Systematic optimisation of modkit parameters enables accurate nanopore-based RNA modification detection

Alexandra Sneddon[1,2,3,4,*], Stefan Prodic[1,2,3], Eduardo Eyras[1,2,3,4]

1 EMBL Australia Partner Laboratory Network at the Australian National University, Canberra ACT 2601, Australia.
2 Centre for Computational Biomedical Sciences, The John Curtin School of Medical Research, Australian National University, Canberra ACT 2601, Australia.
3 The Shine-Dalgarno Centre for RNA Innovation, The John Curtin School of Medical Research, Australian National University, Canberra ACT 2601, Australia.
4 The Australian Research Council Centre of Excellence for the Mathematical Analysis of Cellular Systems, Australia.

*Correspondence to: Alexandra.Sneddon@anu.edu.au

Abstract

Nanopore direct RNA sequencing enables single-molecule detection of RNA modifications, but accurate quantification depends on modkit, which employs a heuristic and frequently suboptimal strategy for filtering erroneous modification calls. We introduce ModkitOpt, a method that systematically scans modkit parameters and stoichiometry cutoffs, evaluates precision and recall against reference sites, and identifies optimal settings. Across mammalian N6-methyladenosine and pseudouridine datasets, ModkitOpt dramatically outperforms modkit's default and provides a robust, context-independent optimisation strategy that is released as an efficient Nextflow workflow at https://github.com/comprna/modkitopt.

Main

Nanopore sequencing enables direct, single-molecule detection of native RNA modifications at single-nucleotide resolution[1]. Oxford Nanopore Technologies (ONT) provides dorado[2] to call modified bases per read, and modkit[3] to aggregate per-read predictions into site-level stoichiometry, defined as the fraction of transcripts that are modified at a given reference position. In downstream analyses, stoichiometry cutoffs are applied to classify sites as modified or unmodified, thereby necessitating accurate stoichiometry estimation and thresholding for correct biological interpretation. However, modkit's default parameters are estimated heuristically from the data, without assessing their impact on site-level accuracy, and no quantitative approach exists to identify the optimal stoichiometry cutoff.

To estimate stoichiometry, modkit *pileup* filters out dorado's low-confidence modification calls. By default, modkit estimates the confidence threshold by considering modification calls for a subset of reads, defining confidence as the maximum predicted probability (modified or non-modified) for each call, and then setting the threshold at the 10th percentile of the confidence values across all calls. Since this heuristic is not based on prediction accuracy, datasets

dominated by low-confidence calls, such as those for rare modifications like pseudouridine (pseU), get assigned insufficiently stringent thresholds, resulting in elevated false discovery rates[4], while datasets dominated by high-confidence calls are filtered too stringently, excluding true sites. Although users can manually specify confidence thresholds for canonical and modified bases (*--filter-threshold* and *--mod-threshold,* respectively), this relies on subjective interpretation of probability distributions output by modkit's *sample-probs*, and no assessment of parameter performance is provided.

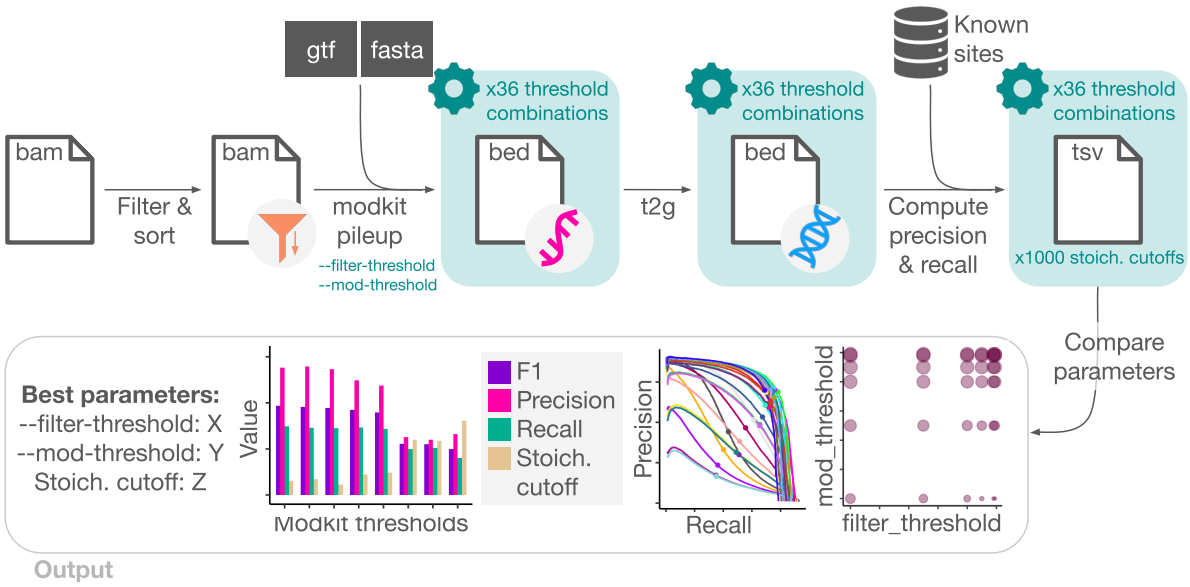

Figure 1. Overview of ModkitOpt. A transcriptome-aligned modBAM file containing per-read modification calls is filtered and sorted, and then processed with modkit *pileup* across a grid of filtering parameters (*--filter-threshold* and *--mod-threshold*) to generate site-level stoichiometry estimates. Transcript sites are converted to genomic coordinates and compared against reference modification sites across 1,000 stoichiometry cutoffs. Precision, recall and F1 score are computed for each parameter combination, and the thresholds and stoichiometry cutoff that maximise F1 score are identified and reported.

Here, we introduce ModkitOpt (**Figure 1**), a computational method to identify the optimal modkit filtering parameters (*--filter-threshold* and *--mod-threshold*) and downstream stoichiometry cutoff for any given nanopore dataset. Taking as input a modBAM file produced by dorado, ModkitOpt systematically scans all combinations of 36 candidate modkit thresholds, and 1,000 candidate stoichiometry cutoffs, and quantifies the precision and recall of predicted sites against a set of known reference sites. For each threshold combination, the stoichiometry cutoff that maximises F1 score is selected for reporting, as described below.

ModkitOpt reports the optimal thresholds and stoichiometry cutoff to the user, and outputs plots that visualise modkit's performance in terms of precision (fraction of predicted sites that are in the reference set), recall (fraction of reference sites that are predicted) and F1 score ($2 \cdot \frac{(Pecision \cdot Recall)}{(Pecision + Recall)}$) across the parameter space, displaying for each threshold combination the stoichiometry cutoff that maximised F1 score. Reference site sets are provided for mammalian m6A and pseU, derived from GLORI-identified m6A sites in HEK293T cells[5] and BACS-identified

pseU sites in HeLa cells[6], respectively. The user can also provide reference sites for other modifications, supporting the broad applicability of ModkitOpt. ModkitOpt is distributed as a Nextflow workflow, enabling reproducible execution on local machines or HPC systems, with resource allocation automatically managed for efficient and scalable analysis.

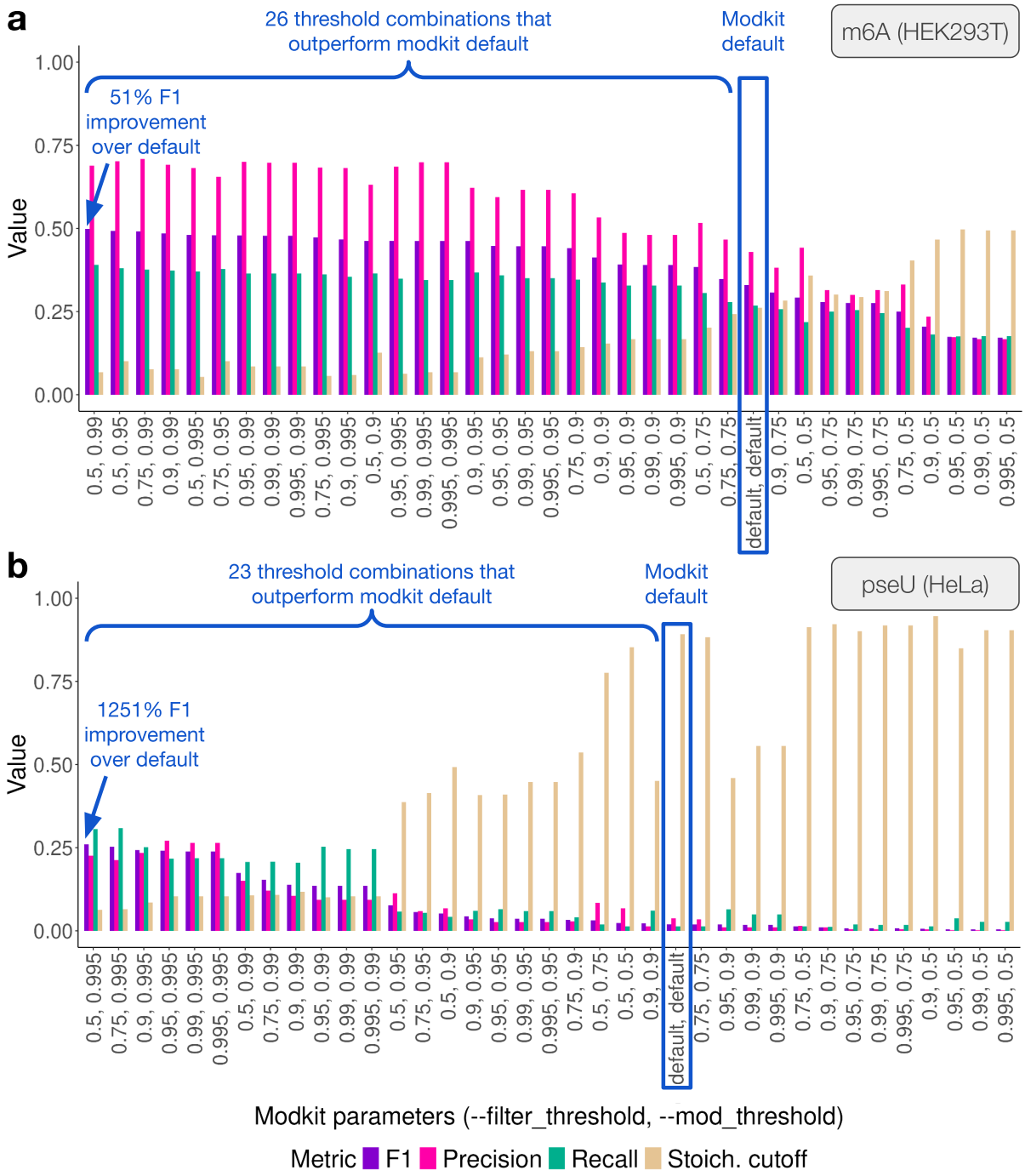

Figure 2. ModkitOpt's visualisation of modkit performance (y-axis) across 36 threshold combinations (x-axis), compared to the modkit default (annotated), in terms of F1 score (purple), precision (pink) and recall (green), for N6-methyladenosine (m6A) in HEK293T cells (**a**) and for pseudouridine (pseU) in HeLa cells (**b**). For each threshold combination, the plots show the performance obtained at the stoichiometry cutoff (tan) that maximised F1 score. For consistency, the reference sites used were detected in the same cell lines, GLORI sites in HEK293T for m6A

and BACS sites in HeLa for pseU.

We first assessed whether modkit *pileup's* default heuristic threshold estimation yields optimal stoichiometry estimates. For consistency, we ran ModkitOpt using modBAM files generated from the same cell line as the corresponding reference site sets, HEK293T for m6A[5] and HeLa for pseU[6]. For both modification types, the default thresholds were suboptimal – for m6A, 26 of the 36 parameter combinations outperformed the default (**Figure 2a**), while 23 of the 36 parameter combinations were superior for pseU (**Figure 2b**). The optimal thresholds were *--filter-threshold=0.5* and *--mod-threshold=0.99* for m6A, and *--filter-threshold=0.5* and *--mod-threshold=0.995* for pseU, resulting in F1 score improvements of 51% and 1251%, respectively, relative to the default. The significant performance improvement for pseU reflects the particularly low confidence of pseU calls under default thresholds.

The F1 score, which quantifies the precision-recall trade-off, was dominated by *--mod_threshold*, with more stringent values improving precision at the expense of recall. For m6A, moderately stringent *--mod_threshold* achieved the highest F1 scores, whereas overly strict or permissive thresholds compromised recall or precision, respectively. In contrast, for pseU, performance was maximised at the most stringent *--mod_threshold*, consistent with the lower confidence of dorado pseU calls, which necessitates stricter filtering. For both modification types, the optimal stoichiometry cutoff increased across threshold combinations as F1 score decreased, suggesting that looser *--mod_threshold* values require stricter stoichiometry cutoffs to mitigate false positives. The impact of *--filter_threshold* was secondary, modulating performance within each *--mod_threshold,* with increased stringency generally reducing the F1 score.

We next evaluated whether ModkitOpt remains effective when reference sites originate from a biological context different from the nanopore dataset. Although such references may not fully overlap with the sample-specific modification landscape, we hypothesised that a shared subset of true sites would suffice for parameter optimisation. Using nanopore datasets from HepG2 and K562 cells[4], and comparing against the original reference sets, we evaluated the same 36 threshold combinations and 1,000 stoichiometry cutoffs. For both m6A and pseU, the relative performance across the parameter space was mostly conserved (**Supplementary Figures 1-4**), and the optimal *--filter-threshold* and *--mod-threshold* values matched those identified in the previous analyses using matched cell lines. These results indicate that ModkitOpt can reliably optimise modkit parameters, even when the biological context of the reference and sample differ, provided that a subset of modification sites is shared.

Finally, we analysed the computational performance of ModkitOpt. When executed on an HPC system (PBSPro Scheduler) using default resource settings (8 CPUs and 8GB RAM for samtools filter and samtools sort; 8 CPUs and 30GB RAM for modkit pileup; 4 CPUs and 8GB RAM for samtools index; 1 CPU, 30GB RAM and 5GB job filesystem disk space for converting transcriptomic to genomic coordinates; and 1 CPU and 8GB RAM for all other tasks), ModkitOpt completed in 22 minutes for a 7GB modBAM file and 33 minutes for a 21GB modBAM file (**Supplementary Table 1**), demonstrating that comprehensive parameter optimisation can be performed efficiently on typical nanopore datasets.

Accurate estimation of RNA modification stoichiometry is essential for extracting reliable biological insight from nanopore direct RNA sequencing. ModkitOpt provides a systematic and

reproducible computational method for optimising modkit filtering thresholds through quantitative evaluation of modification calls against reference sites. By exhaustively scanning the parameter space and identifying threshold combinations that maximise precision and sensitivity, ModkitOpt overcomes the limitations of modkit's heuristic default filtering and the subjectivity of manual parameter tuning. We demonstrate that modkit's default thresholds are frequently suboptimal and that ModkitOpt identifies parameters that substantially improve the accuracy of RNA modification detection. Importantly, ModkitOpt can estimate optimal parameter combinations even when reference sites originate from a biological sample different from the nanopore dataset, enabling broad applicability. Implemented as an efficient Nextflow workflow, ModkitOpt integrates seamlessly into existing nanopore analysis pipelines and enables more accurate and reliable reporting of RNA modifications across diverse experimental settings. ModkitOpt is available at https://github.com/comprna/modkitopt.

Methods

ModkitOpt workflow

ModkitOpt (**Figure 1**) takes as input a transcriptome-aligned modBAM[8] file generated by dorado, a reference transcriptome (FASTA), gene annotation (GTF or GFF3), and a specified modification type (e.g., m6A, pseU, m5C, inosine). Input modBAM files are filtered to retain primary alignments only, then sorted and indexed using samtools (v1.21)[9]. Modkit (v0.6.0) pileup is subsequently executed across a predefined grid of filtering parameters, spanning all combinations of *--filter-threshold* and *--mod-threshold* values (0.5, 0.75, 0.9, 0.95, 0.99, 0.995), producing transcriptome-coordinate bedMethyl files with site-level stoichiometry estimates. We also allow advanced users to optionally specify the --filter-threshold and --mod-threshold values to test.

Predicted sites are converted to genomic coordinates for comparison with reference sites. When multiple transcriptomic positions map to the same genomic site, stoichiometry values are aggregated using a weighted average across transcripts. ModkitOpt evaluates 1,000 candidate stoichiometry cutoffs uniformly spaced between 0 and 0.999, retaining sites with estimated stoichiometry exceeding each cutoff.

Precision (fraction of the predicted sites that are in the reference set), recall (fraction of the reference sites that are predicted) and F1 score ($2 \cdot \frac{(Pecision \cdot Recall)}{(Pecision + Recall)}$) are calculated for every combination of modkit parameters and stoichiometry cutoff. For each parameter combination, the stoichiometry cutoff that maximises the F1 score is identified, and performance across all parameter combinations is compared to determine the combination of *--filter-threshold, --mod-threshold,* and stoichiometry cutoff that achieves the highest F1 score for the given dataset.

Reference site sets are provided for mammalian m6A and pseU, derived from GLORI-identified m6A sites in HEK293T cells[5] and BACS-identified pseU sites in HeLa cells[6], respectively. ModkitOpt is distributed as a Nextflow workflow, enabling reproducible execution on local machines or HPC systems, with resource allocation automatically managed for efficient and scalable analysis.

ModkitOpt outputs

ModkitOpt reports the optimal *--filter-threshold*, *--mod-threshold* and stoichiometry cutoff to the user and produces a bar plot depicting the precision, recall and F1 score across the parameter space (**Figure 1b**)[10]. The bar plot's data is also written in .tsv format. More detailed plots are additionally provided; a scatter plot visualising the F1 score across all combinations of *--mod-threshold* and *--filter-threshold,* using the optimal stoichiometry cutoff per threshold combination (**Supplementary Figure 5**); and the set of precision-recall curves for each threshold combination, annotated with the maximum F1 score per curve (**Supplementary Figure 6**).

References

1. Garalde, D. R. *et al.* Highly parallel direct RN A sequencing on an array of nanopores. *Nat Methods* 15, 201–206 (2018).
2. nanoporetech/dorado: Oxford Nanopore's Basecaller. https://github.com/nanoporetech/dorado.
3. nanoporetech/modkit: A bioinformatics tool for working with modified bases. https://github.com/nanoporetech/modkit.
4. Zou, Y. *et al.* A comparative evaluation of computational models for RNA modification detection using nanopore sequencing with RNA004 chemistry. *Brief Bioinform* 26, (2025).
5. Liu, C. *et al.* Absolute quantification of single-base m6A methylation in the mammalian transcriptome using GLORI. *Nat Biotechnol* 41, 355–366 (2023).
6. Xu, H. *et al.* Absolute quantitative and base-resolution sequencing reveals comprehensive landscape of pseudouridine across the human transcriptome. *Nat Methods* 21, 2024–2033 (2024).
7. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100 (2018).
8. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
9. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* 4, 1686 (2019).

Funding

This research was supported by the Australian Research Council (ARC) Discovery Project grants DP250100070 and DP250103133; by the National Health and Medical Research Council (NHMRC) Ideas Grant 2018833; by a PhD scholarship and an Innovator Grant from the ANU TALO Computational Biology Talent Accelerator, and by the ARC Centre of Excellence for the Mathematical Analysis of Cell Systems (MACSYS) (CE230100001). This research was also

indirectly supported by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) through access to computational resources provided by the National Computational Infrastructure (NCI) through the National Computational Merit Allocation Scheme (NCMAS) and the ANU Merit Allocation Scheme (ANUMAS). The funding bodies had no role in study design, data collection, or data analysis.

Author Contributions

AS, SP and EE conceptualised the study. AS led the software development and data analysis, with essential contributions from SP & EE. AS wrote the manuscript, and SP and EE reviewed the manuscript.