SUPPLEMENTAL FIGURES

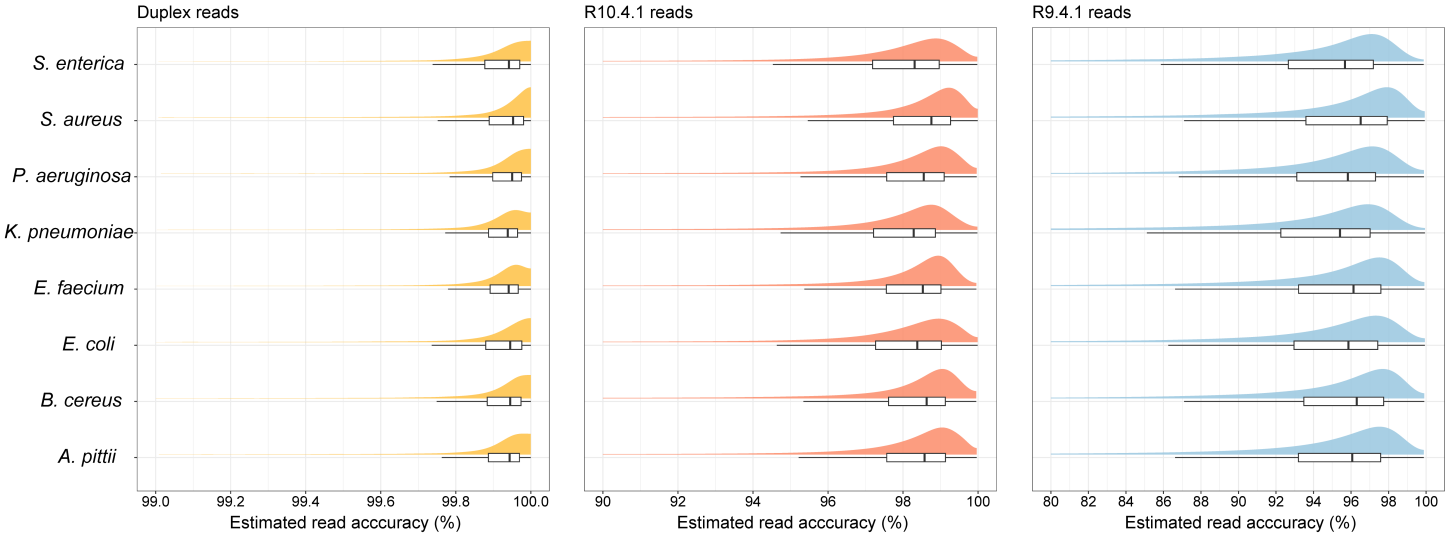

Figure S1. The distribution of WGS data quality including duplex, R10.4.1, and R9.4.1 reads for each bacterial sample.

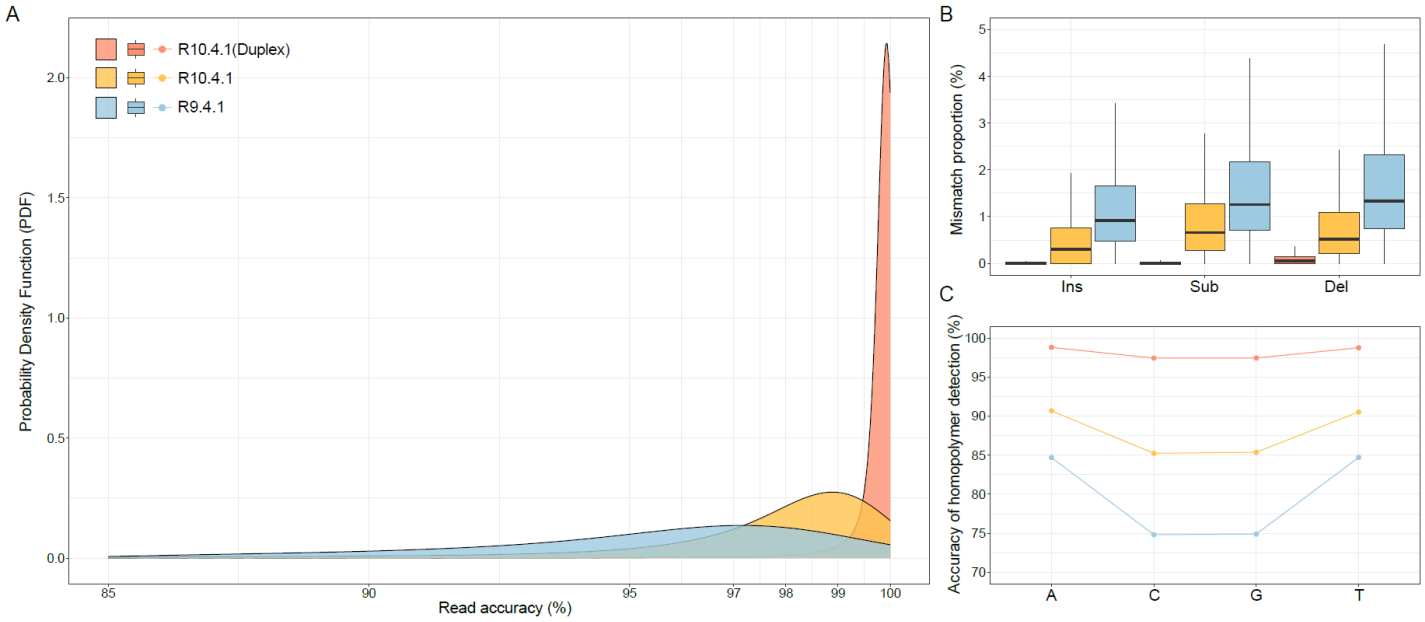

Figure S2. (A) Observed read accuracy of R10.4.1, R9.4.1, and duplex reads. **(B)** Mismatches distribution among insertions, nucleotide substitution, and deletions for each read type. **(C)** Homopolymer detection accuracy showcases perfect matches for each read type's A, C, G, and T bases. Ins: insertion, Sub: substitution, Del: deletion.

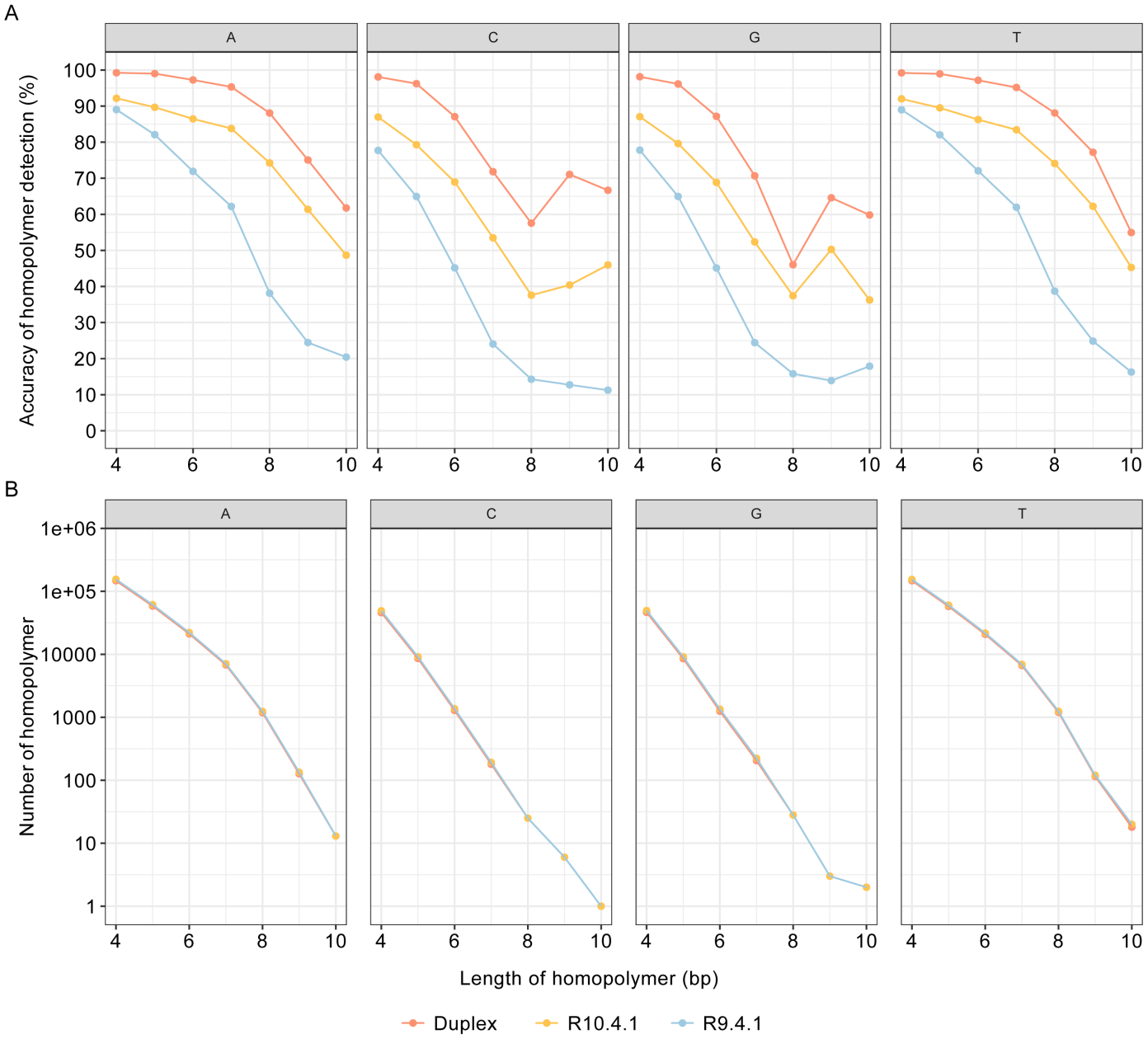

Figure S3. (A) Homopolymer detection accuracy and **(B)** homopolymer number for R10.4.1, R9.4.1, and duplex reads in different lengths.

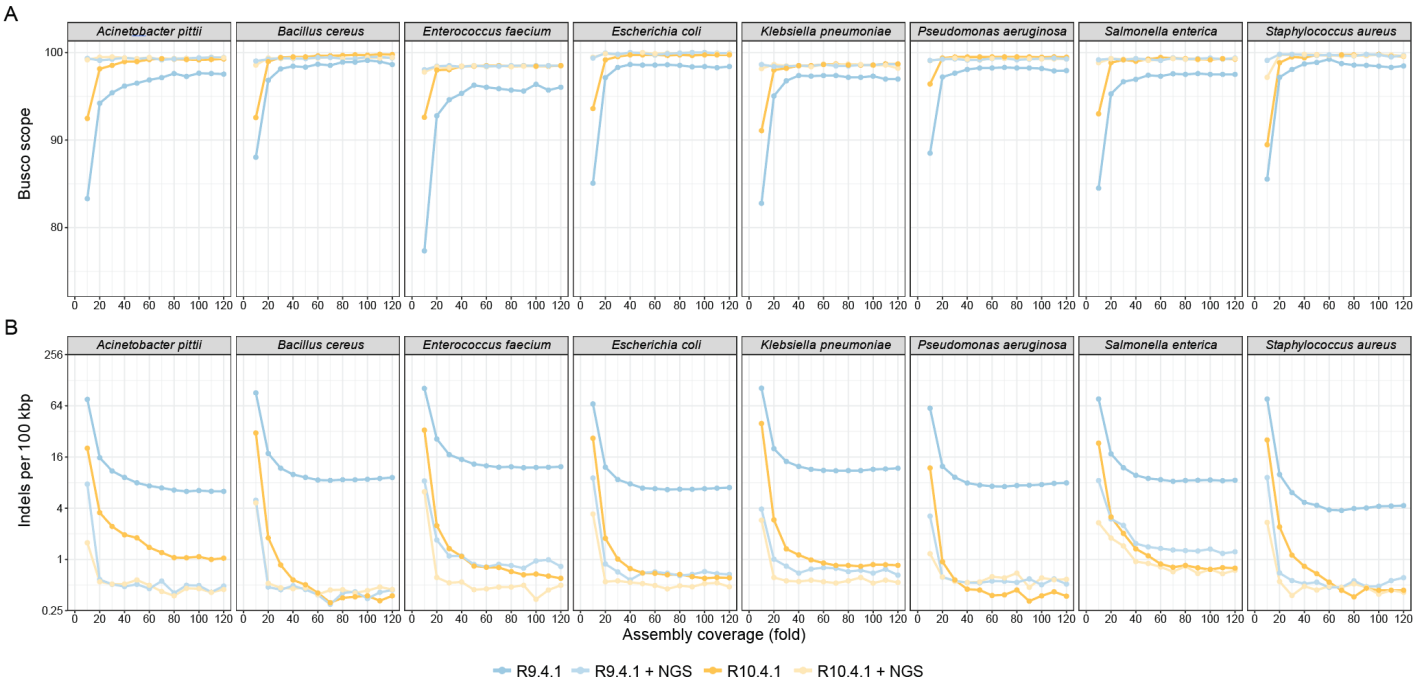

Figure S4. (A) Busco scores and **(B)** indels per 100 Kbps of assemblies generated using different coverage of R10.4.1 and R9.4.1 reads, with or without short-read data polishing. The X-axis shows the subsampled read coverage for ONT reads. Busco: Benchmarking Universal Single-Copy Orthologs.

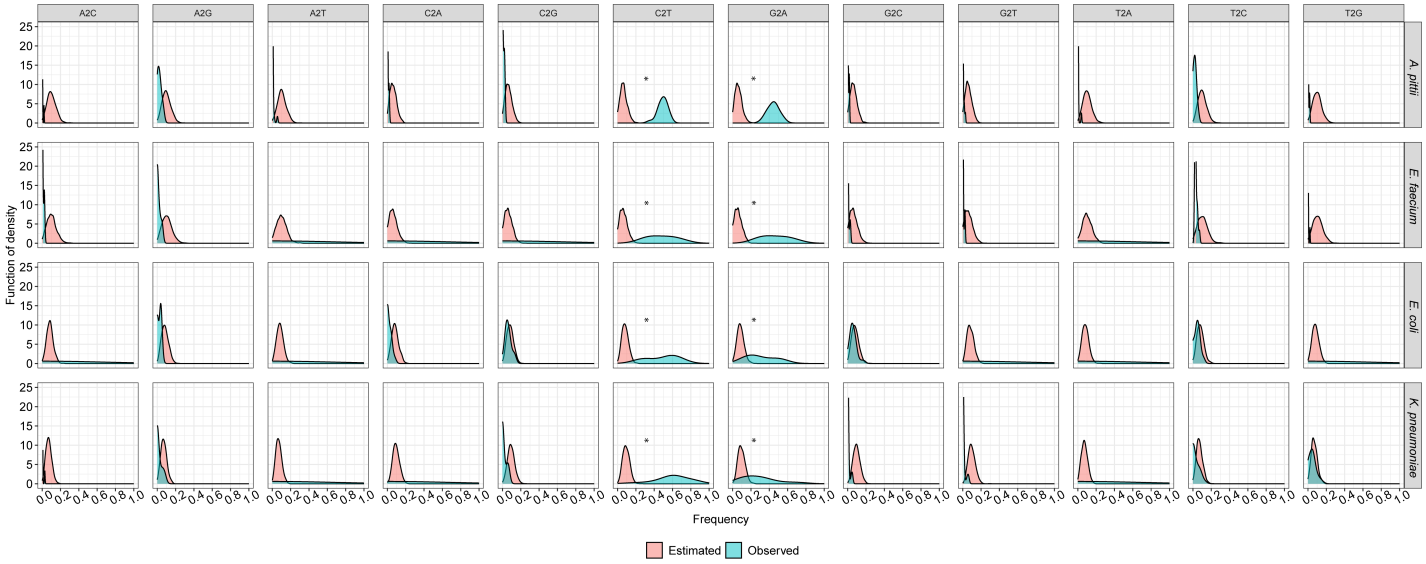

Figure S5. The substitution frequency comparison between estimation (n=1,000) and observation (n=35). The estimation was built by 1,000 times simulations to generate the count distribution of all 12 substitution types in four bacterial genomes, with the parameter of "one random substitution per 100 kb". The observed substitution frequency was obtained from the 35 assemblies from each bacterial sample. * p-value < 0.01, Student's t-test.

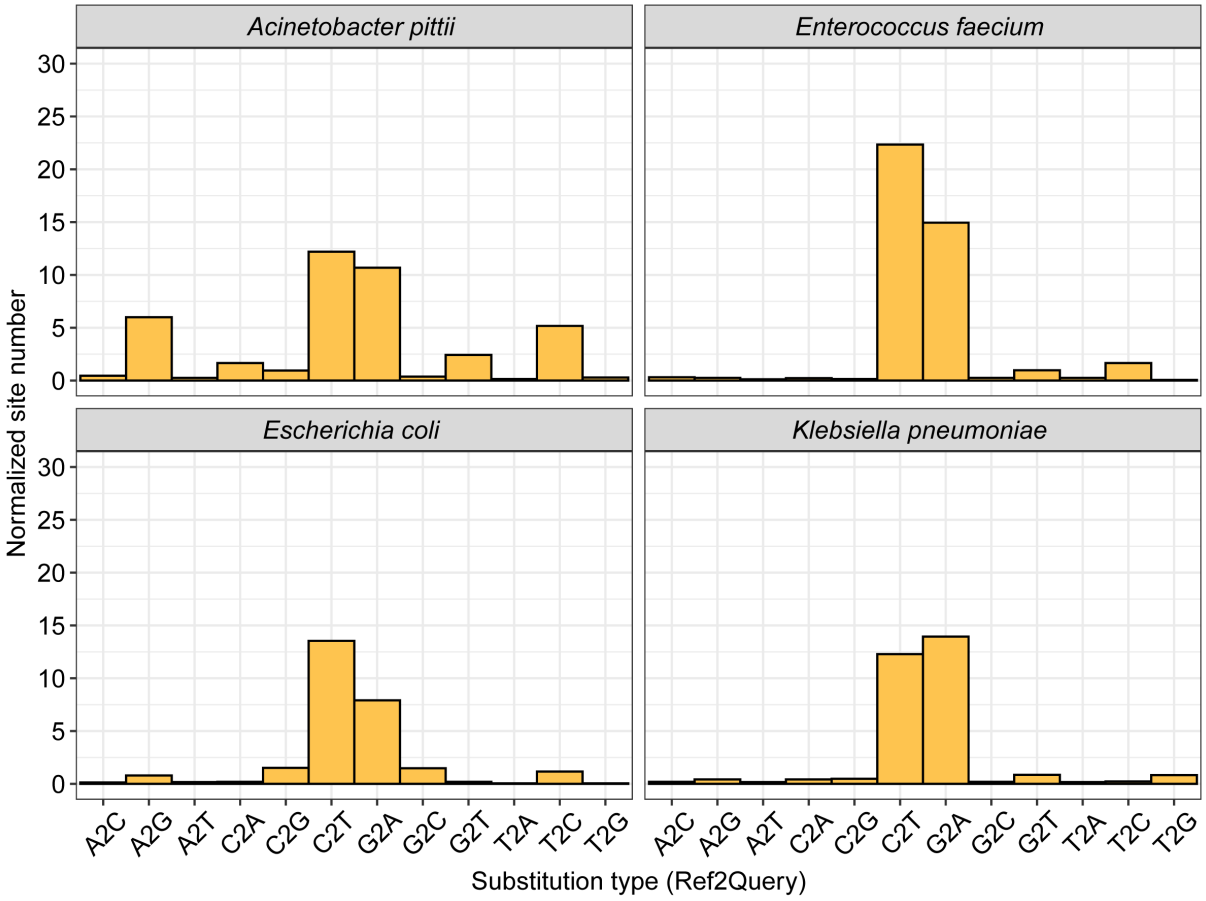

Figure S6. The normalized number of twelve single nucleotide substitution (SNS) types in assembles of four bacterial species using R10.4.1 read-based assemblies with an additional polish processing using Racon.

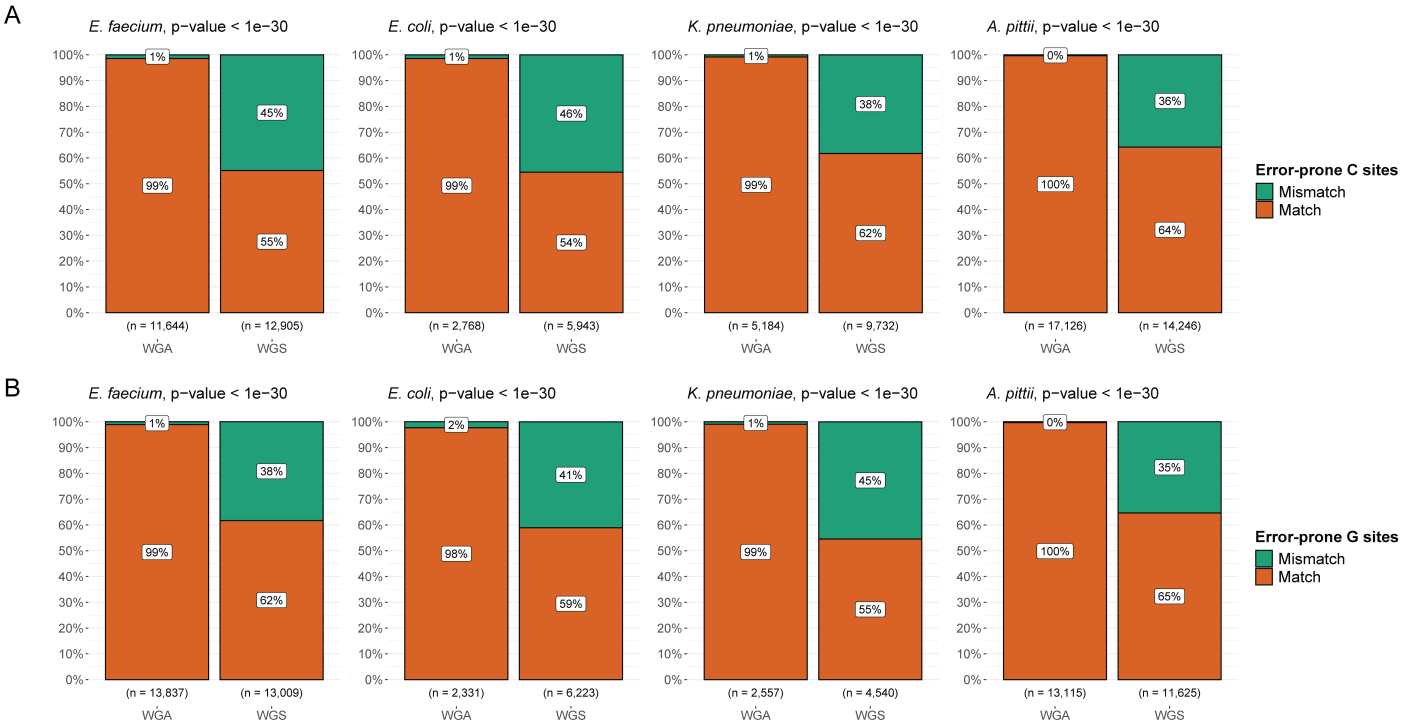

Figure S7. (A), (B) Proportions and counts of accurately and inaccurately mapped R10.4.1 WGS and WGA reads in *E. faecium*, *E. coli*, *K. pneumoniae*, and *A. pittii* at error-prone C sites and G sites, respectively. All the p-values using Fisher's exact test were less than 1e-30. WGS: whole-genome sequencing; WGA: whole-genome amplification.

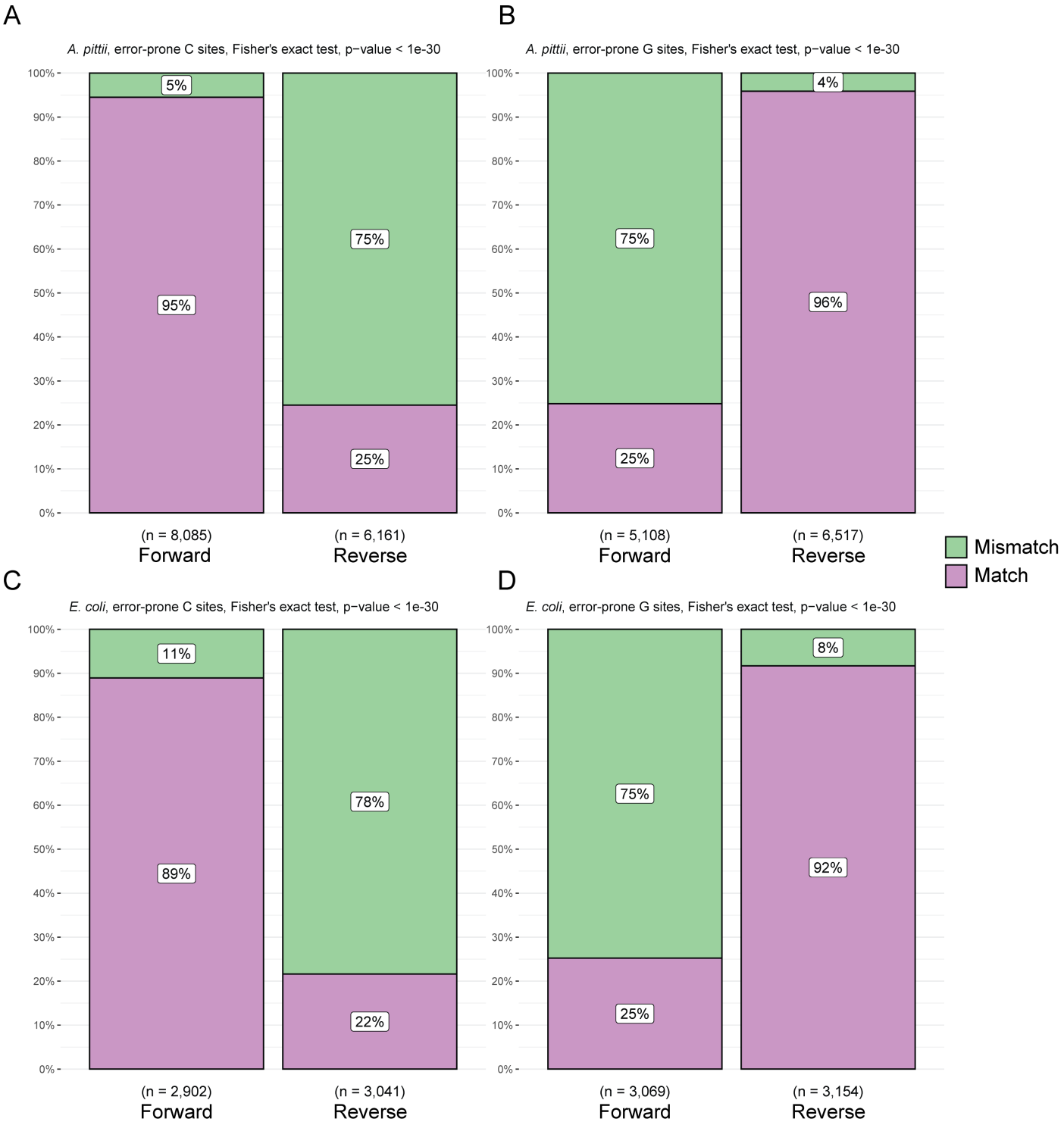

Figure S8. (A) and **(B)** Proportions and counts of accurately and inaccurately mapped R10.4.1 forward and reverse reads at error-prone C and G sites in *A. pittii* and *E. coli*, respectively. All the p-values of the Fisher test less than 1e-30.

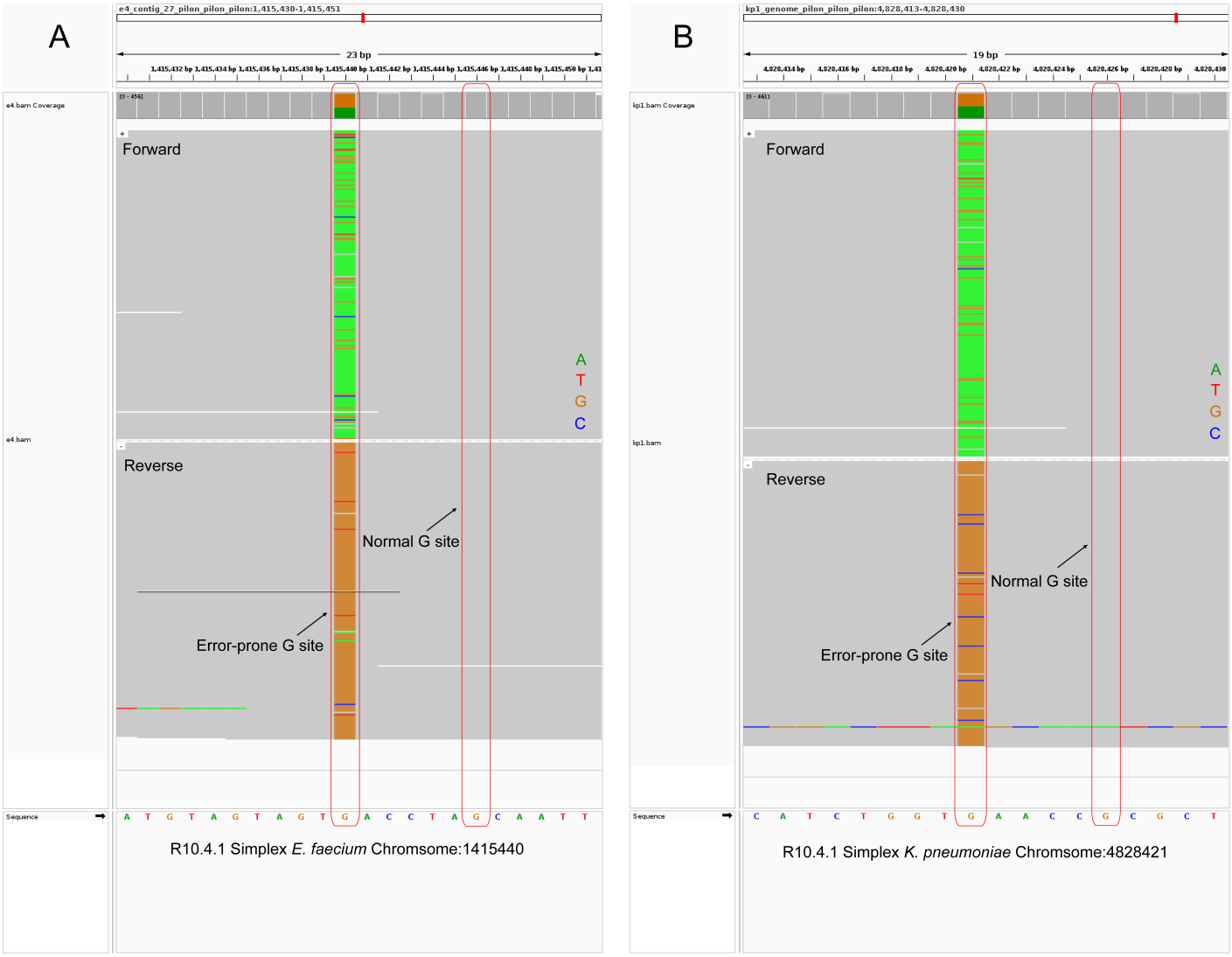

Figure S9. (A) and **(B)** IGV snapshots of R10.4.1 mismatches at an error-prone G site in *E. faecium* (chromosome position 1,415,440) and *K. pneumoniae* (chromosome position 4,828,421) genome, respectively. The forward strand reads show a high proportion of mismatch (G2A). The normal G sites of 1,415,446 in *E. faecium* chromosome and 4,828,426 in *K. pneumoniae* chromosome were used as the negative control to show the influence of methylation on strands. The grey color indicates matched bases. Only the A, T, C, and G at the error-prone sites are highlighted in different colors. The p-values of the Fisher Test were less than 1e-30 between the error-prone G site and another normal G site.

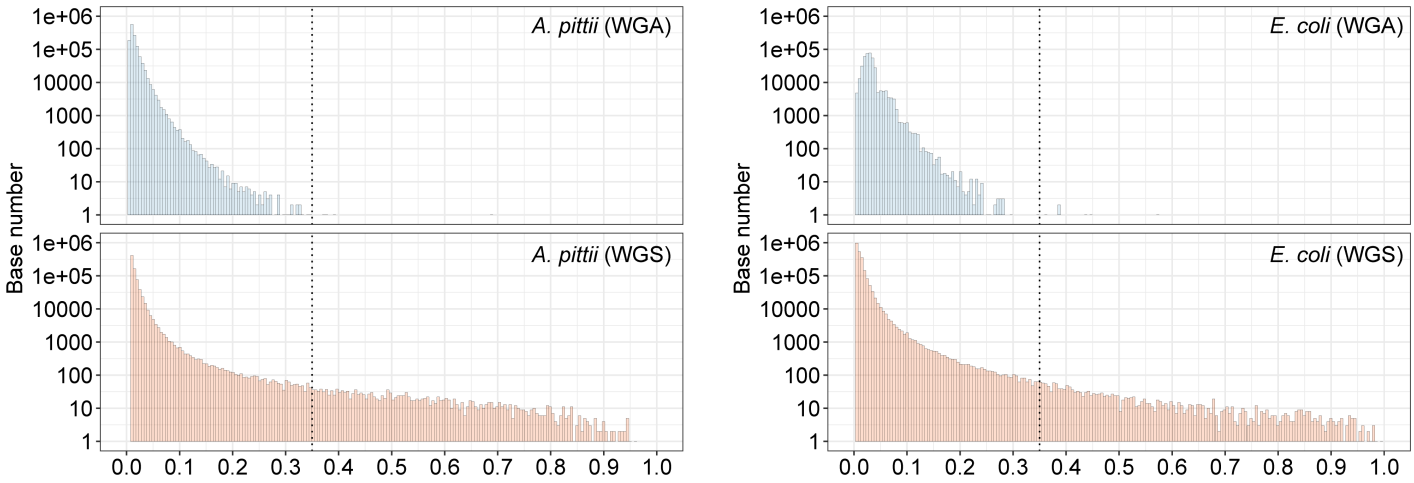

Figure S10. The distribution of site difference index in WGA and WGS sequencing for *A. pittii* and *E. coli*, respectively. The WGA sequencing, representing random read errors, serves as a background filter. High discrepancies between forward and reverse strands in WGS reads suggest potential DNA modifications. A cutoff of 0.35 (FDR < 1e-06 in WGA reads) is used here to identify possible DNA modification sites in WGS reads.

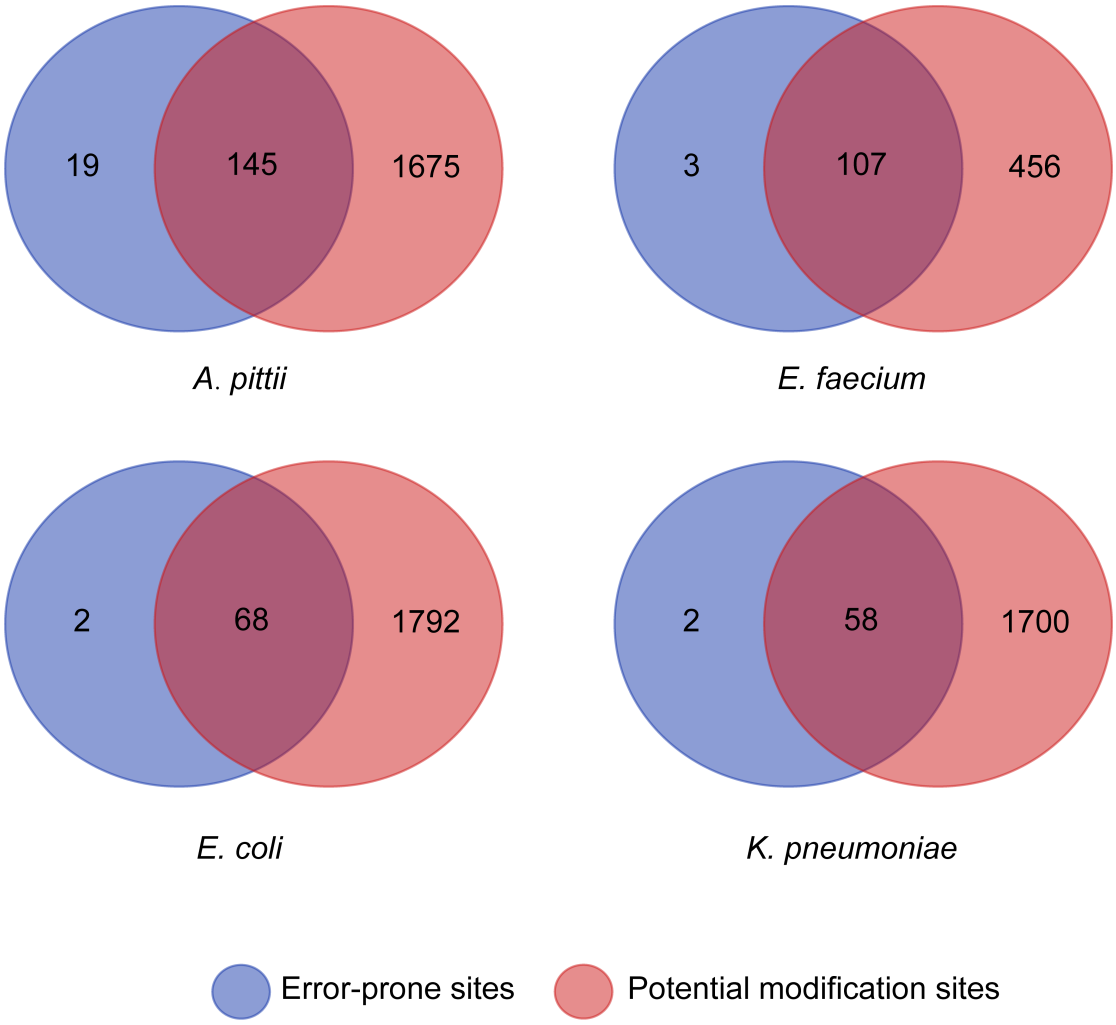

Figure S11. The intersection of error-prone sites and potential modification sites in *A. pittii*, *E. faecium*, *E. coli*, and *K. pneumoniae*.

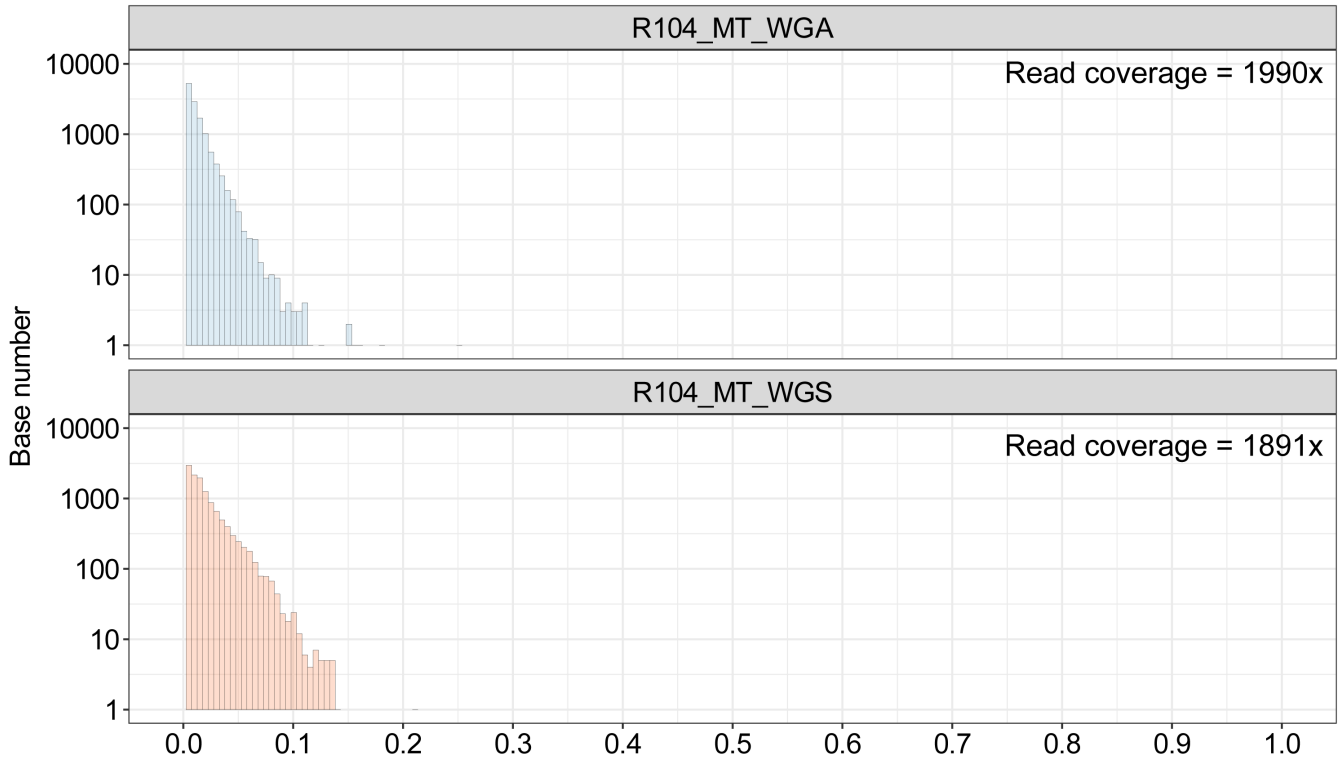

Figure S12. The distribution of site difference index in WGA (32,981,816 bp) and WGS (31,338,505 bp) sequencing for human R10.4 datasets. Based on the limitation of read coverage, only the sites in the mitochondrial genome (MT, size=16,569 bp) were used.

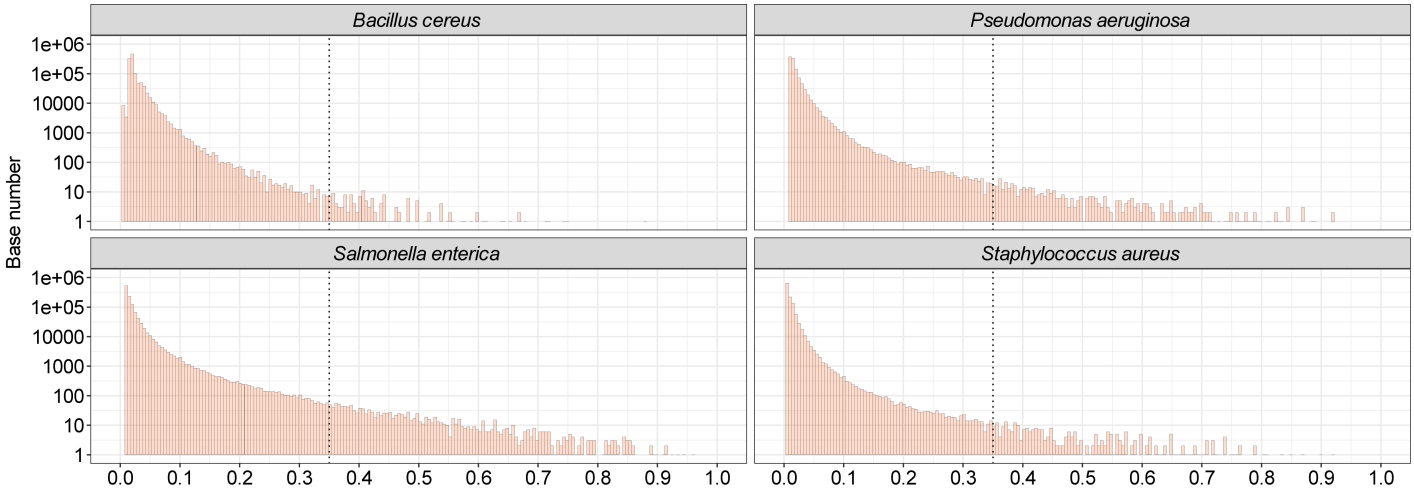

Figure S13. The count of strand nucleotide difference between forward and reverse reads in WGS sequencing reads in *B. cereus*, *P. aeruginosa, S. enterica,* and *S. aureus*, respectively.

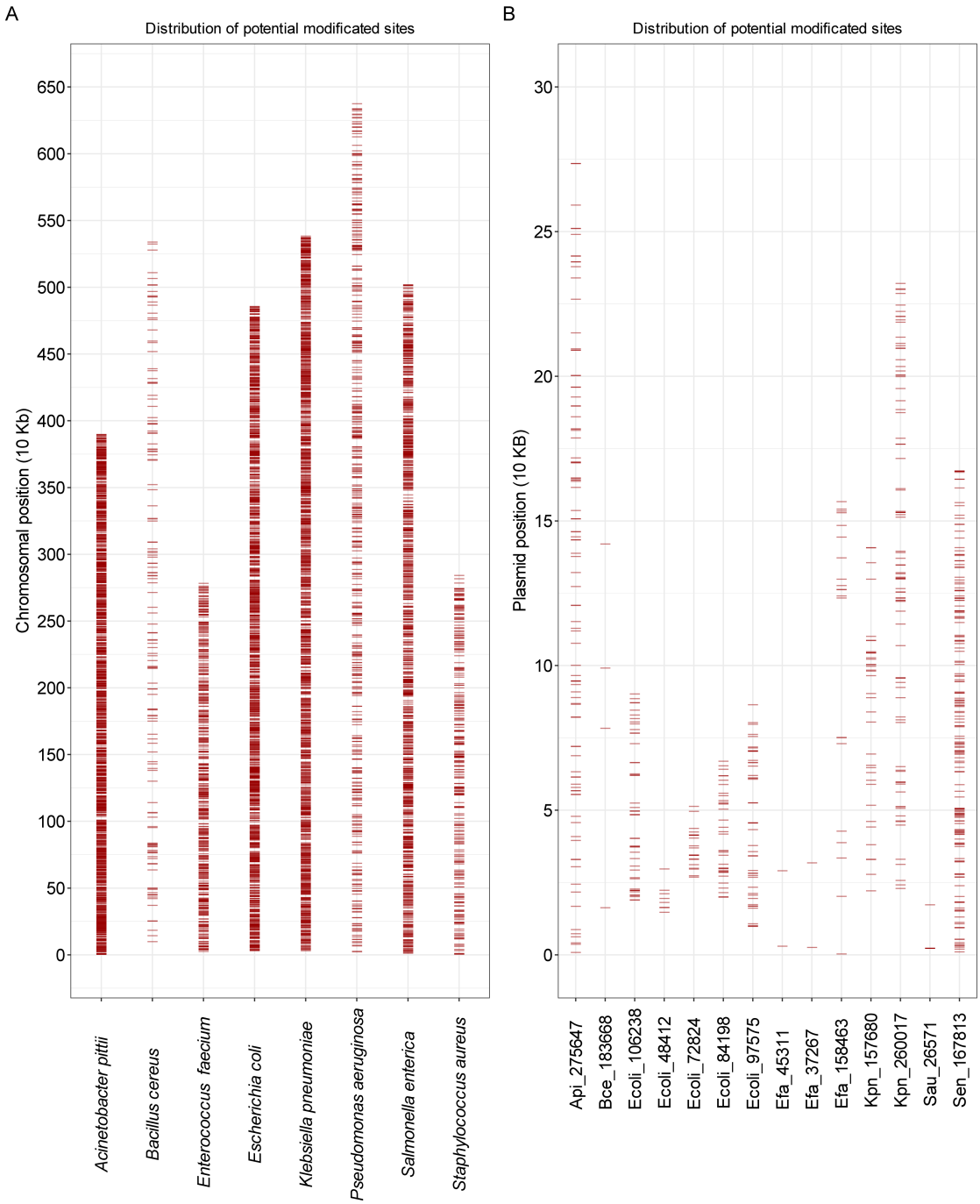

Figure S14. The distribution of potential modification sites in **(A)** chromosome and **(B)** plasmids. Notably, only plasmids with length exceeding 20K bp were chosen for analysis. The identification of each plasmid was determined by combining the bacterium's name with the plasmid length.

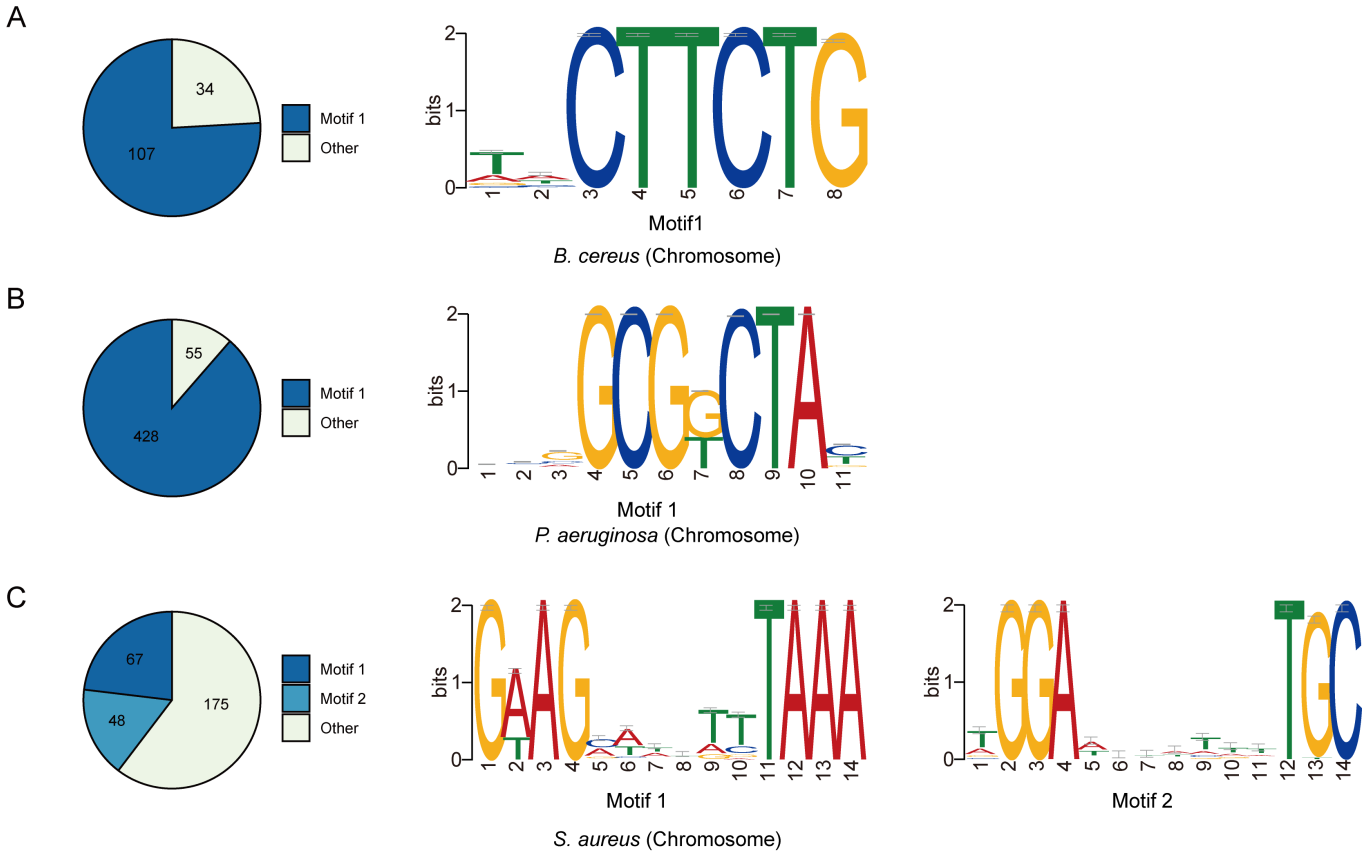

Figure S15. (A), **(B)**, and **(C)** The enriched motif for possible DNA modification sites identified by strand DNA accuracy comparison in chromosome sequences in *B. cereus*, *P. aeruginosa,* and *S. aureus*, respectively.

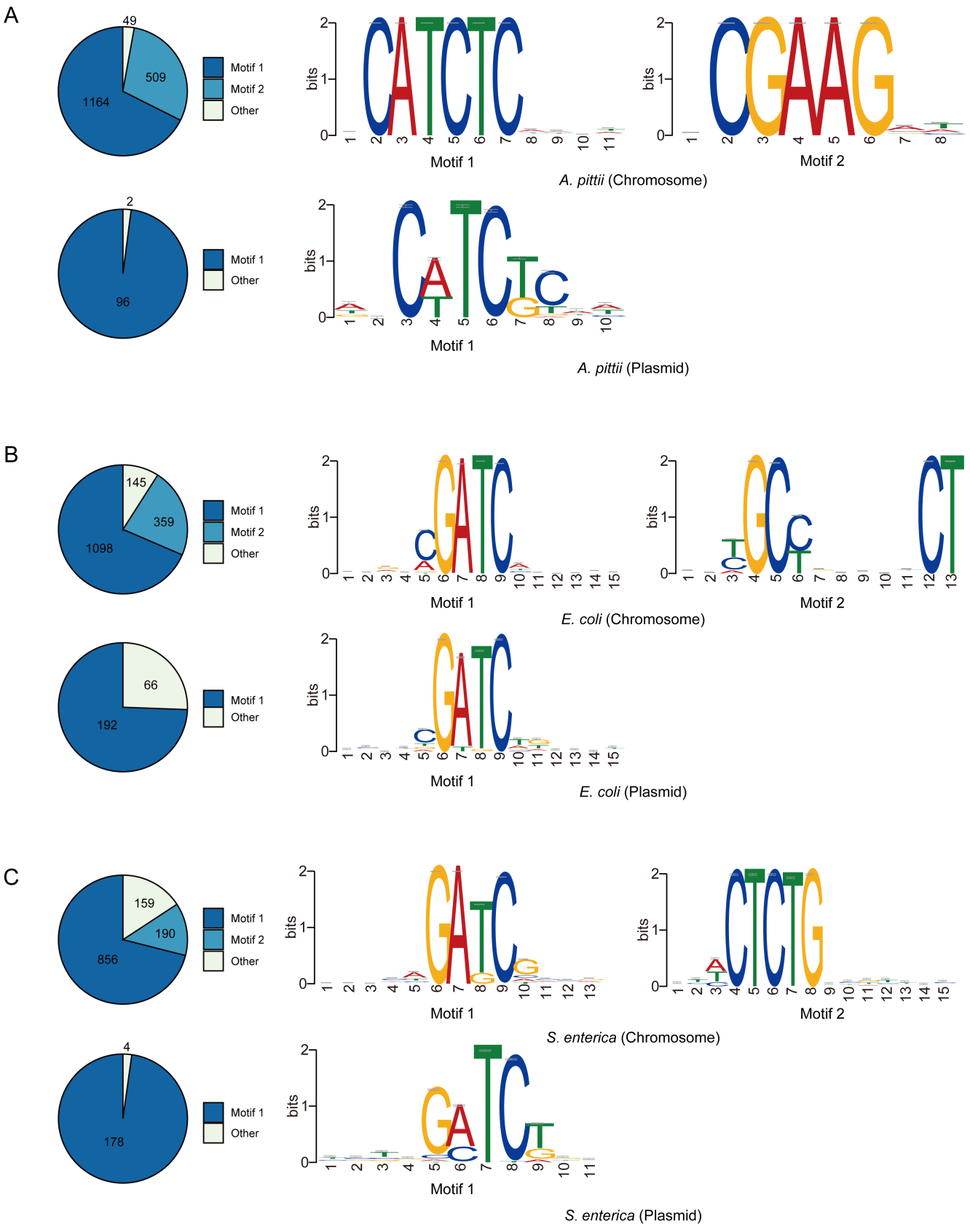

Figure S16. (A), **(B)**, and **(C)** The enriched motif for possible DNA modification sites identified by strand DNA accuracy comparison in chromosome and plasmid sequences in *A. piittii*, *E. coli*, and *S. enterica* respectively.

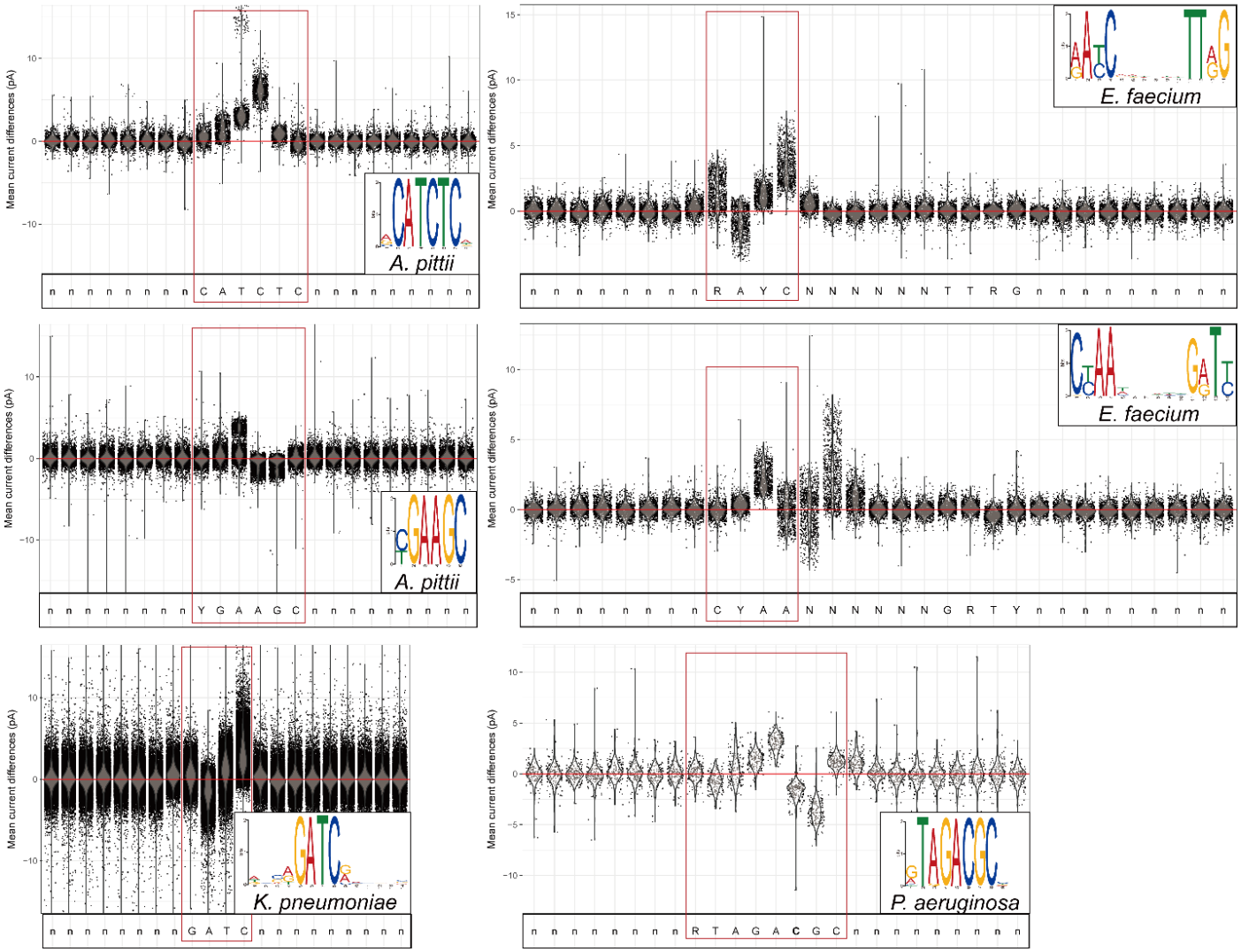

Figure S17. The current signal difference was observed in six shared motifs which were identified by Hammerhead and Nanodisco. The red reference line represents a current signal difference of zero and bases that exhibited a difference were highlighted with a red frame.

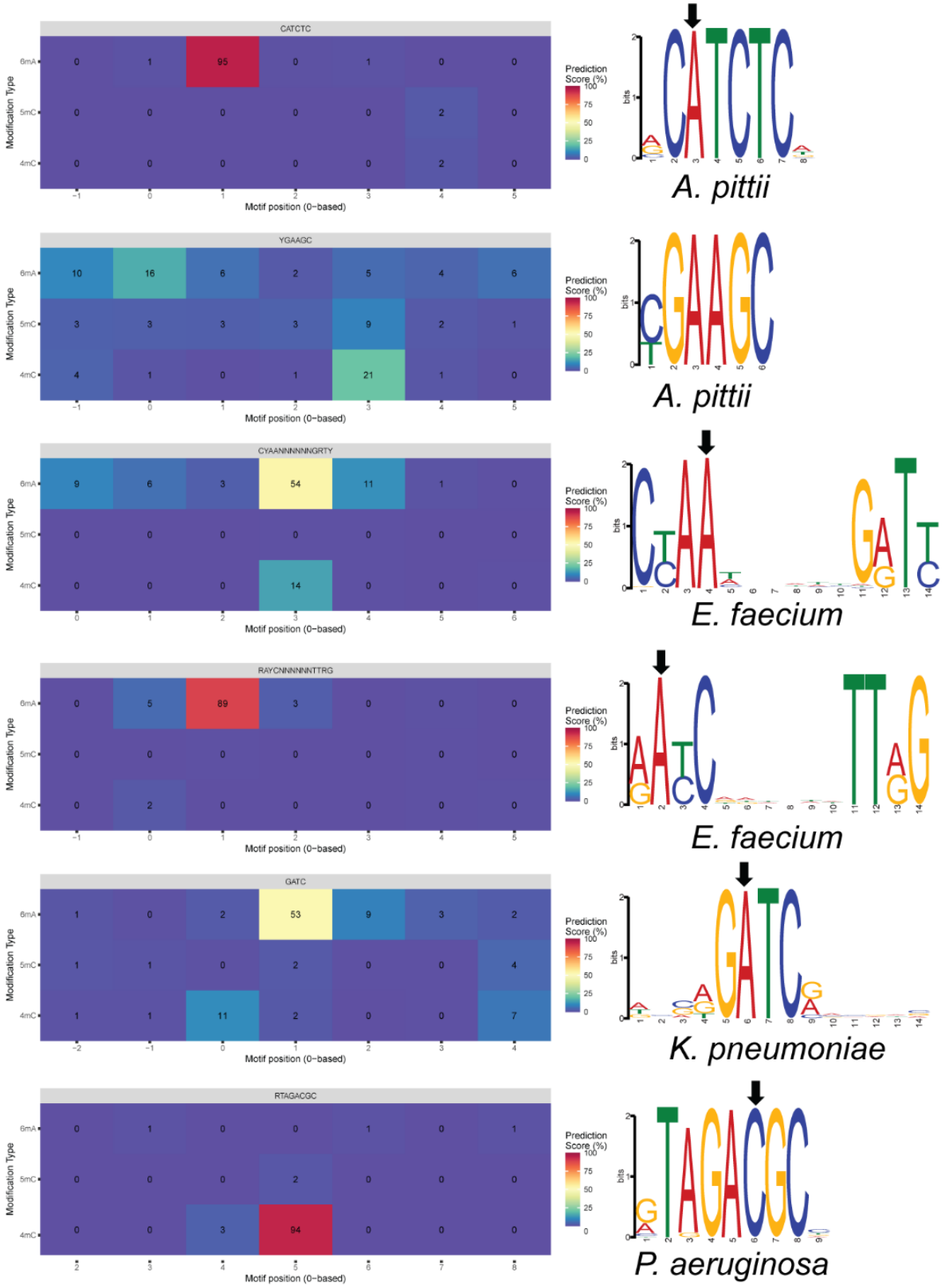

Figure S18. The methylation type predicted by Nanodisco for six shared motifs. The 6mA methylation was predicted at the CATCTC (*A. pittii*), CYAANNNNNNGRTY (*E. faecium*), RAYCNNNNNNTTRG (*E. faecium*), and GATC (*K. pneumoniae*) motifs. The 4mC methylation was predicted at the RTAGACGC (*P. aeruginosa*) motif. The modified base in the motif was highlighted.

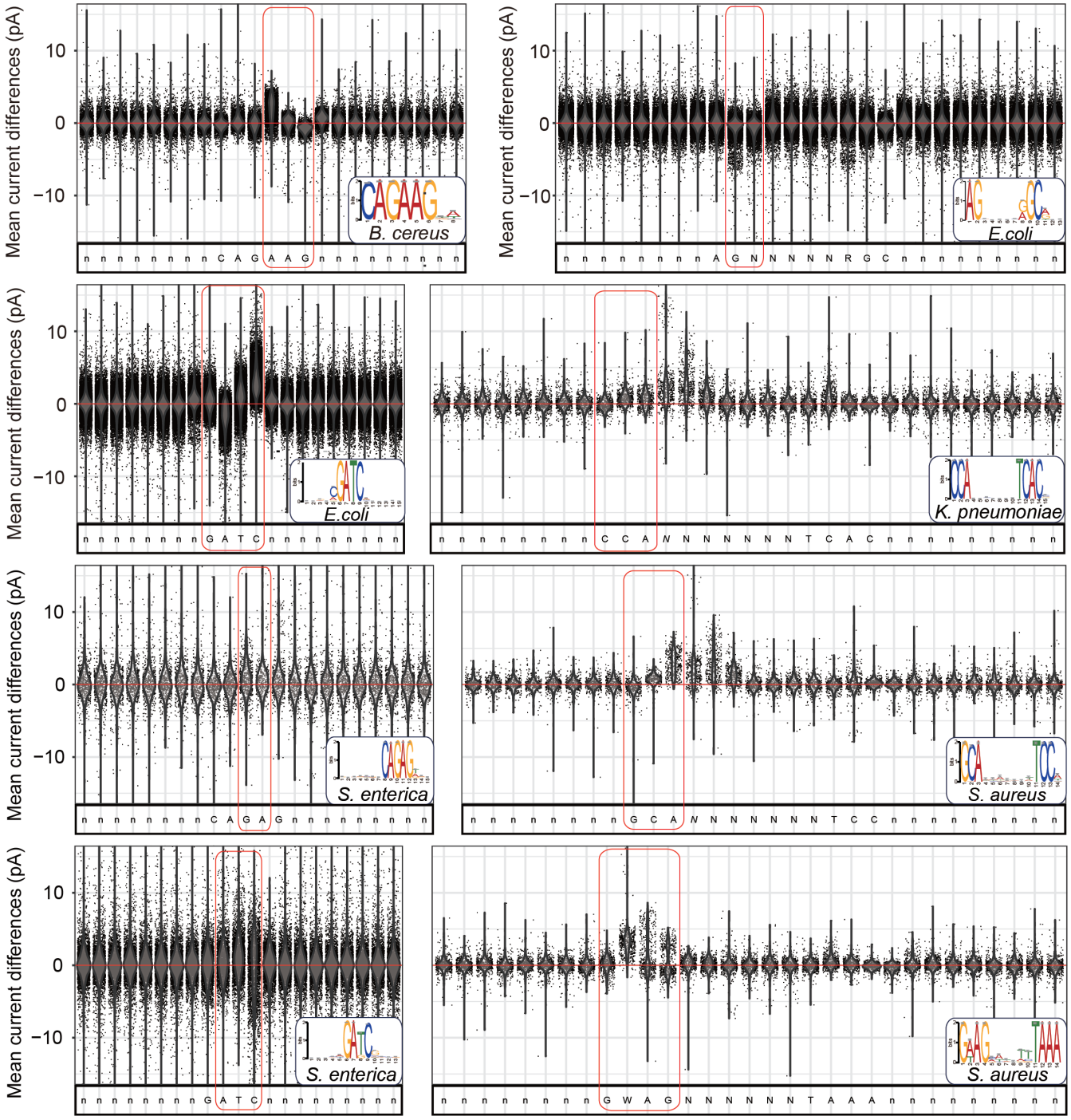

Figure S19. The current signal difference observed in the eight motifs in R9.4.1 reads. These motifs were uniquely detected by Hammerhead using R10.4.1 reads. The current signal differences were calculated by Nanodisco, utilizing the comparison between WGS reads and WGA reads. The red reference line represents a current signal difference of zero. The closer the trend is to the red line, the smaller the difference in the current signal. The bases that exhibited a difference were highlighted with a red frame. The "n" base surrounding the motifs can represent any of four DNA bases (A, T, G, or C), depending on the specific base neighboring the motifs in the genome.

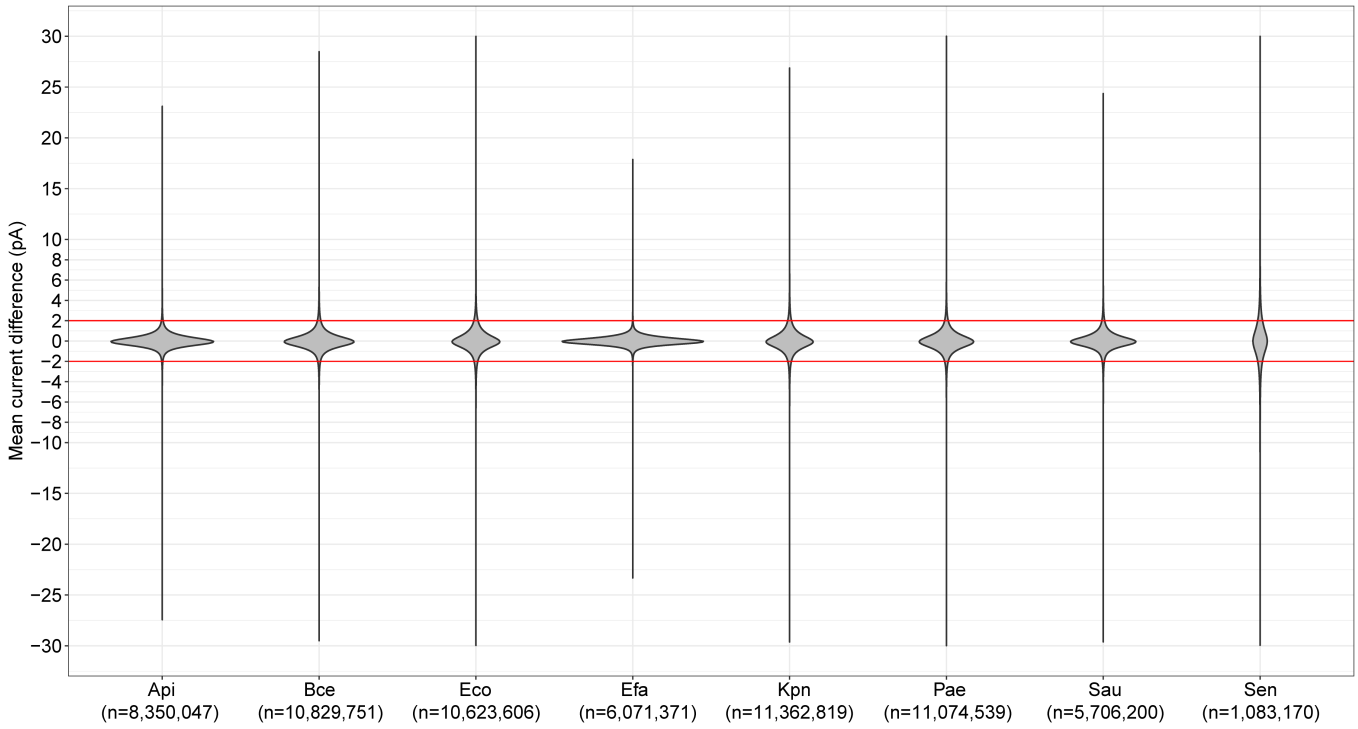

Figure S20. The distribution of mean current differences for each bacterial sample. All the values were profiled from the RDS file generated by Nanodisco. The sites with an absolute current difference over 2 pA were selected to compare with potential modification sites selected by Hammerhead. The red line means the mean current difference is -2 or 2 pA. Api: *Acinetobacter pittii*; Bec: *Bacillus cereus*; Eco: *Escherichia coli*; Efa: *Enterococcus faecium*; Kpn: *Klebsiella pneumoniae*; Pae: *Pseudomonas aeruginosa*; Sau: *Staphylococcus aureus*; Sen: *Salmonella enterica*.

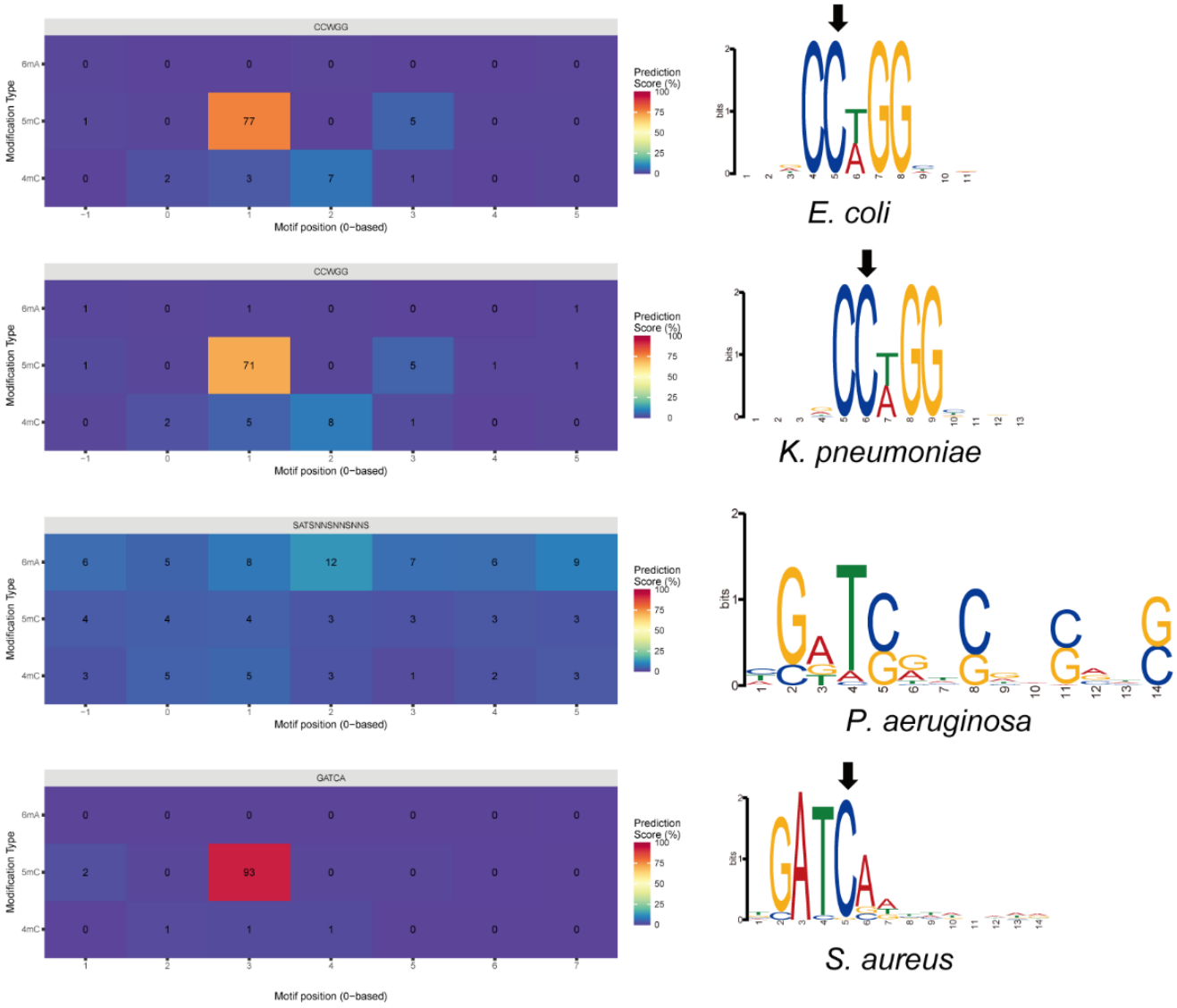

Figure S21. The methylation type predicted by Nanodisco for four unique motifs identified by Nanodisco. The second and fourth C base was predicted as 5mC in CCWGG and GATCA motifs, respectively.

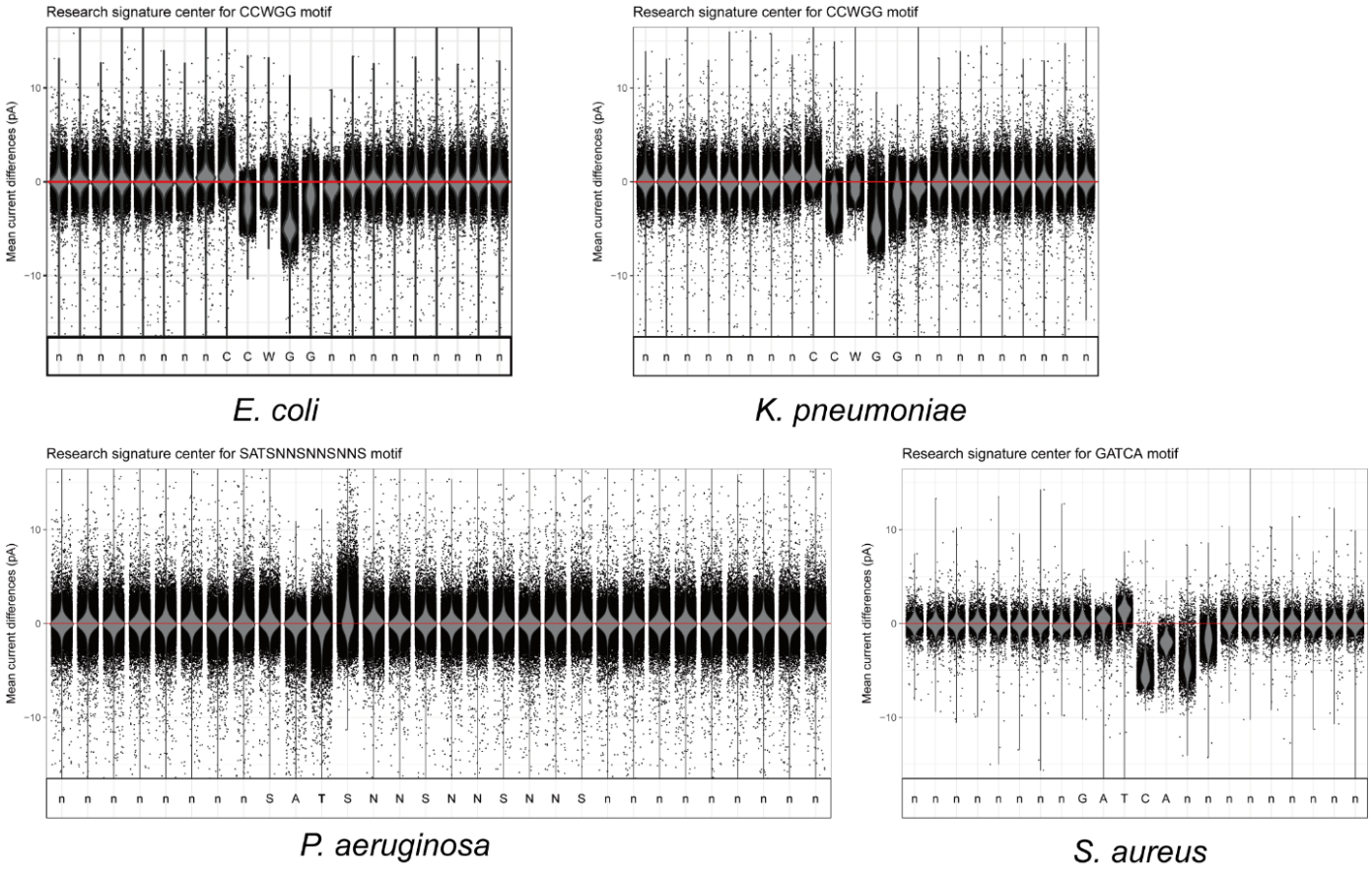

Figure S22. The current signal difference observed in four unique motifs identified by Nanodisco. The red reference line represents a current signal difference of zero.

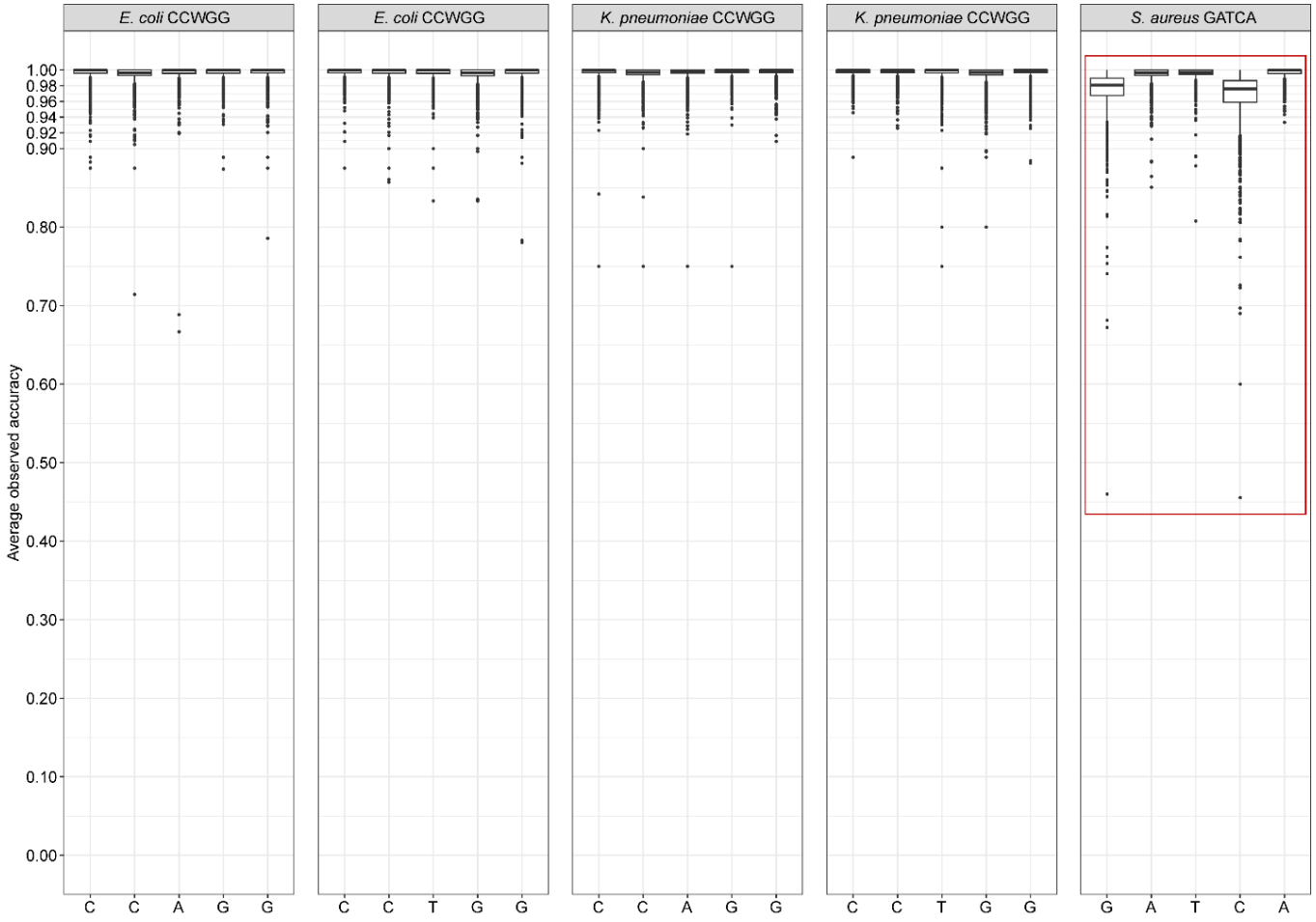

Figure S23. The observed read accuracy for the CCWGG motif within *E. coli* and *K. pneumoniae* genome and GATCA motif within *S. aureus* genome. The accuracy was calculated using R10.4.1 WGS reads.

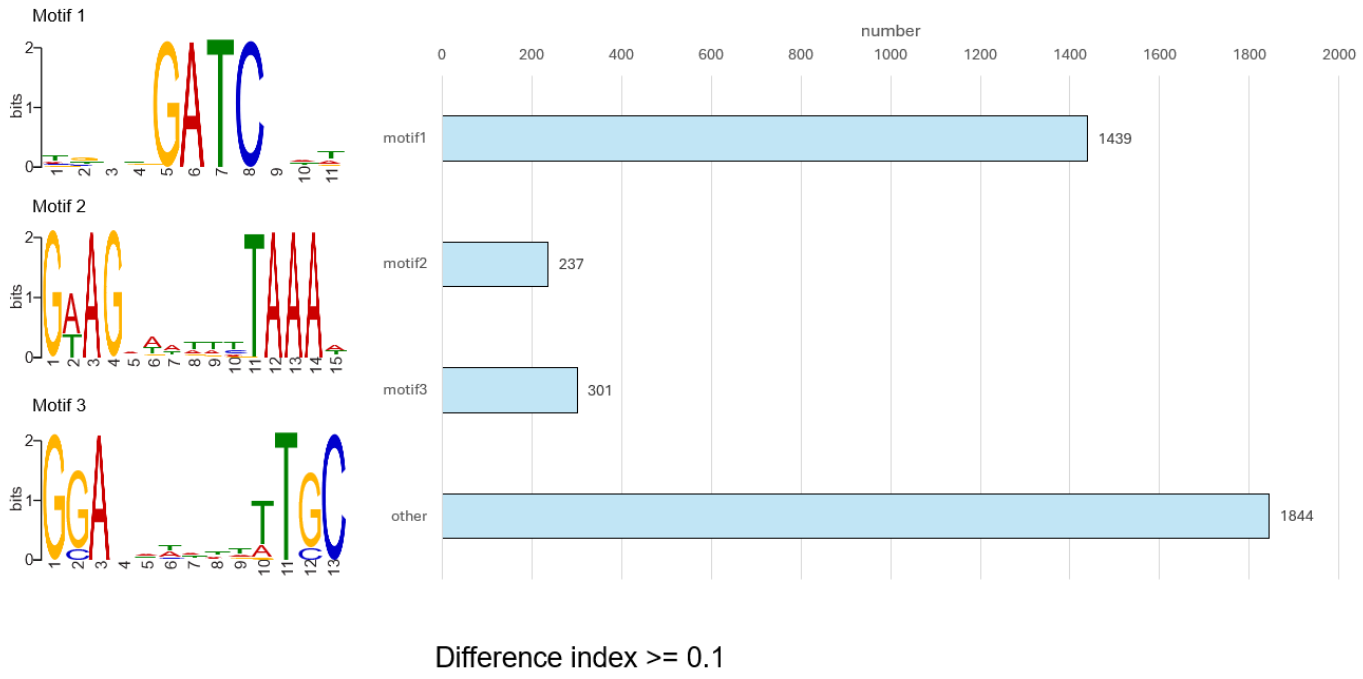

Figure S24. The motif enriched using the sequences near the potential modification sites (-10 bp to 10 bp) with a difference index over 0.1. Hammerhead pipeline was used to calculate the difference index using the R10.4.1 WGS reads of *S. aureus*.

Figure S25. The IGV of the *dam* and *dcm* genes are found in our *E. coli* strain genome.

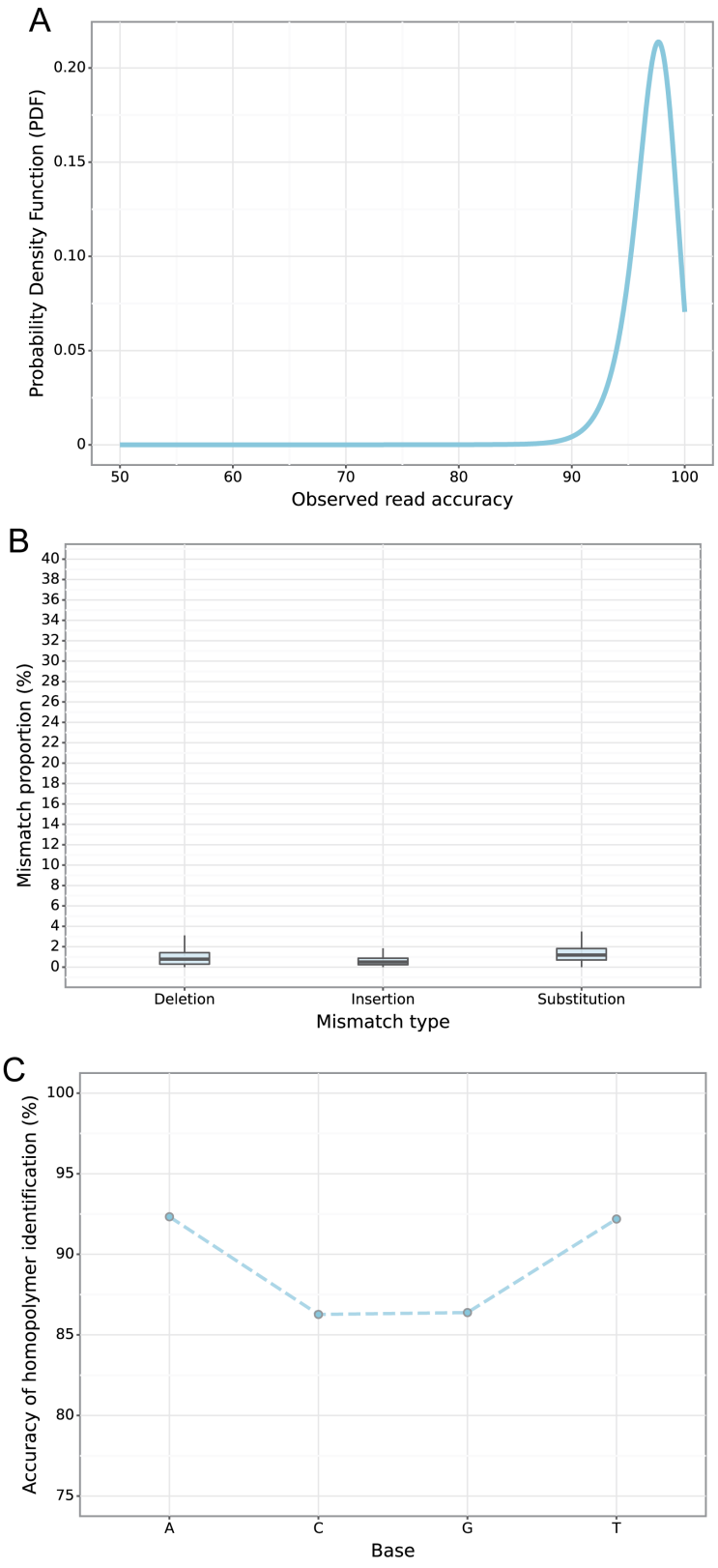

Figure S26. The read quality of re-basecalled WGS reads of *E. coli* using the "modification aware" model. (A) The density of read accuracy. **(B)** The proportion of three types of mismatches. **(C)** The accuracy of homopolymer identification.

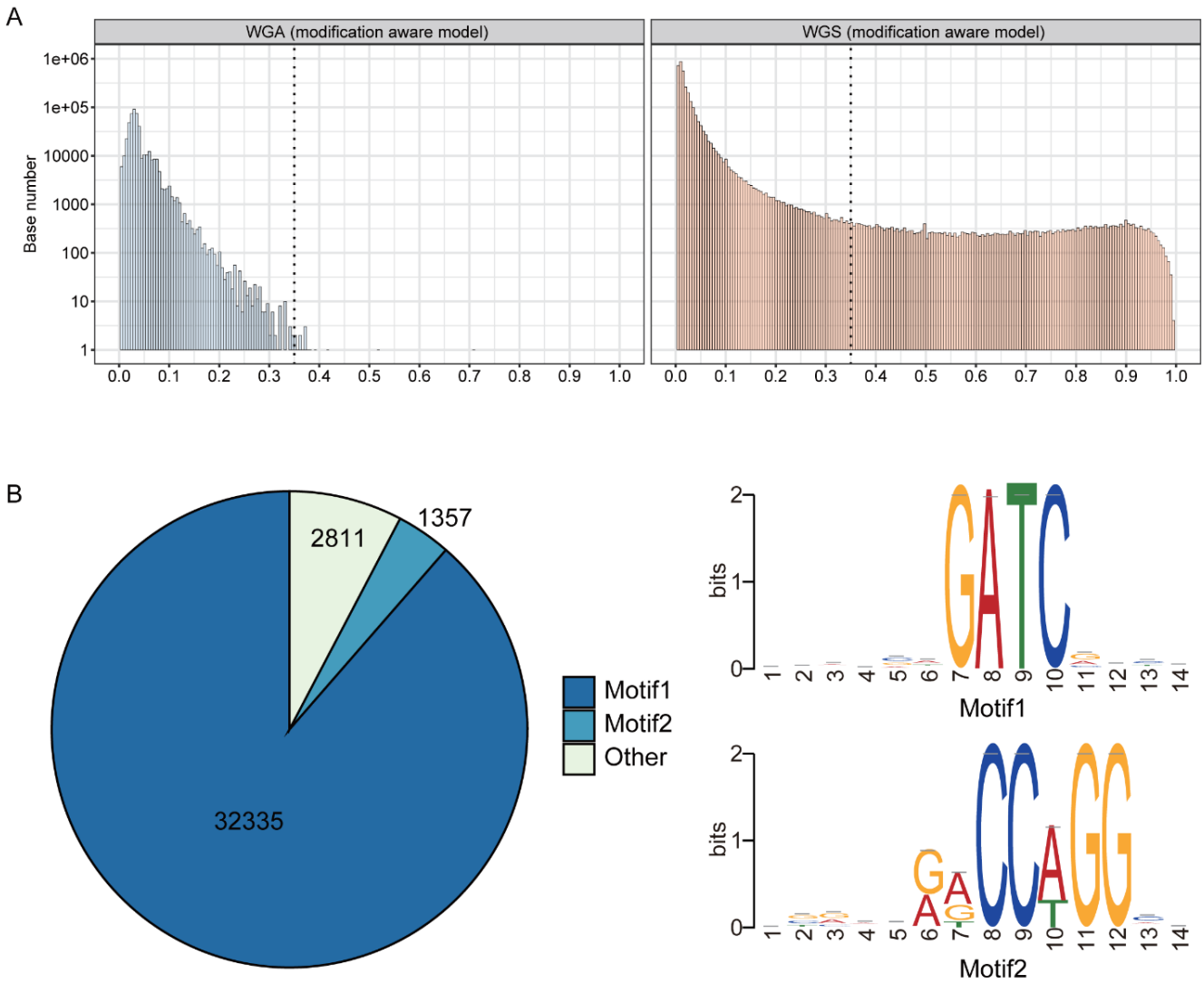

Figure S27. The ability of methylation identification using the Hammerhead with the "modification aware" model in our *E. coli* dataset. (A) The distribution of difference index of WGA reads and WGS reads. **(B)** The enriched motif for possible DNA modification sites identified by Hammerhead. Note: The GATC and CCWGG (CCTGG and CCAGG) motifs can be methylated by Dam and Dcm methylase, respectively.

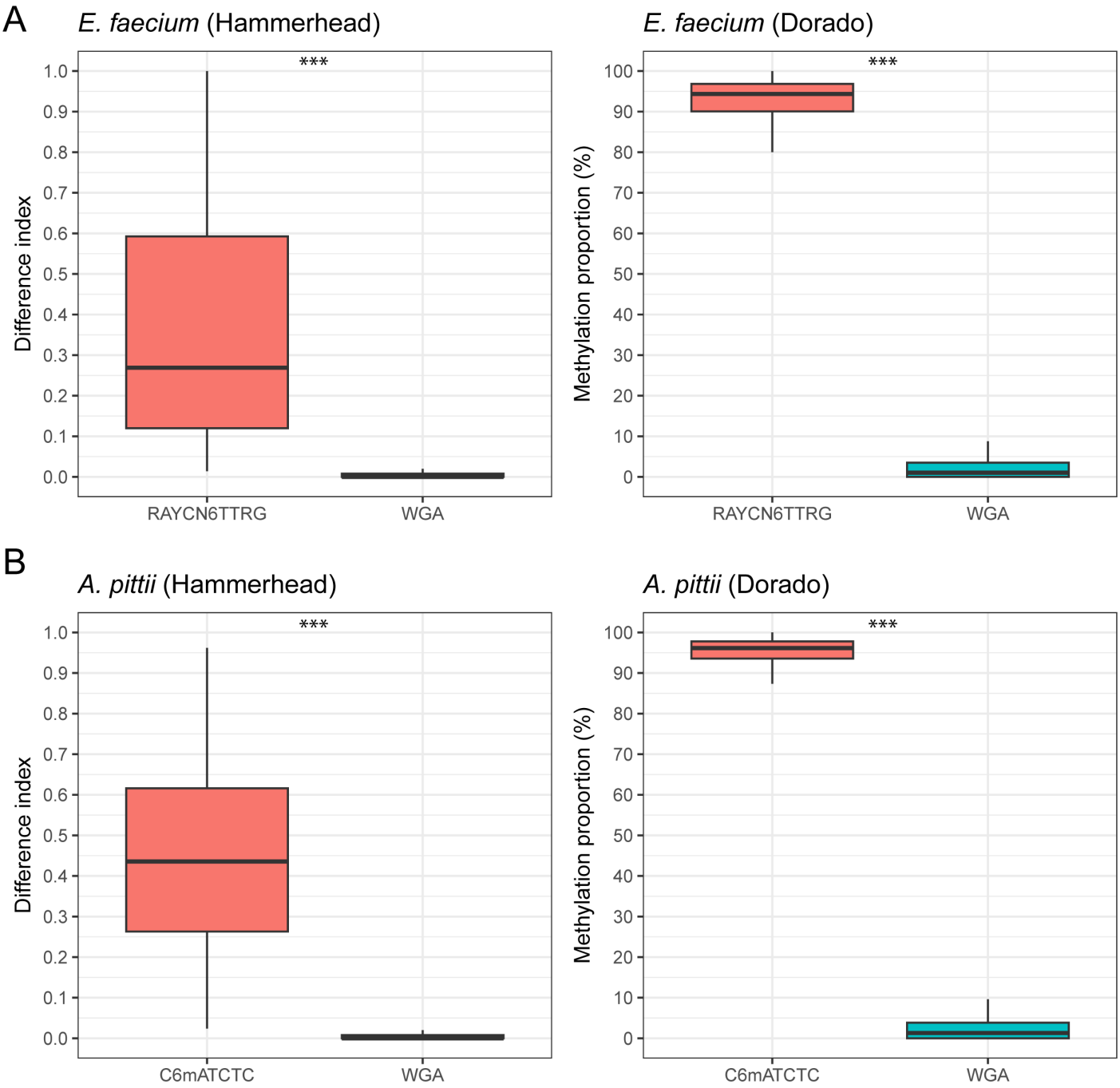

Figure S28. The distribution of the difference index (calculated by Hammerhead) and methylation p roportion (calculated by Dorado) between modified A sites in WGS reads and total sites in WGA r eads. *** p-value < 2.2e-16, Student's t-test.

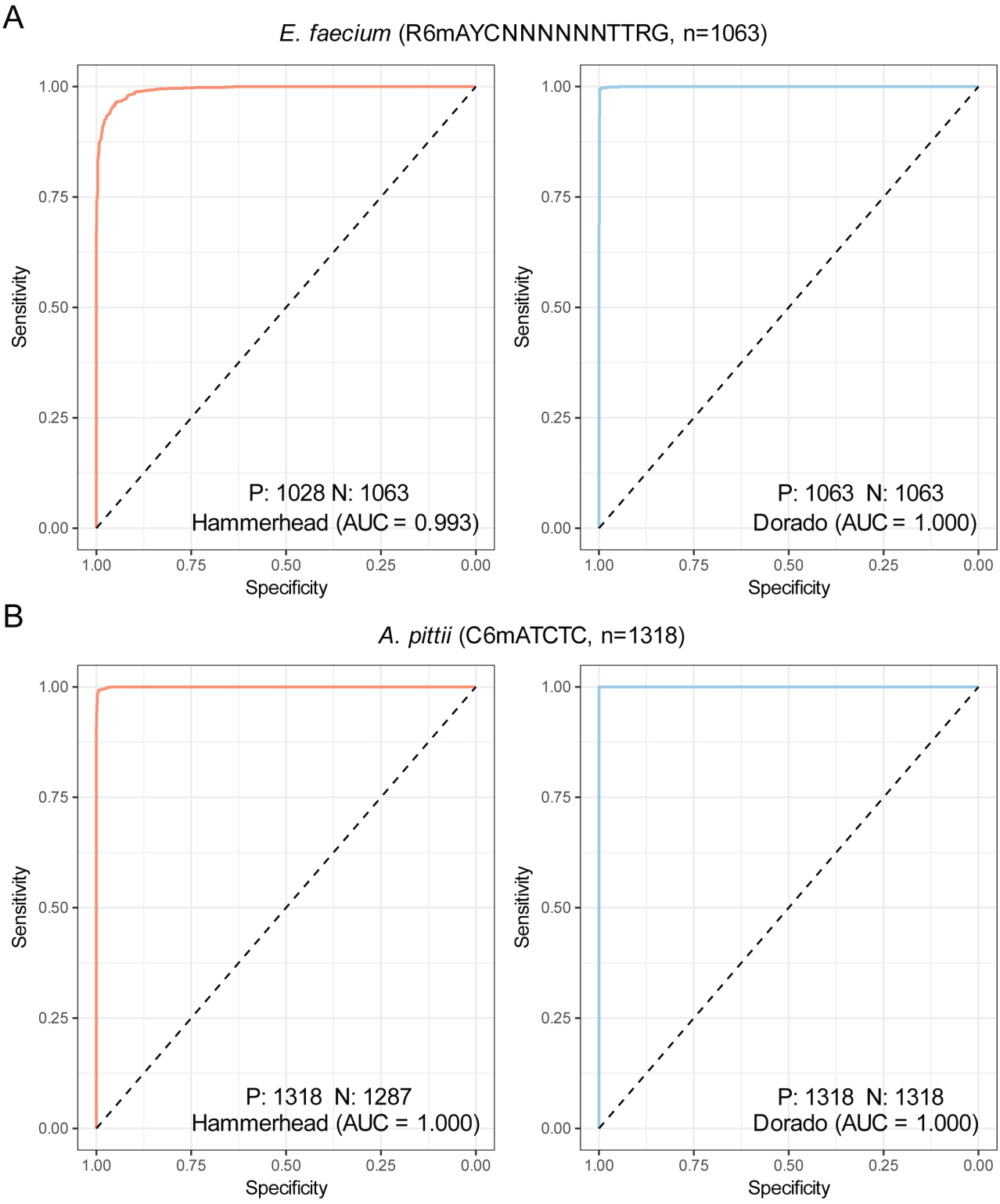

Figure S29. (A) and **(B)** The receiver operating characteristic (ROC) curve of Hammerhead and Dorado using 1063 6mA sites within type I motif RAYCNNNNNNTTRG in *E. faecium* and 1318 6mA sites within type II motif CATCTC in *A. pittii*. P and N mean the number of positives and negatives, respectively. AUC: area under the ROC curve.

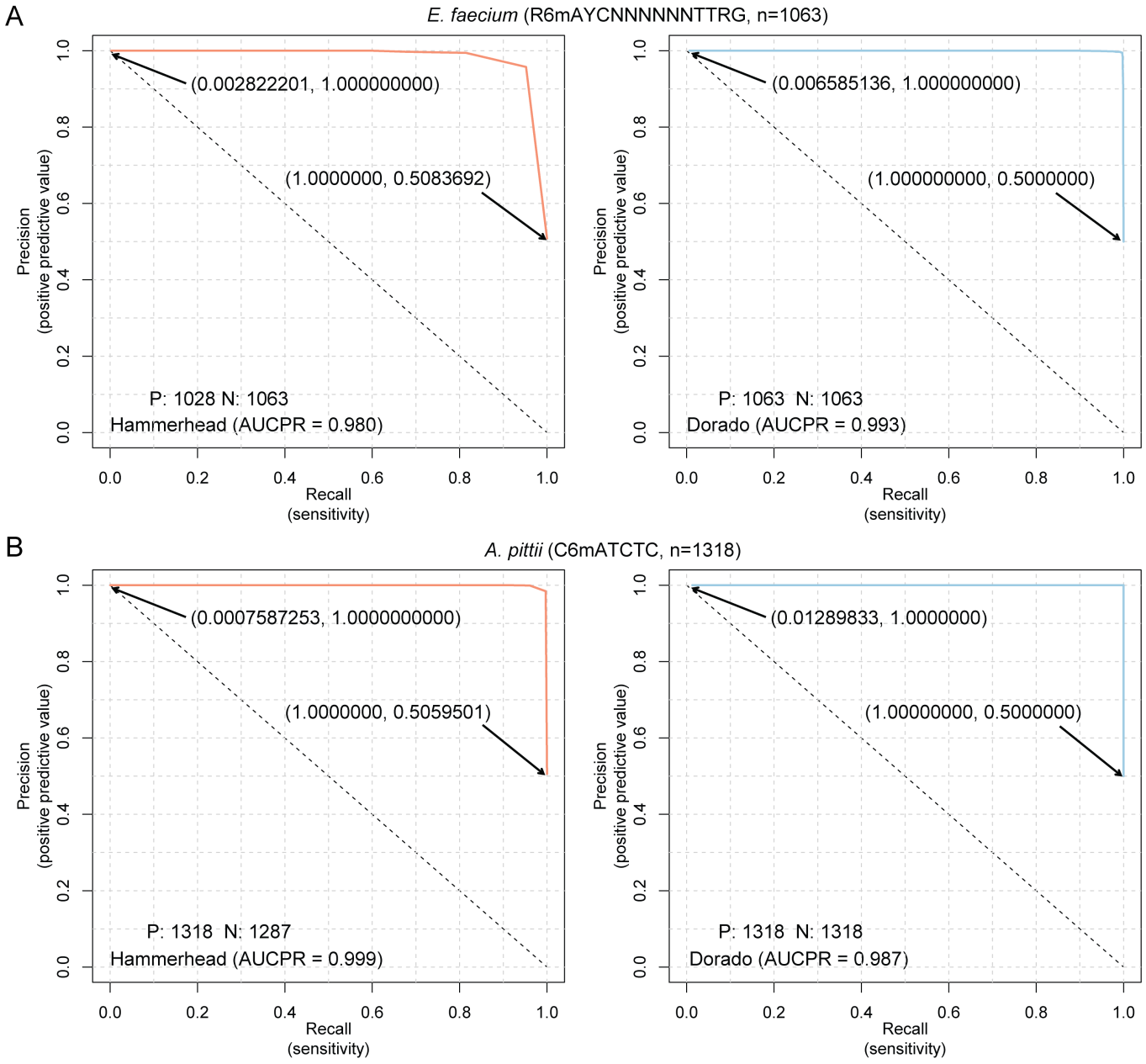

Figure S30. (A) and **(B)** The precision-recall (PR) curve of Hammerhead and Dorado using 1063 6mA sites within type I motif RAYCNNNNNNTTRG in *E. faecium* and 1318 6mA sites within type II motif CATCTC in *A. pittii*. P and N mean the number of positives and negatives, respectively. AUCPR: area under the precision-recall curve.

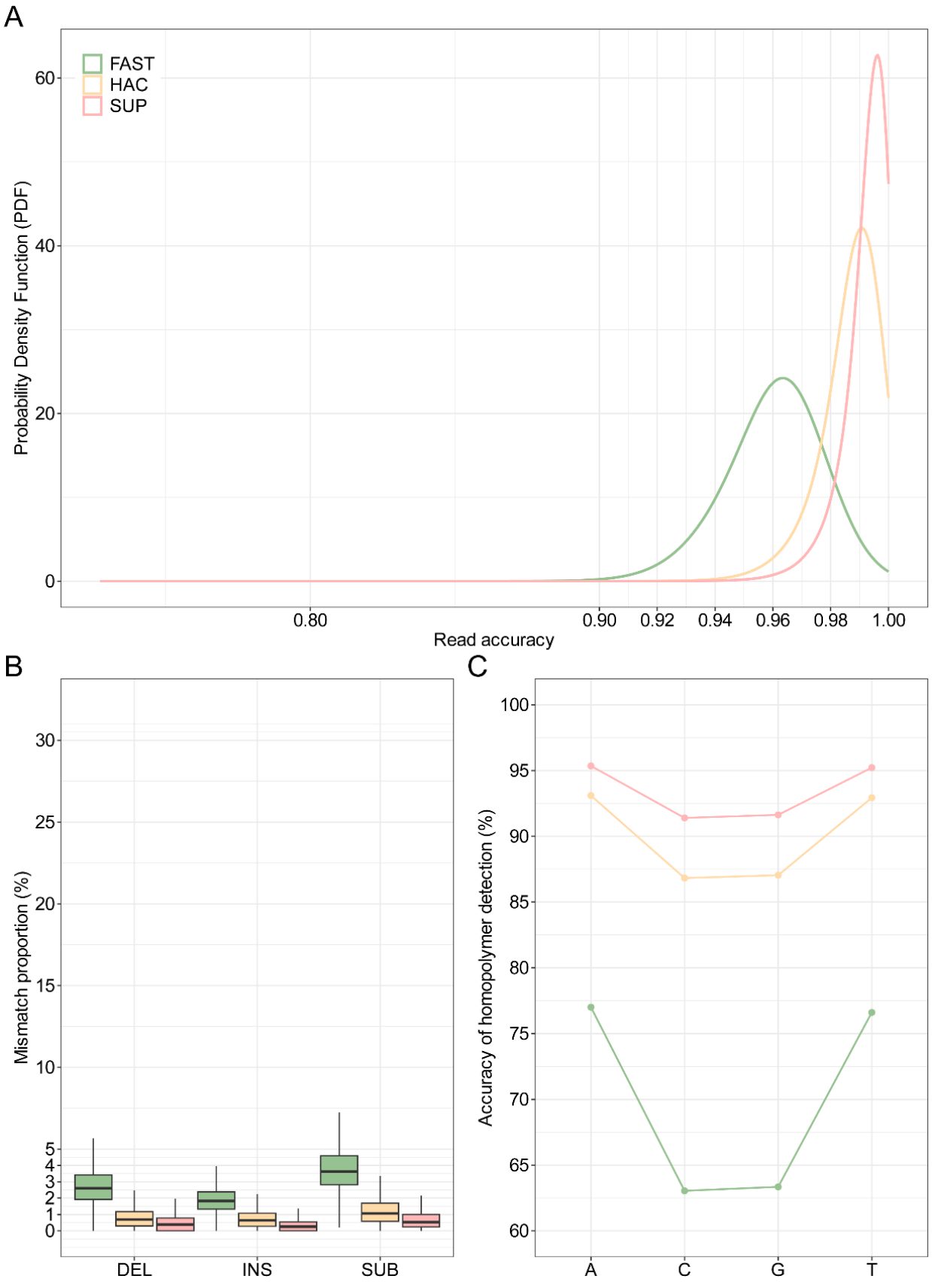

Figure S31. The read quality of WGS reads from *E. coli* using the SUP, HAC, and FAST models. **(A)** The density of read accuracy. **(B)** The proportion of three types of mismatches. **(C)** The accuracy of homopolymer identification. DEL: deletion, INS: insertion, SUB: substitution.

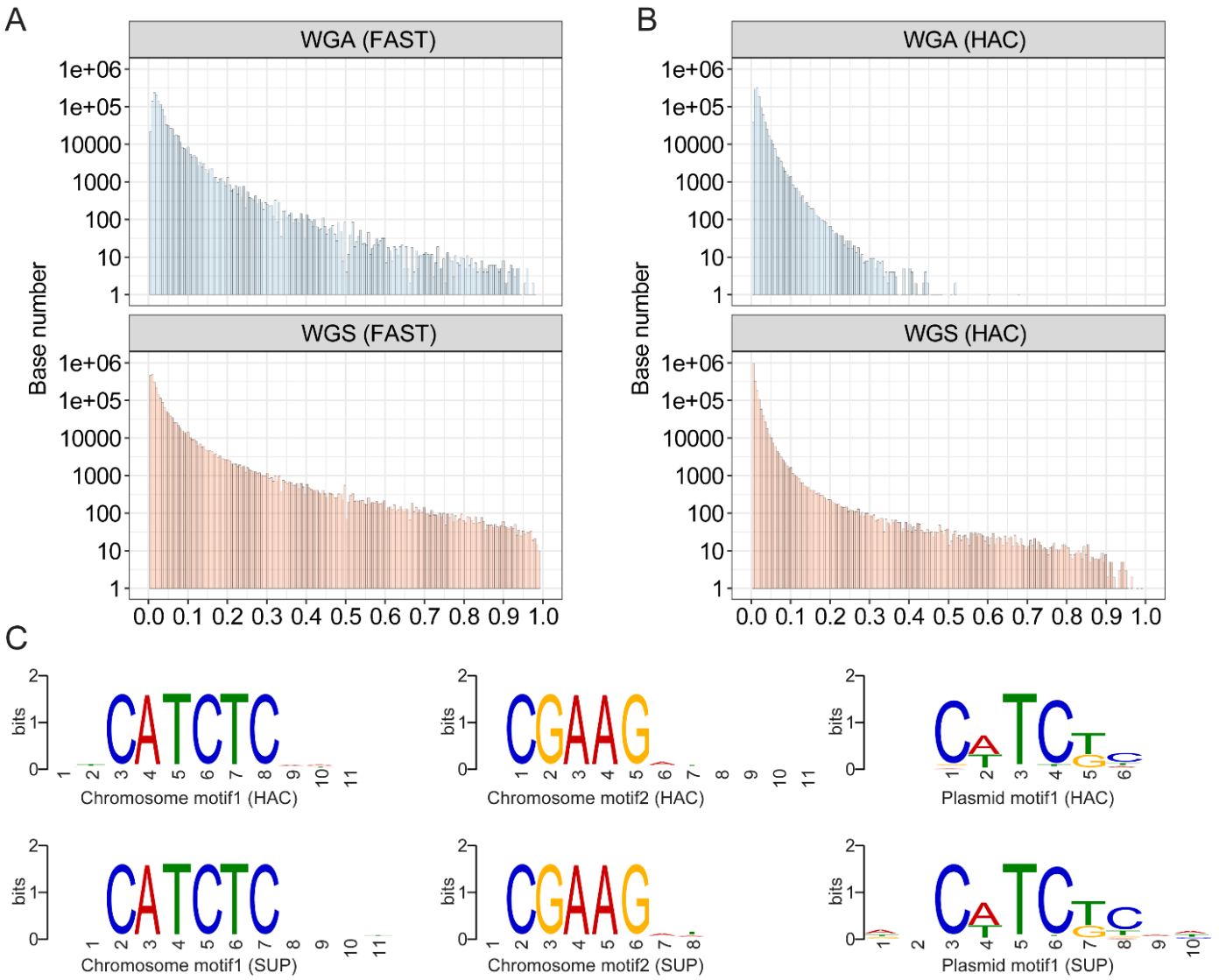

Figure S32. The ability of methylation identification using the Hammerhead with FAST and HAC model in the *A. pittii* dataset. (A) and **(B)** The distribution of the difference index of WGA reads and WGS reads for the FAST and HAC models, respectively. **(C)** The comparison of enriched motifs for possible DNA modification sites identified by Hammerhead between the HAC model and the SUP model. All E-values of motifs were less than 0.05.

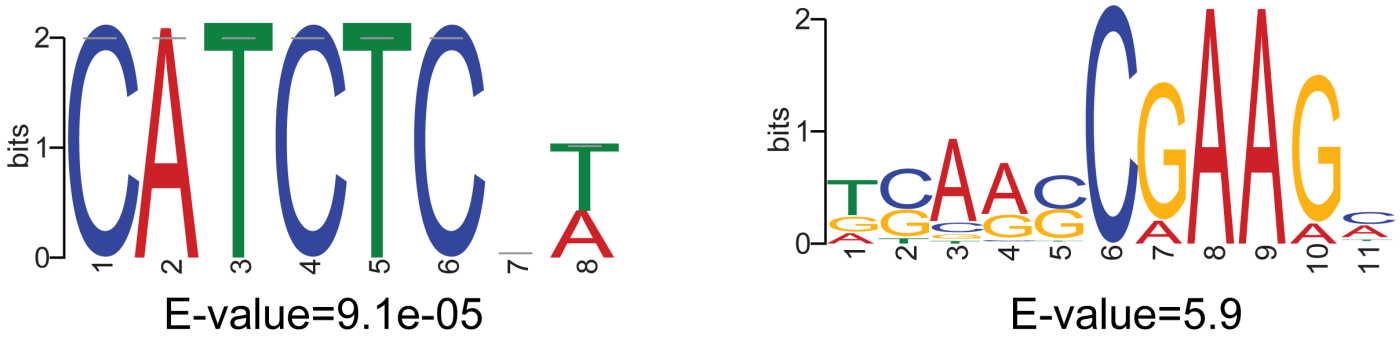

Figure S33. The motif was enriched using Hammerhead with the FAST model in the *A. pittii*. The MEME E-value less than 0.05 is significant.

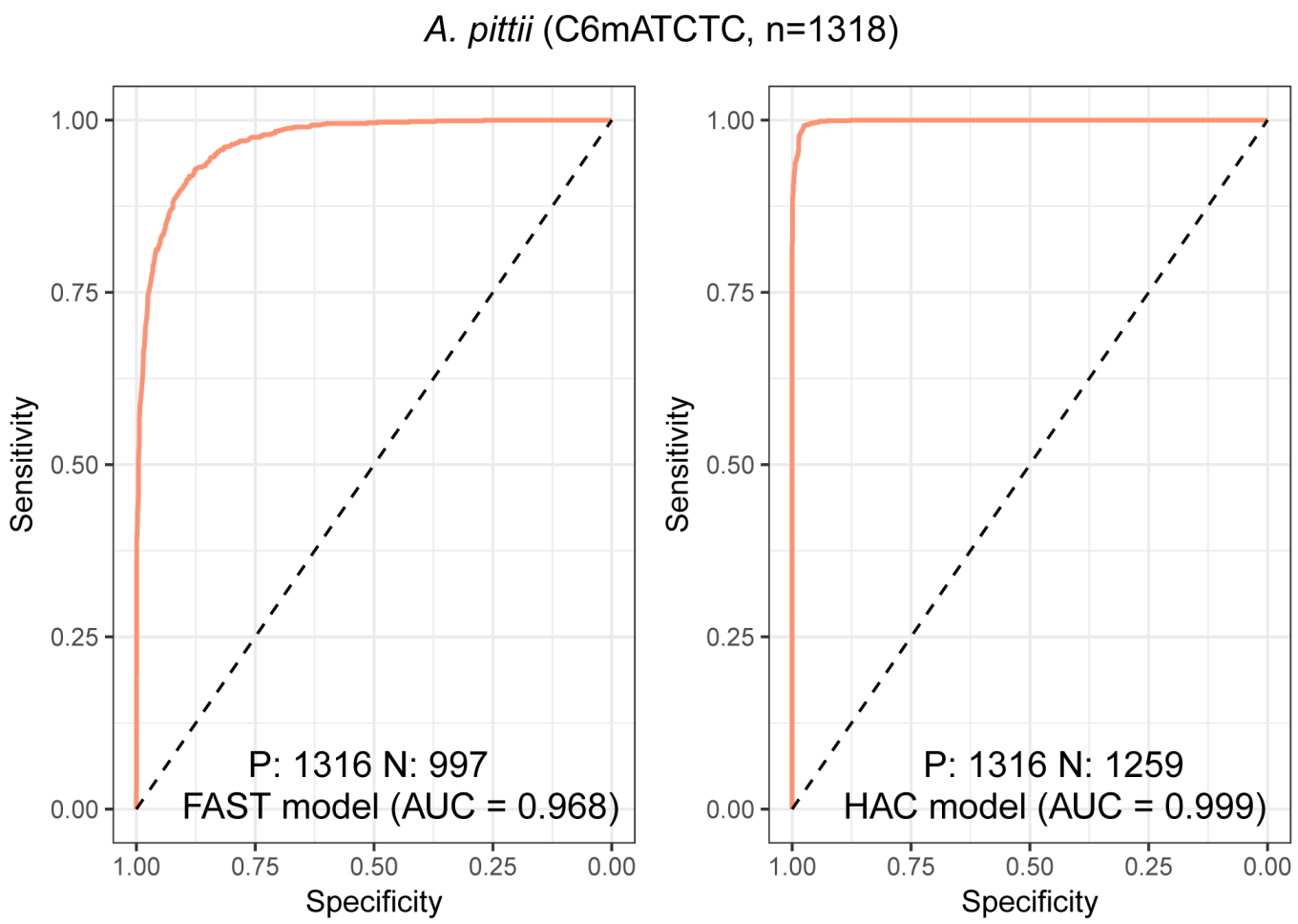

Figure S34. The receiver operating characteristic (ROC) curve of FAST, HAC, and SUP models using 1318 6mA sites within type II motif CATCTC in *A. pittii*. P and N mean the number of positives and negatives, respectively. AUC: area under the ROC curve.

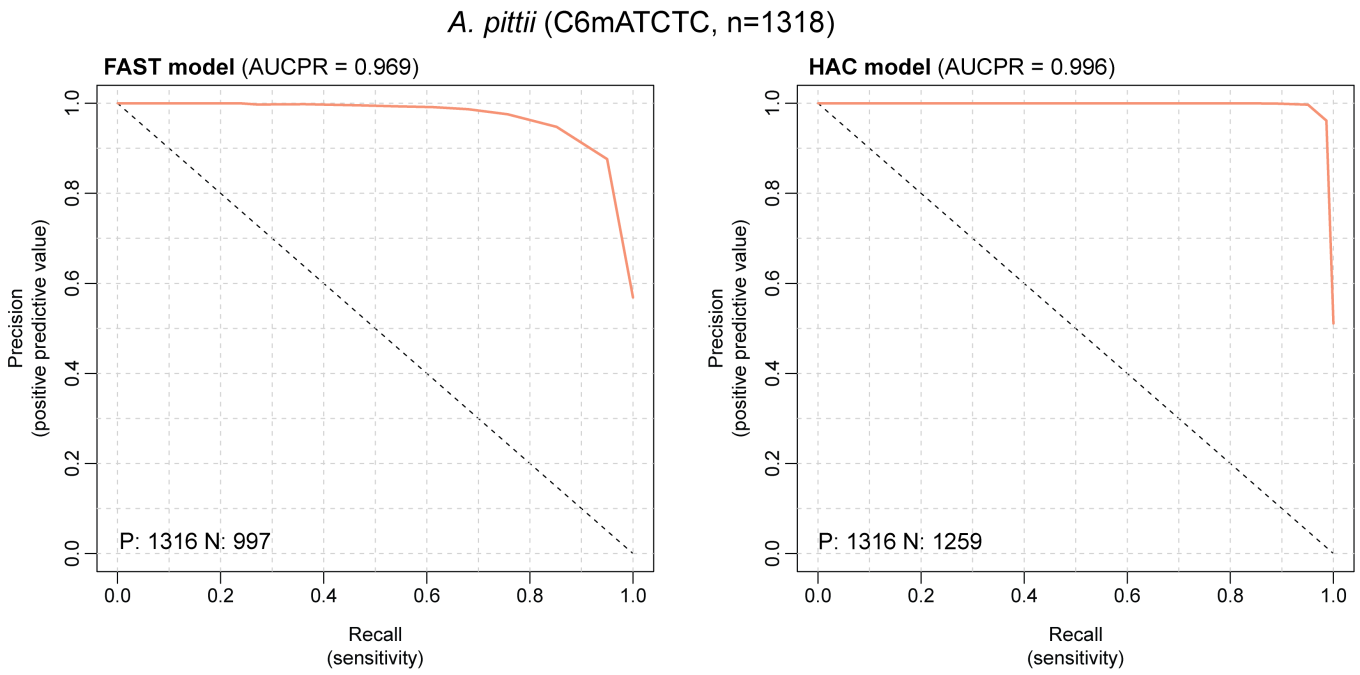

Figure S35. The precision-recall (PR) curve of FAST, HAC, and SUP models using 1318 6mA sites within type II motif CATCTC in *A. pittii*. P and N mean the number of positives and negatives, respectively. AUCPR: area under the precision-recall curve.

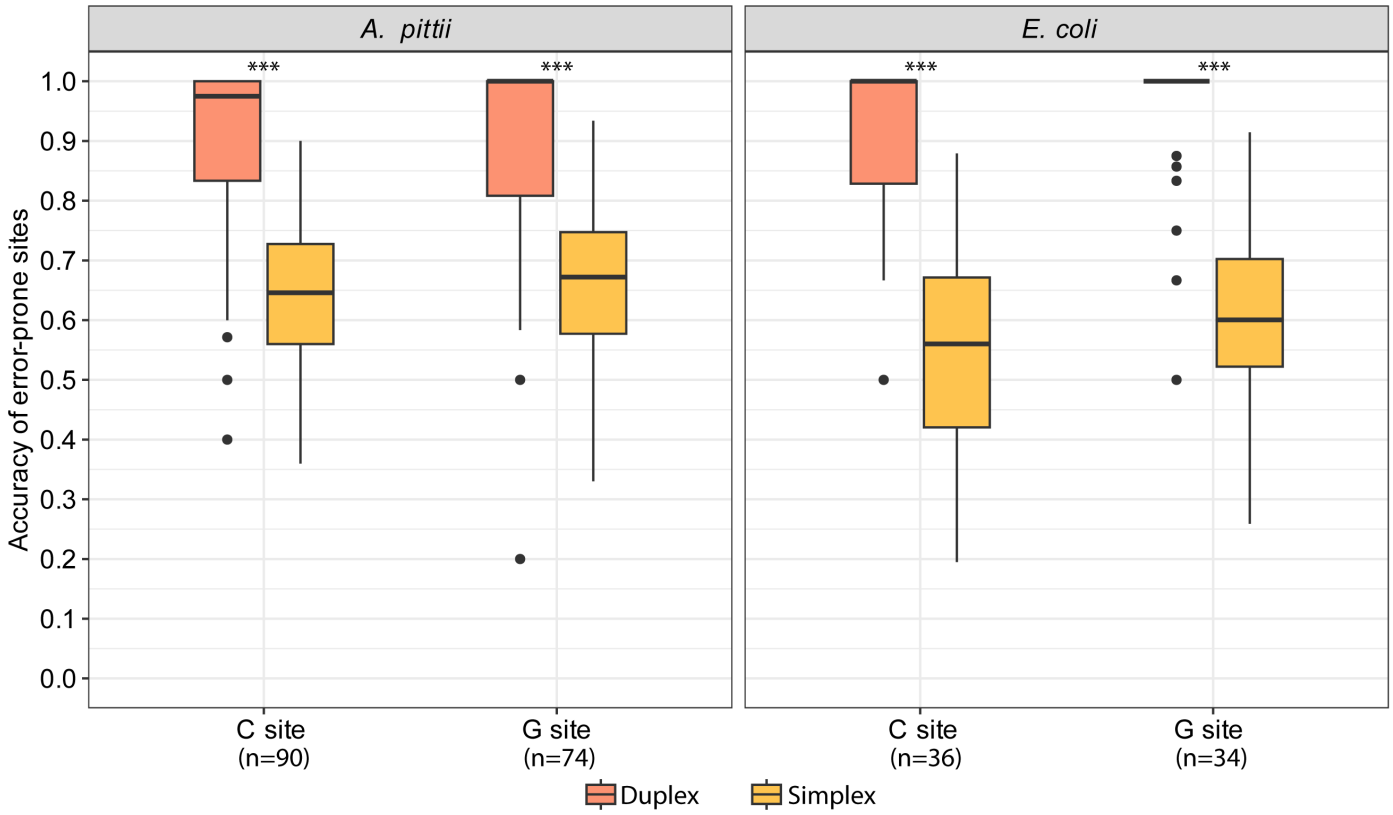

Figure S36. The accuracy of error-prone C and G sites in *A. pittii* and *E. coli* for R10.4.1 reads and Duplex reads. *** p-value < 1e-06, Student's t-test.

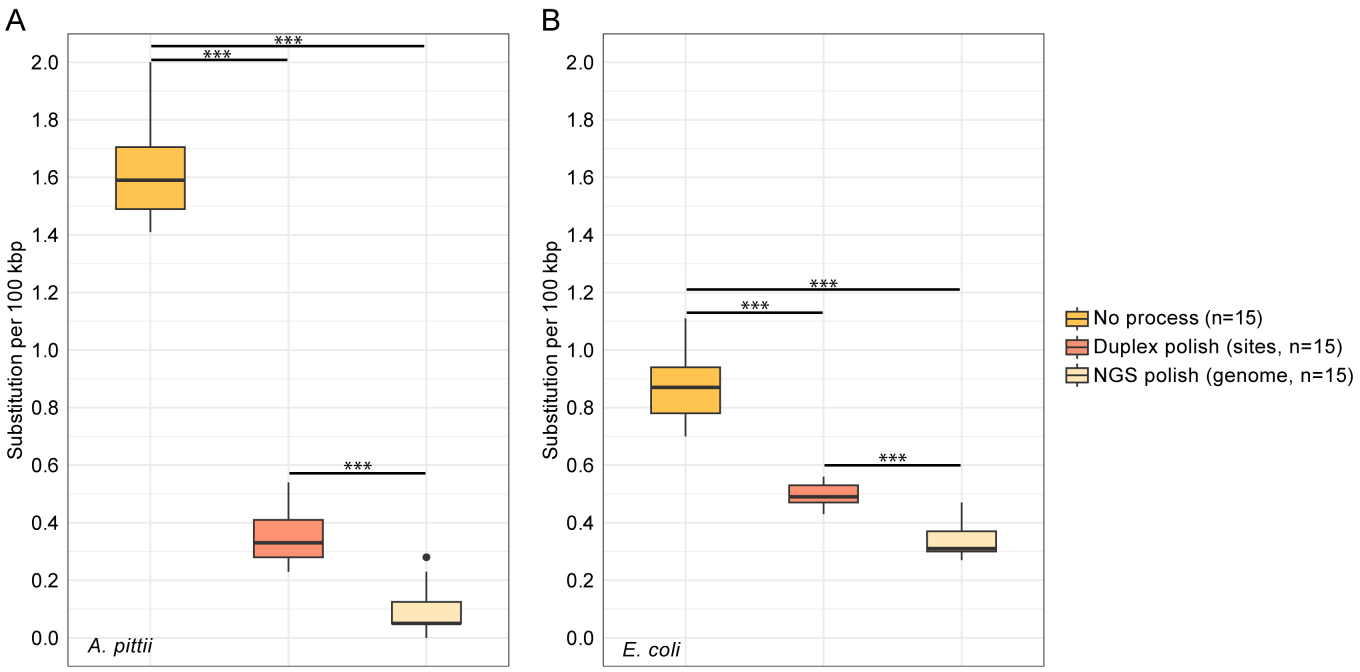

Figure S37. (A) and **(B)** Substitution per 100 Kbps of assemblies generated from R10.4.1 reads with three polish methods including control (no process), Duplex polish (only targeting the potential modification sites), and short-read polish (targeting the whole genome) in *A. pittii* and *E. coli*. *** p-value < 1e-06, Student's t-test.

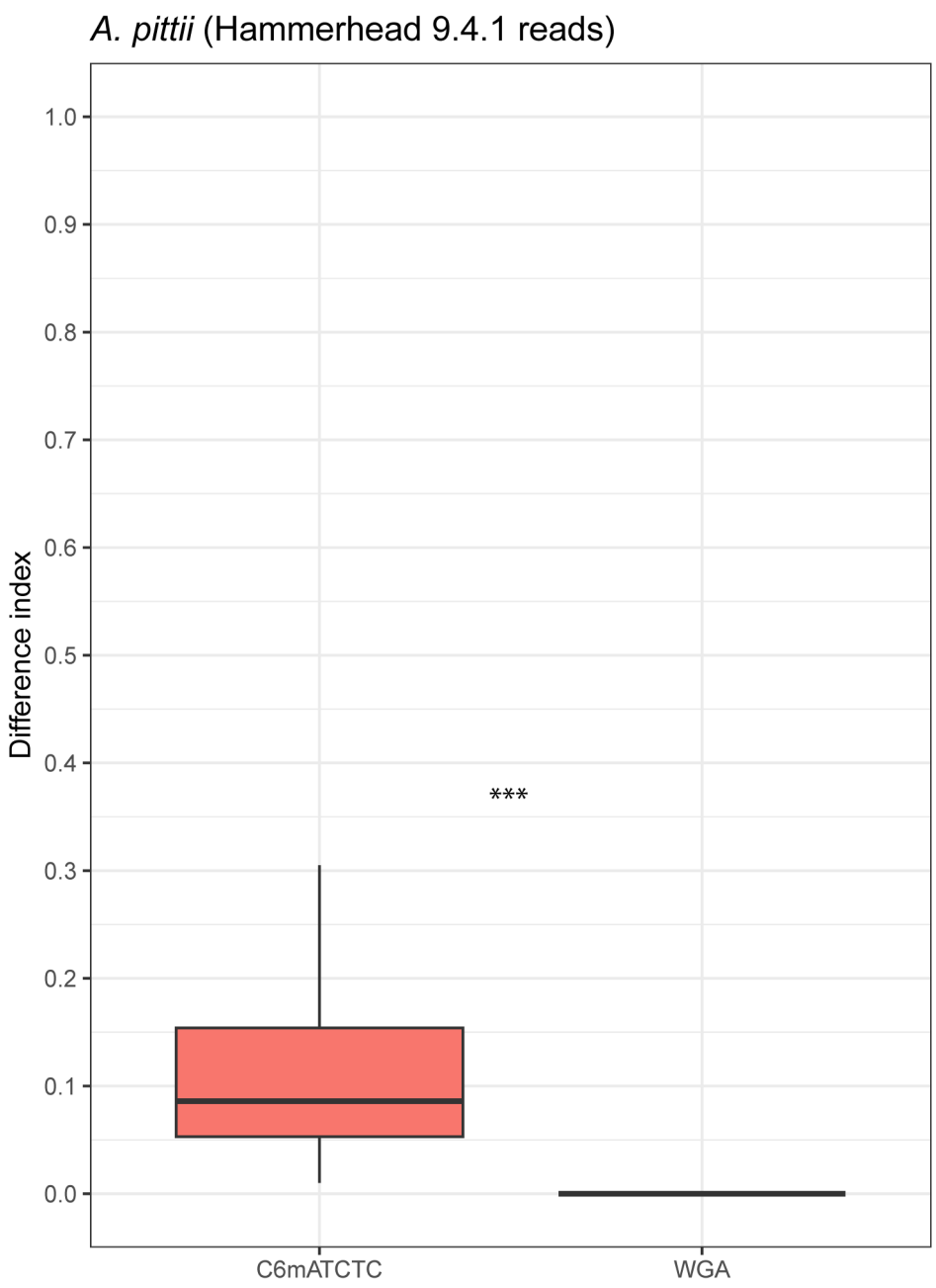

Figure S38. The distribution of the difference index (calculated by Hammerhead) between modified A sites in WGS reads and total sites in WGA reads using R9.4.1 reads. *** p-value < 2.2e-16, S tudent's t-test.