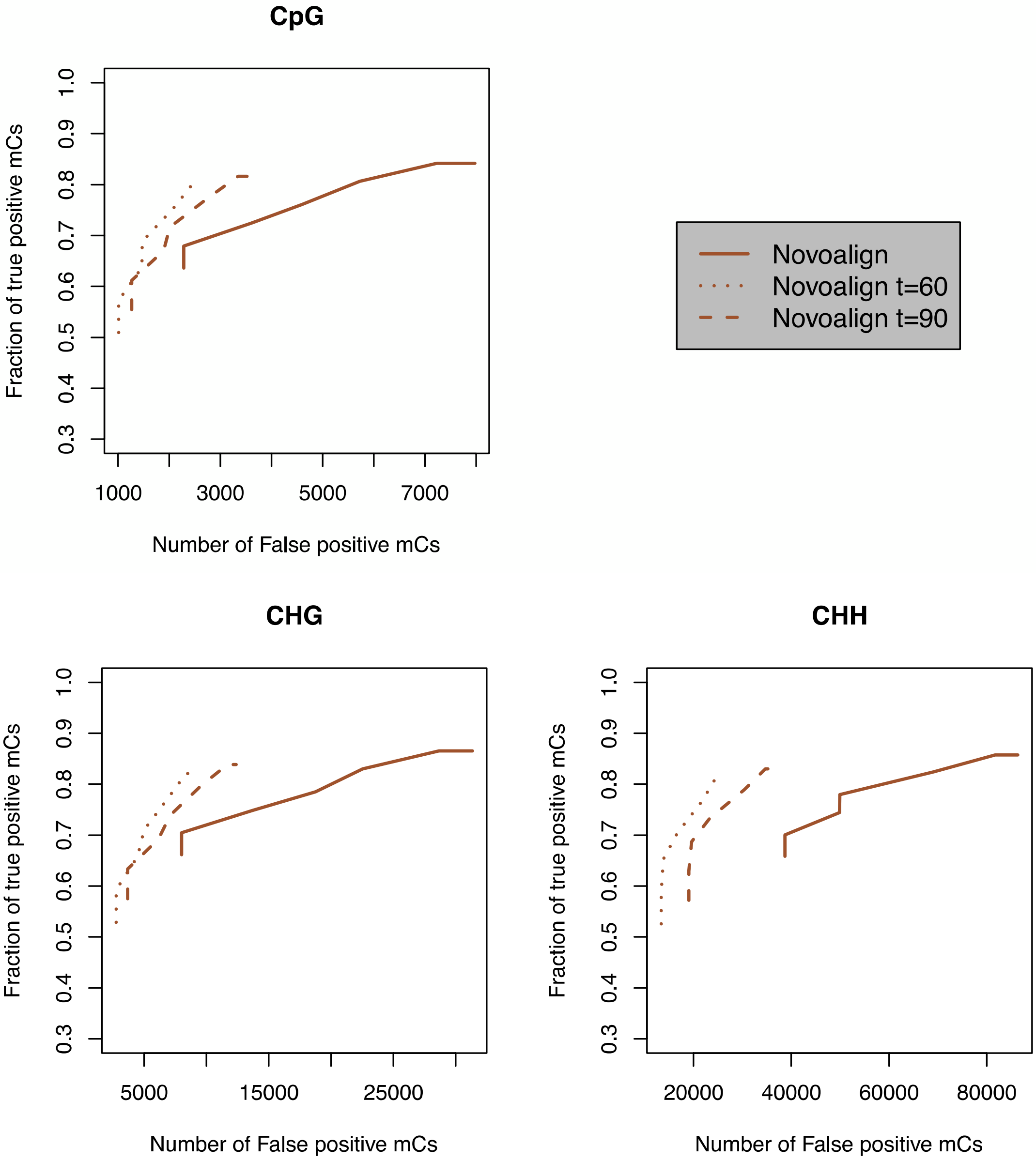


Supplementary Figure 14 (continued)



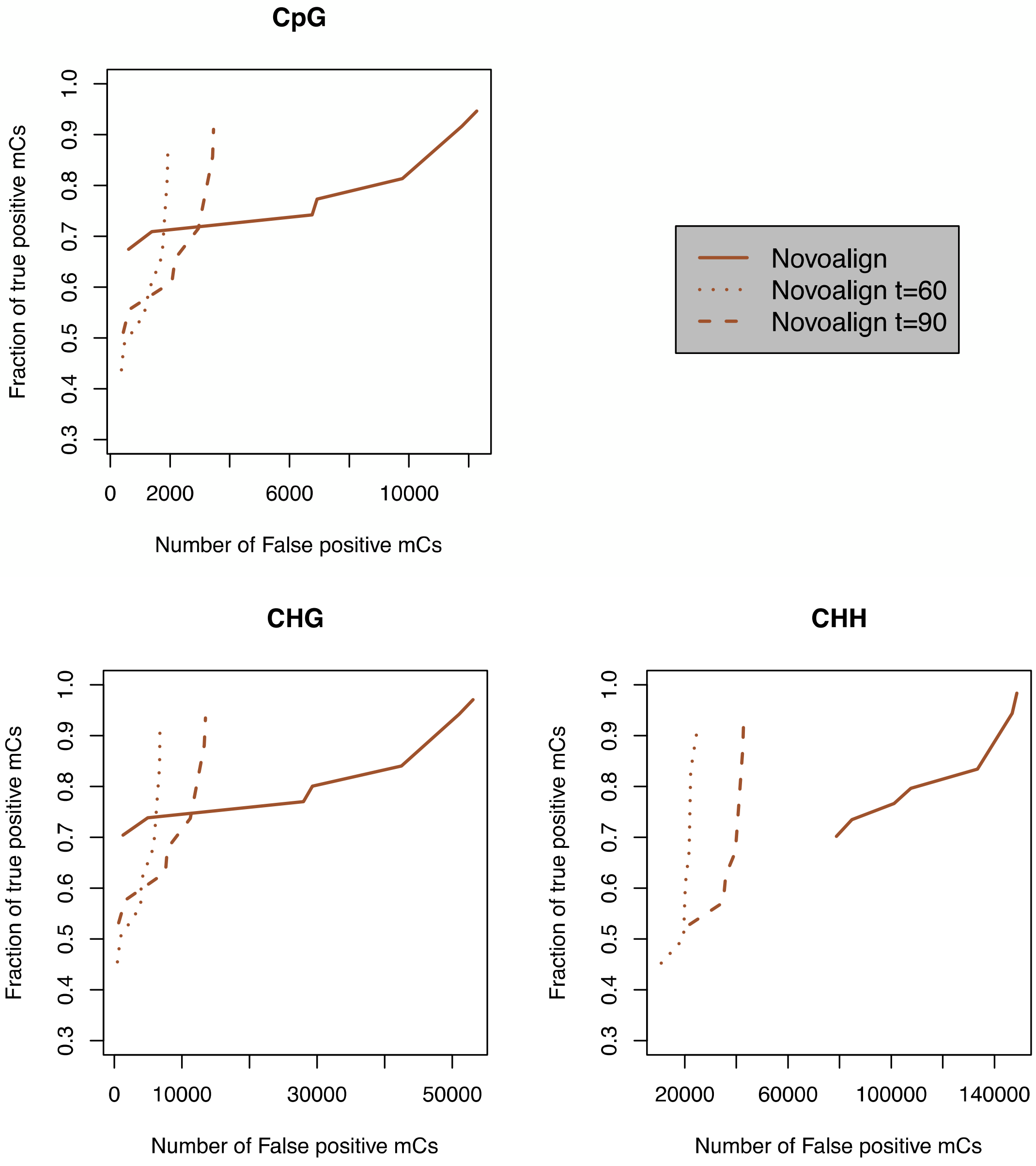
Supplementary Figure 15

Dataset A

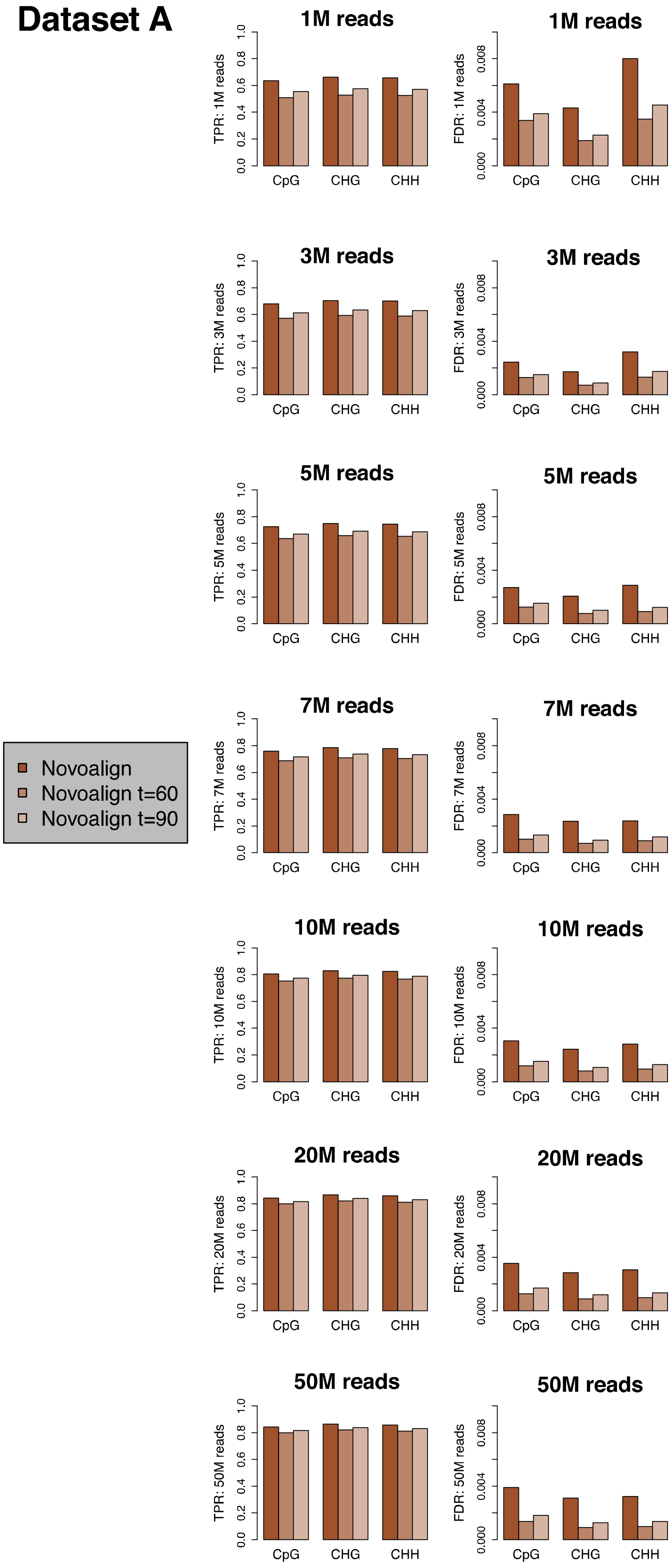


Supplementary Figure 15 (continued)

Dataset B

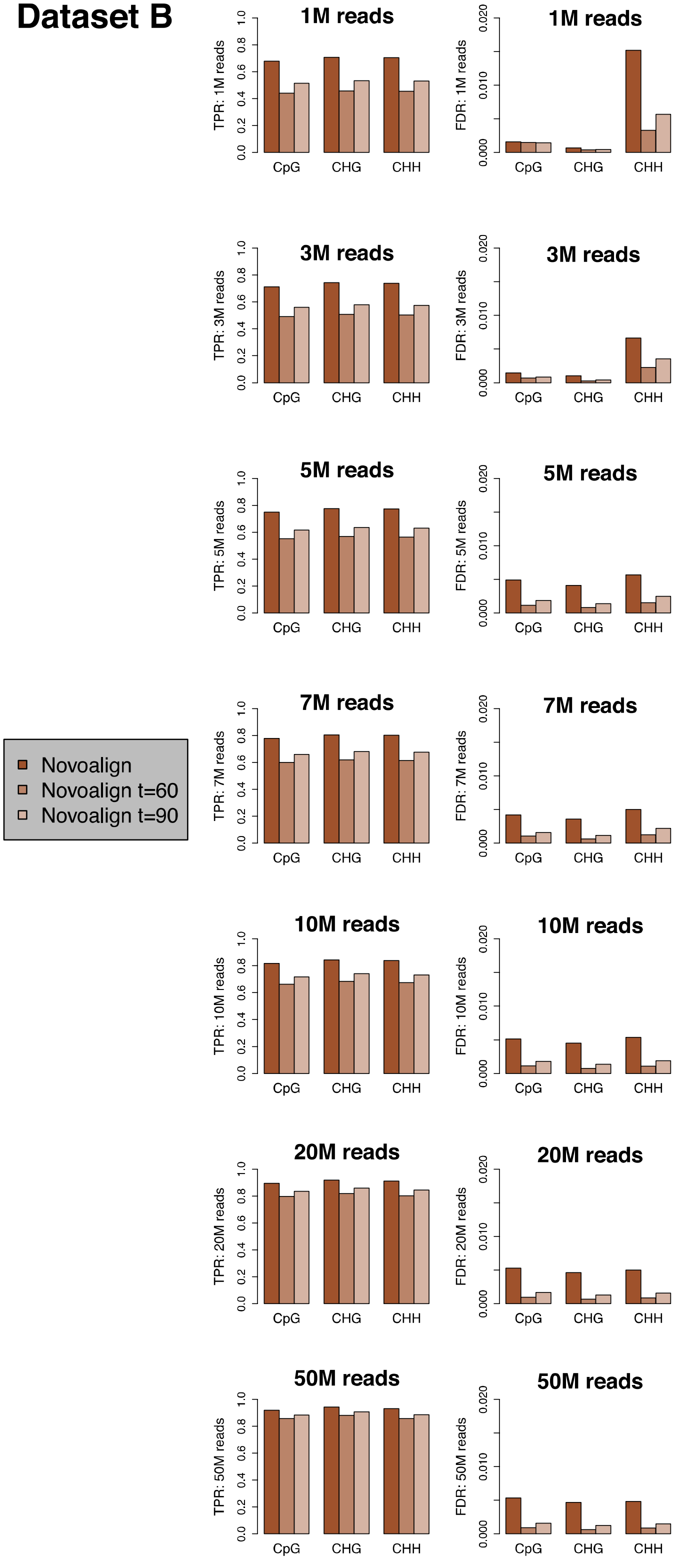


Supplementary Figure 15 (continued)

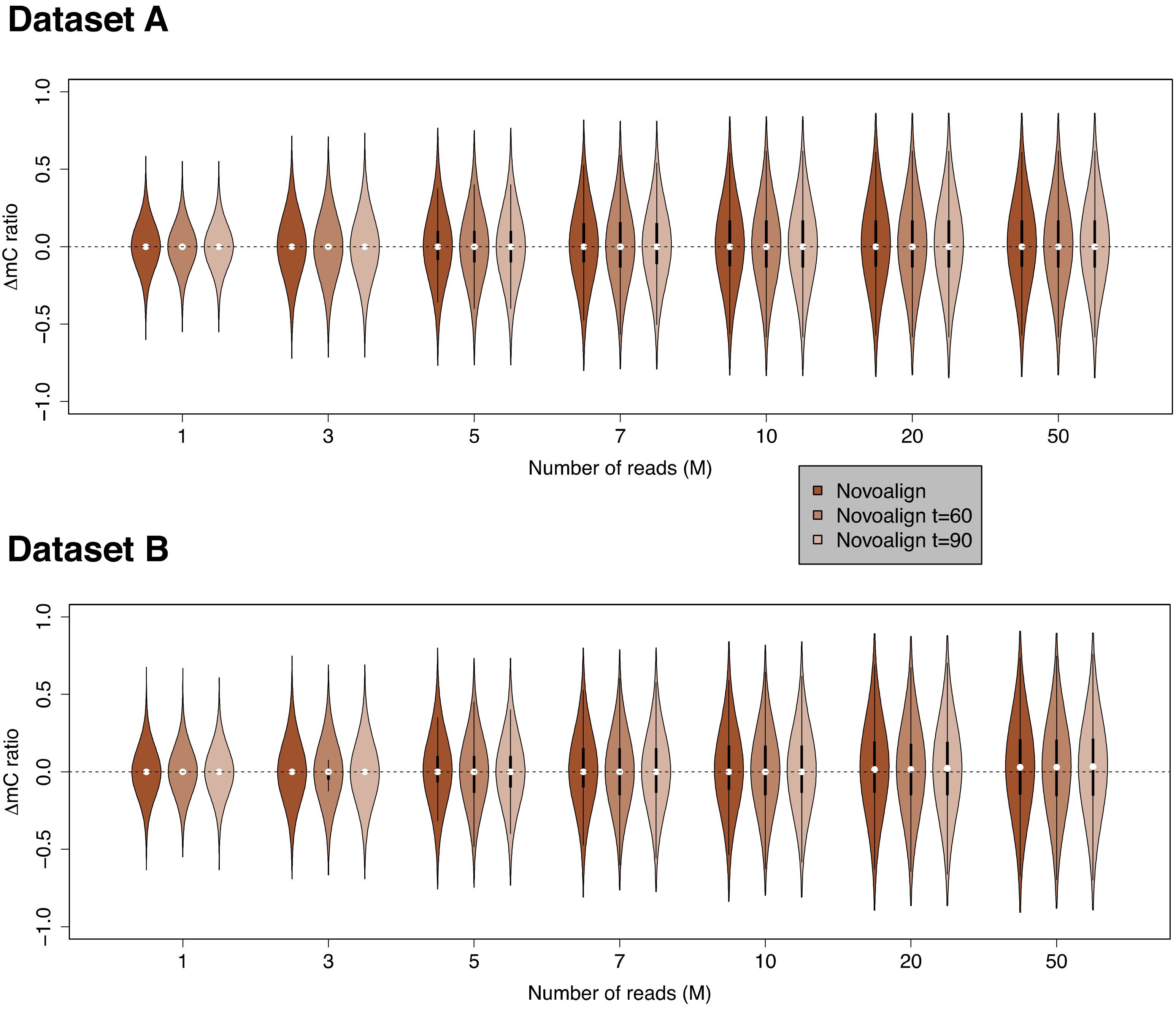


Supplementary Figure 15 (continued)

Dataset B



Supplementary Figure 15 (continued)


Dataset A
ΔmC ratio
Number of reads (M)
■ Novoalign
■ Novoalign t=60
■ Novoalign t=90
Dataset B
ΔmC ratio
Number of reads (M)