Strand-specific 5hmC methylated base detection using PacBio HiFi sequencing

Daniel M. Portik[1], Anupam K. Chakravarty[1], Aaron Wenger[1], and Guilherme de Sena Brandine[1]
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

PacBio HiFi sequencing observes a polymerase in real time as it incorporates fluorescently labeled nucleotides to synthesize a DNA strand. When the polymerase encounters methylated bases (e.g., chemical modifications to canonical DNA bases), the resulting kinetic signals are different from unmodified bases. The kinetic changes affect pulse widths and interpulse durations (Fig. 1). They occur at the methylated base and several surrounding bases, creating a unique kinetic signature.

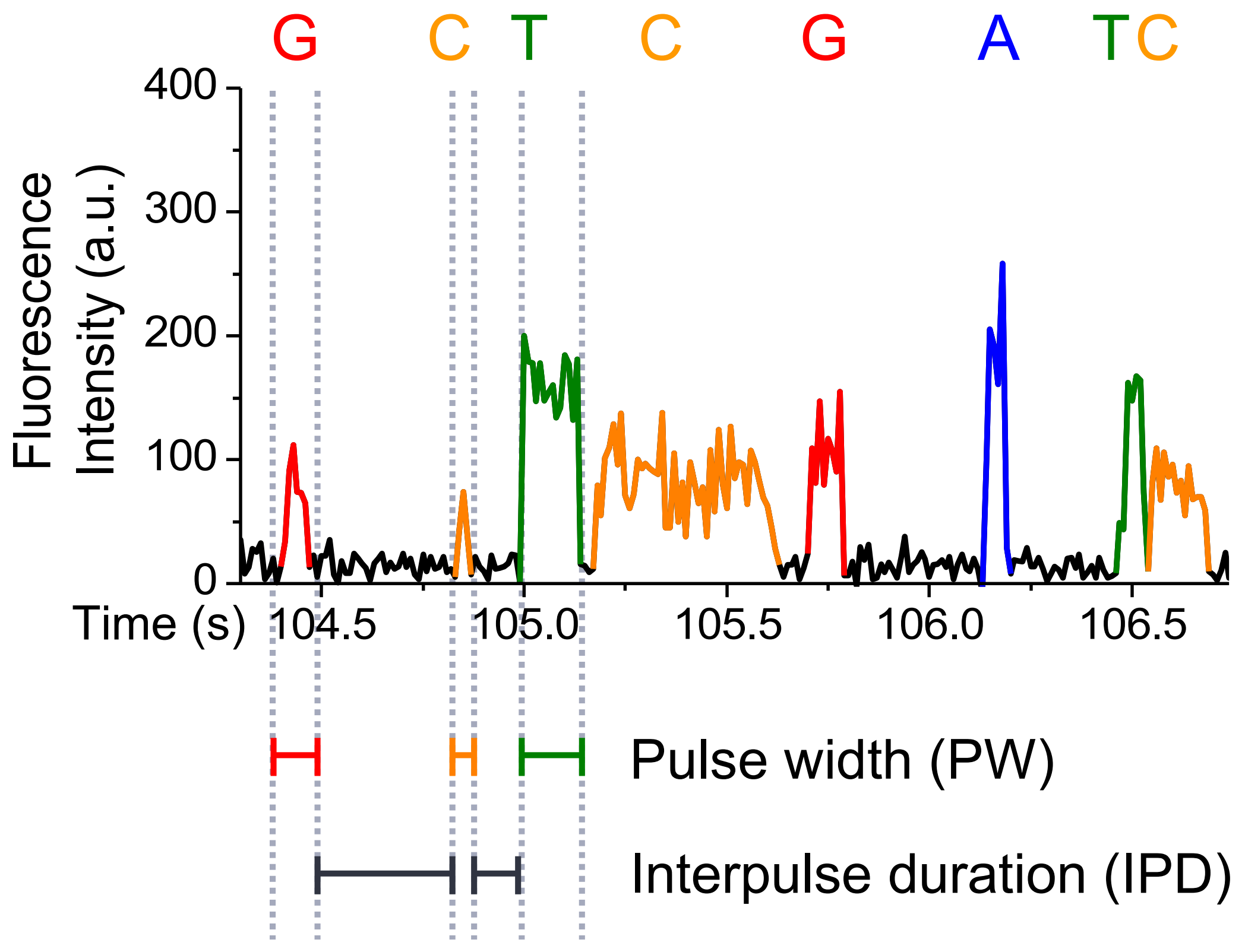


Figure 1. Example DNA trace showing pulse width (time of incorporation) and interpulse duration (time between adjacent incorporations). Image modified from Flusberg et al. (2010)[1].

PacBio sequencers currently offer single-molecule detection of 5-methylcytosine (**5mC**) and N6-methyladenine (**6mA**). Beyond these, 5-hydroxymethylcytosine (**5hmC**) is another modification which plays an important role in gene expression in vertebrates. It is enriched in neuronal and developmental tissues, where it contributes to cell-type-specific gene expression patterns[2,3]. The strand and haplotype context of 5hmC provides insight into allele-specific regulation and parent-of-origin effects, making phasing an important aspect of interpretation[4].

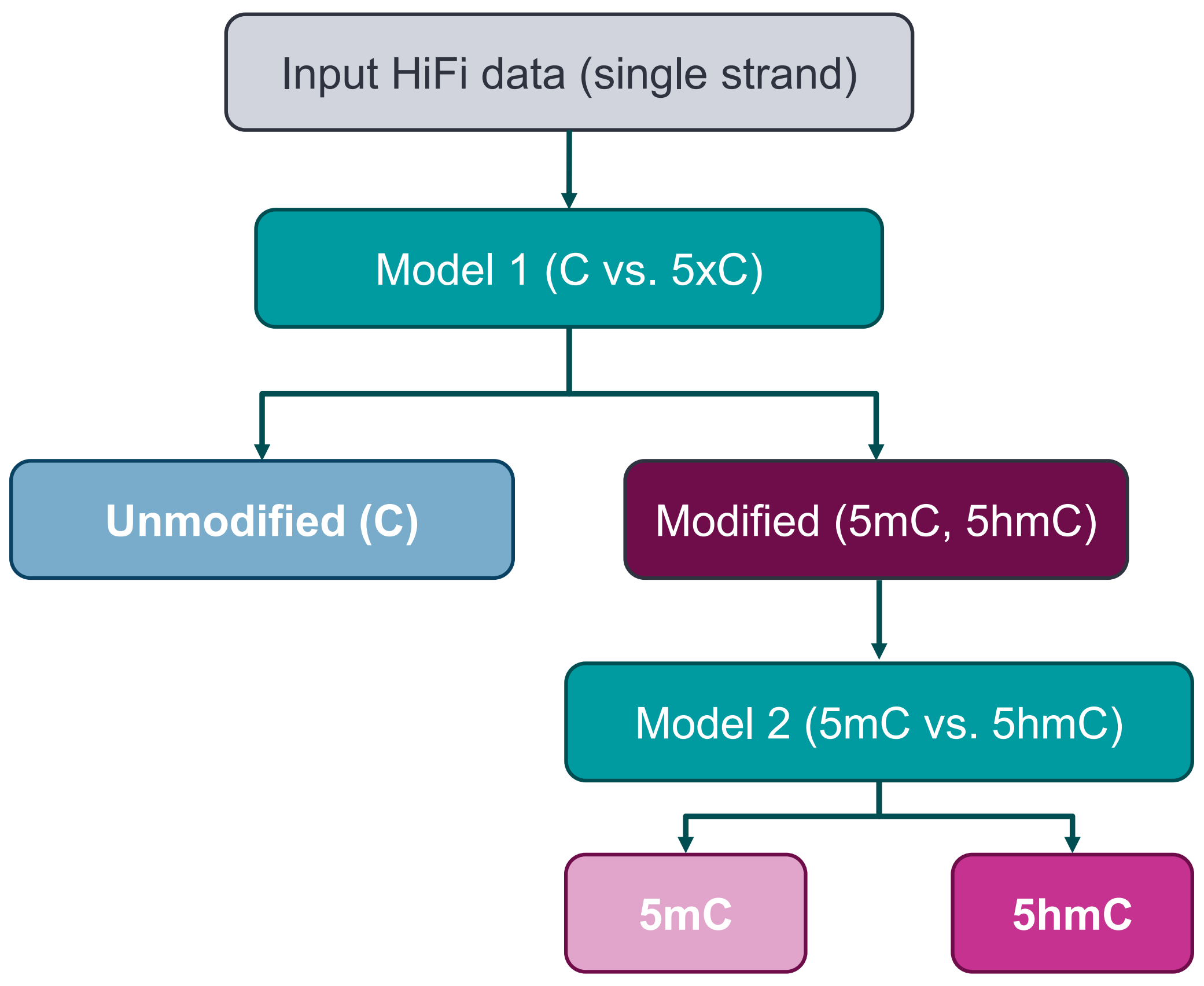
Here, we demonstrate an approach to detect strand-specific 5hmC in PacBio whole genome sequencing.

Methods

We sequenced three types of control data using SPRQ and SPRQ-Nx chemistry on the Revio system. Control datasets were comprised of unmodified and modified cytosines.

- **5mC**: created from methyltransferase-treated HG002 DNA (MSssI), modifying all C's in CpG contexts to 5mC
- **5hmC**: used an adapter-ligation process[5], which creates one 5hmCpG site joining two fragments of HG002 DNA
- **C**: generated using whole genome amplification of HG002, eliminating native 5mC sites

Strand-specific 5hmC workflow


Input HiFi data (single strand)

Model 1 (C vs. 5xC)

Unmodified (C)

Modified (5mC, 5hmC)

Model 2 (5mC vs. 5hmC)

5mC

5hmC

Figure 2. A visual overview of a single-strand 5hmC workflow. Here, 5xC refers to the combination of 5mC and 5hmC sites.

- To call 5hmC in CpG contexts, we implemented a two-step approach based on deep-learning models composed of convolutional neural networks with transformer layers[5] (Fig. 2).
- Feature sets included strand-specific kinetics and sequences, with added normalizations.
- We used several combinations of training datasets to investigate the effects on model performance.
- Model evaluations were performed on non-overlapping control datasets.
- We investigated the effects of setting minimum probability thresholds for each workflow step.

Results

- We accurately detect single-strand 5hmC using a variety of model training conditions and post-analysis filters.
- Over 50% of the training and filtering conditions display <1% false positive rate and >50% recall (Fig. 3).

5hmC false positive rate vs. recall

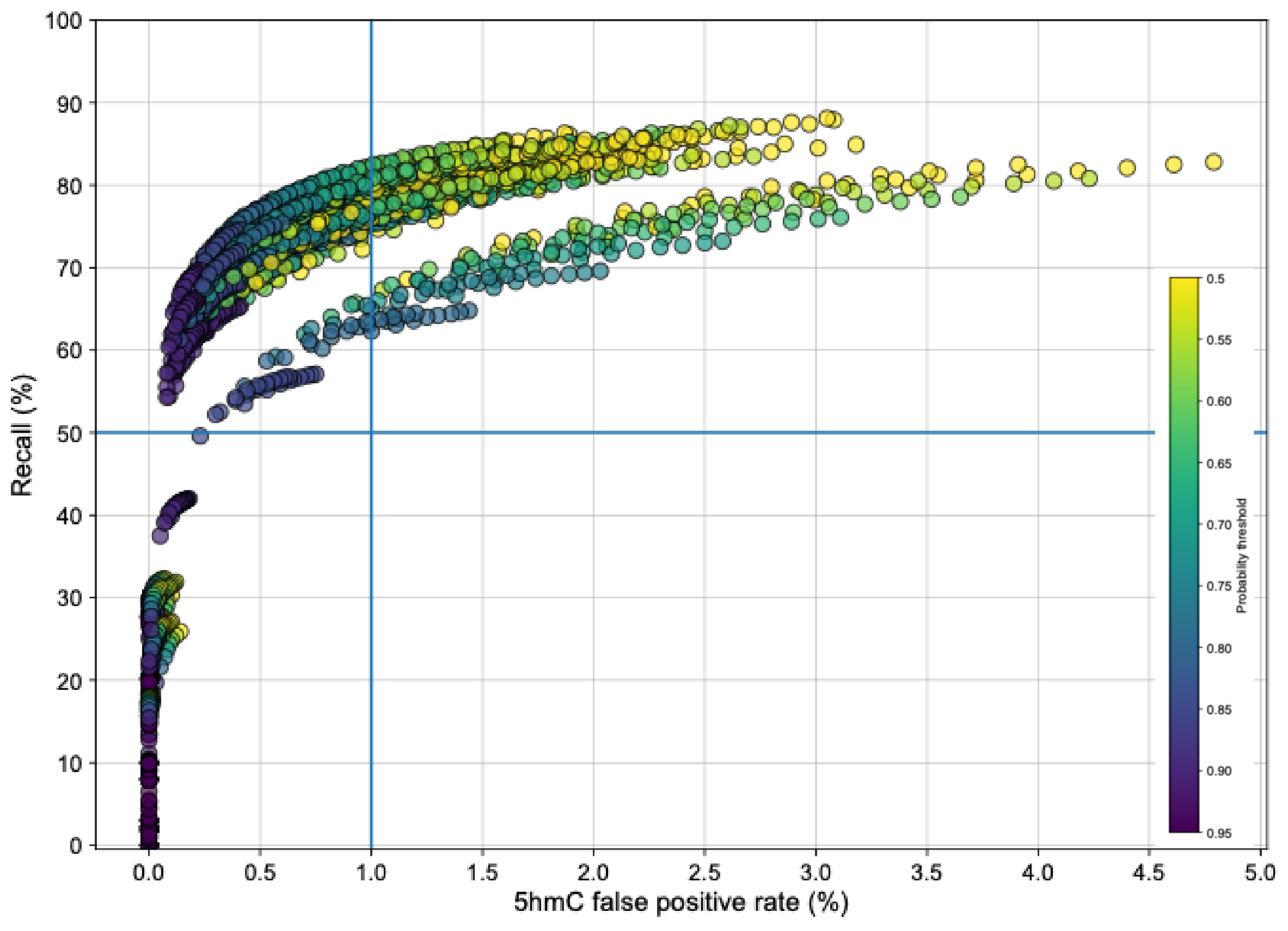


Figure 3. Scatterplot of 5hmC false positive rate vs. recall. Each point represents a unique model training strategy and combination of probability filters. Yellow color represents the lowest probability thresholds whereas darker colors represent higher values..

- We examined the relationship between 5hmC recall and false positive rates in our preferred model combination (Fig. 4).
- At ~0.1% FPR, we achieve 60-65% recall.
- At ~1% FPR, we achieve 80-82% recall.

Best recall across false positive rates

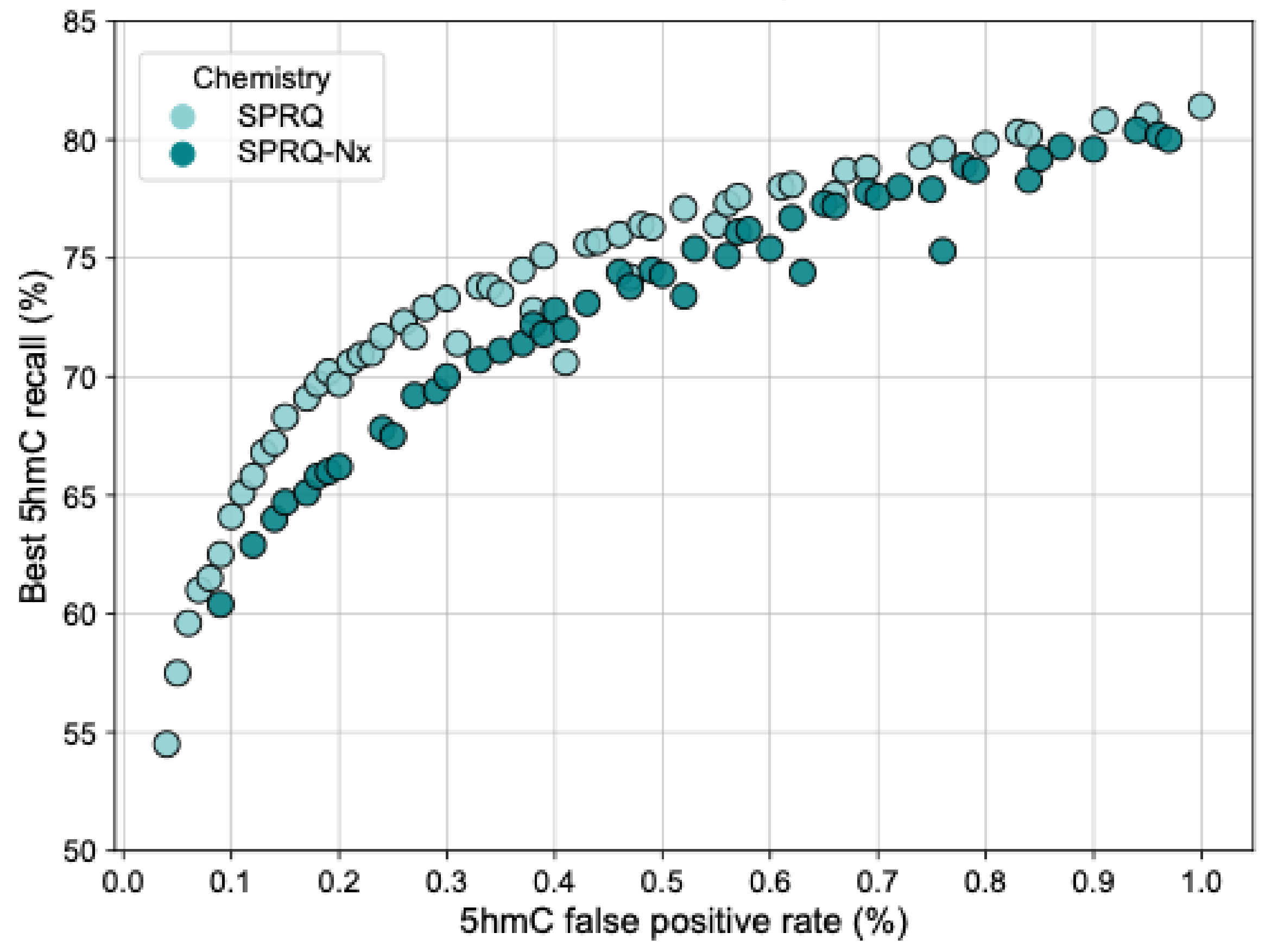
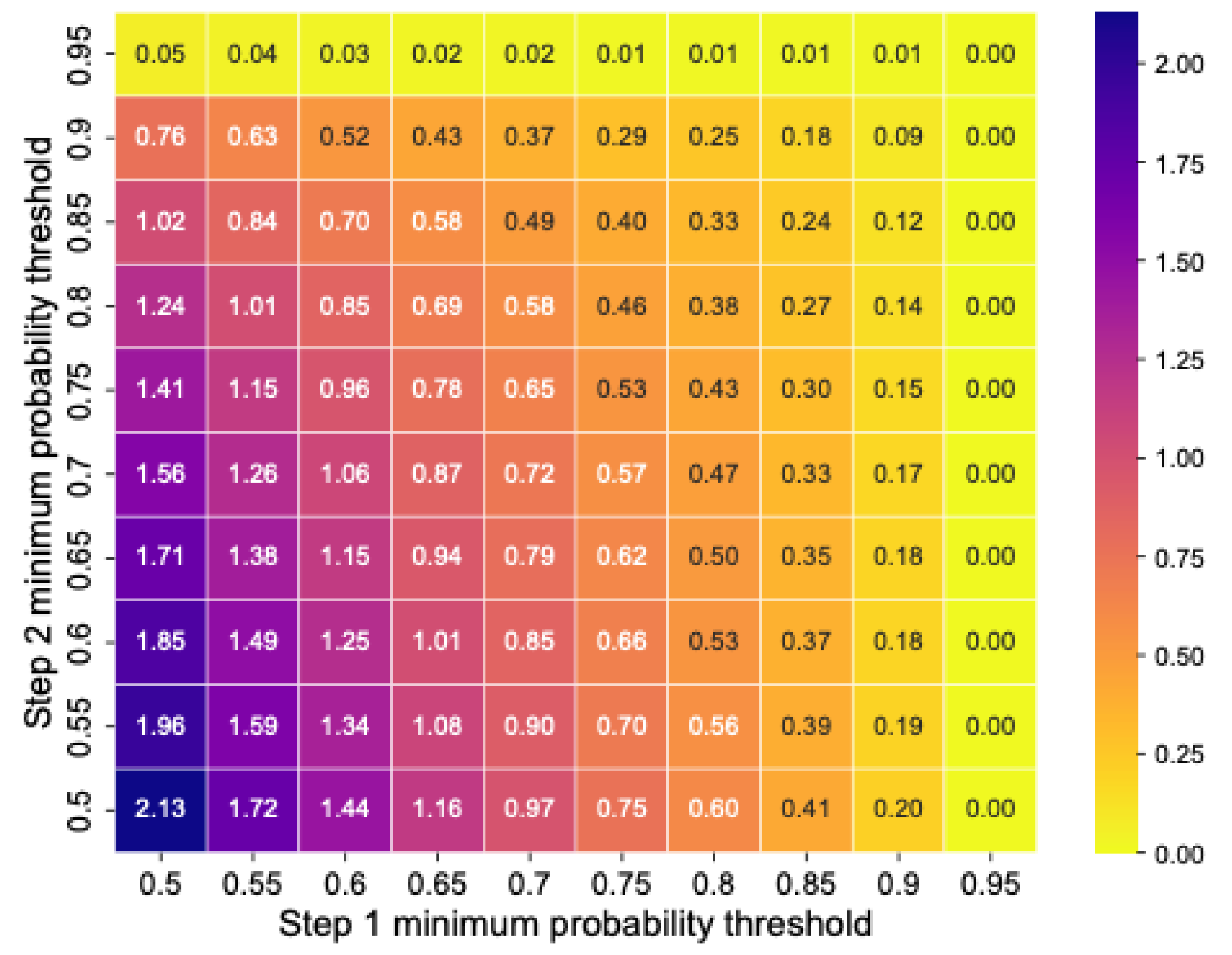


Figure 4. Scatterplot of top recall scores across false positive rates from our preferred model combination. Light blue = SPRQ chemistry, dark blue = SPRQ-NX.

- False positive rates can be controlled using probability thresholds for each step (Fig. 5).
- The reported 5hmC probabilities (e.g., step 2 probabilities) can be filtered to adjust the false positive rate for specific applications.

5hmC false positive rate heatmap



Step 2 minimum probability threshold	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
0.95	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.00
0.9	0.76	0.63	0.52	0.43	0.37	0.29	0.25	0.18	0.09	0.00
0.85	1.02	0.84	0.70	0.58	0.49	0.40	0.33	0.24	0.12	0.00
0.8	1.24	1.01	0.85	0.69	0.58	0.46	0.38	0.27	0.14	0.00
0.75	1.41	1.15	0.96	0.78	0.65	0.53	0.43	0.30	0.15	0.00
0.7	1.56	1.26	1.06	0.87	0.72	0.57	0.47	0.33	0.17	0.00
0.65	1.71	1.38	1.15	0.94	0.79	0.62	0.50	0.35	0.18	0.00
0.6	1.85	1.49	1.25	1.01	0.85	0.66	0.53	0.37	0.18	0.00
0.55	1.96	1.59	1.34	1.08	0.90	0.70	0.56	0.39	0.19	0.00
0.5	2.13	1.72	1.44	1.16	0.97	0.75	0.60	0.41	0.20	0.00
Step 1 minimum probability threshold	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95

Figure 5. Heatmap showing false positive rates across different combinations of minimum probability thresholds. Yellow represents the lowest FPR whereas darker colors represent higher FPR.

Conclusions

We demonstrate the ability to accurately detect 5hmC in CpG contexts with HiFi sequencing. This base modification can be detected in standard PacBio sequencing libraries without the need for any additional chemical treatment.

References

1. Flusberg et al. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7: 461–465.
2. Cui et al. 2020. A human tissue map of 5-hydroxymethylcytosinesexhibits tissue specificity through gene andenhancer modulation. *Nature Communications*, 11: 6161.
3. He et al. 2021. Tissue-specific 5-hydroxymethylcytosinelandscape of the human genome. *Nature Communications*, 12: 4249.
4. Fu et al. 2025. Computational analysis of DNA methylation from long-read sequencing. *Nature Reviews Genetics*, 26: 620-634.
5. Hu et al. 2025. Transformer-based deep learning for accurate detection of multiple base modifications using single molecule real-time sequencing. *Communications Biology*, 8: 606.