Sequencing accuracy and systematic errors in nanopore direct RNA sequencing

Wang Liu-Wei [1,2*], Wiep van der Toorn [1], Patrick Bohn [3], Martin Hölzer [4], Redmond Smyth [3,5], and Max von Kleist [1,6*]

* To whom correspondence should be addressed.
1 Systems Medicine of Infectious Disease (P5), Robert Koch Institute, Berlin, Germany
2 International Max-Planck Research School for Biology and Computing (IMPRS-BAC), Berlin, Germany
3 Helmholtz Institute for RNA-based Infection Research, Helmholtz Centre for Infection Research, Würzburg, Germany
4 Genome Competence Center (MF1), Robert Koch Institute, Berlin, Germany
5 Faculty of Medicine, University of Würzburg, Würzburg, Germany
6 Department of Mathematics and Computer Science, Freie Universität, Berlin, Germany

Abstract

Motivation: Direct RNA sequencing (dRNA-seq) on the Oxford Nanopore Technology platforms has become increasingly popular in recent years, with promising outlook to transform the field of epi-transcriptomics. Reads produced from dRNA-seq can cover up to full-length gene transcripts while containing decipherable information about RNA base modifications and poly-A tail lengths. Although many studies have been published exploring and expanding the potential of dRNA-seq, the sequencing accuracy and error patterns remain understudied and less characterized compared to DNA sequencing.

Results: We evaluated the sequencing accuracy and characterized the systematic error patterns of dRNA-seq on public datasets including native RNA samples from diverse organisms, as well as synthetic *in vitro* transcribed RNAs. The median read accuracy is about 90% for most organisms, although some species are more challenging. Deletions account for the majority of errors and are twice as common as mismatches or insertions. Apart from the well-known homopolymer sequencing errors, there are systematic biases across all organisms at both single nucleotide and 2-mer motif level. In general, cytosines and uracils are more likely to be erroneous than guanines and adenines. Moreover, the systematic errors are found to be strongly dependent on the local sequence contexts, with complex interactions between adjacent positions. We further evaluated the accuracy of sequencing homopolymers, read quality scores as an estimate of error rates, and the consequences of failing to detect the DNA adaptors. Lastly, we discuss the relevance of such error patterns for the downstream applications of dRNA-seq data, such as transcript identification and base modification detection.

Availability: `https://github.com/KleistLab/nanopore_dRNAseq`
Contact: wangl@rki.de, kleistm@rki.de

1 Introduction

Envisioned in the 1980s and developed over the next three decades, the nanopore sequencing technology has transformed the DNA and RNA sequencing landscape in the recent years [Deamer et al., 2016, Marx, 2023]. Sequencing platforms released by the Oxford Nanopore Technologies (ONT) has enabled economically viable high-throughput long-read sequencing of single DNA/RNA molecules. Recently, it was reported that DNA reads from the latest R10.4 chemistry achieved comparable sequencing accuracy on bacterial genomes without having to be polished with short reads, thanks to significant improvements in the accuracy of homopolymer regions [Sereika et al., 2022].

In addition to DNA sequencing, ONT also offers currently the only commercial platform for the direct sequencing of RNA molecules (dRNA-seq) [Garalde et al., 2018, Jain et al., 2022]. Ligated to a DNA adapter containing a helicase motor, a polyadenylated RNA molecule is translocated in the 3'–5' direction through a protein nanopore embedded in a electrically charged membrane. The translocation causes systematic disruptions to the ionic current flow, characteristic to the nucleotides passing through the pore at the time. The current signals are then basecalled into nucleotide sequences with machine learning algorithms, commonly known as "basecallers". Ideally, these basecallers are trained on diverse organisms for better generalization and less biases to rare sequences and species.

Traditionally, high-throughput RNA sequencing pro-

tocols have relied on reverse transcription and/or amplification, which introduce various errors and biases that confound downstream analyses [Hansen et al., 2010, Schulz et al., 2021]. In comparison, nanopore dRNA-seq avoids these biases and at the same time produces reads up to thousands of bases in length. The ability to cover full-length gene transcripts at single read resolution has the potential to improve significantly on analyses traditionally complex with short read RNA-seq, such as identifying transcript isoforms and quantifying Poly-A tail lengths [Workman et al., 2019, Soneson et al., 2019, Chen et al., 2021]. So far, nanopore dRNA-seq has been applied to diverse organisms such as DNA and RNA viruses [Depledge et al., 2019, Viehweger et al., 2019, Kim et al., 2020, Price et al., 2020], bacteria [Grünberger et al., 2022a], archaea [Grünberger et al., 2022b], plants [Parker et al., 2020, Rousseau-Gueutin et al., 2020, Gao et al., 2021b], yeast [Liu et al., 2019, Jenjaroenpun et al., 2021], fish [Begik et al., 2022a], mouse [Bilska et al., 2020] and humans [Workman et al., 2019, Hendra et al., 2022].

Another promising application of dRNA-seq is the characterisation of the "epitranscriptome". Although over 100 types of RNA modifications are known, only a small subset of them, such as m^6A, m^5C, and pseudouridine, are transcriptome-wide detectable [Mattick and Amaral, 2023], albeit with limited accuracy [Grozhik and Jaffrey, 2018]. By sequencing RNA molecules directly, dRNA-seq has the potential of producing characteristic signals not only for the four canonical RNA bases, but also the diverse family of RNA base modifications. So far, several functionally important RNA modifications are found to be detectable with dRNA-seq, with a heavy focus on m^6A [Liu et al., 2019, Pratanwanich et al., 2021, Gao et al., 2021a, Hendra et al., 2022, Price et al., 2020], but also pseudouridine [Begik et al., 2021] and inosine [Nguyen et al., 2022]. Typically, modifications are detected by computationally identifying systematic deviations in signal levels or basecalling errors at potentially modified positions. This is typically performed by including unmodified control samples (gene knockouts or *in vitro* transcribed RNA) as baseline, due to the abundance of background errors even in unmodified RNAs.

Despite the growing interest in this technology, nanopore dRNA-seq has been widely perceived as error-prone, with reported read accuracies of around 90% for single species [Rousseau-Gueutin et al., 2020, Jain et al., 2022, Grünberger et al., 2022a], which limits its wider popularization. So far, there are no studies that systematically characterise dRNA-seq accuracy and error patterns. Here, we re-analysed a diverse array of datasets covering a wide taxonomic range starting from the raw signal data and present a first comprehensive benchmark of dRNA-seq accuracy using standardized metrics. Notably, systematic errors exist at both single base and motif levels that are reproducible across the organisms, and such errors show strong and complex dependency on their local sequence contexts. In addition, we examined the relationship between read quality scores and error rates, and how adaptor detection failure can impact the read quality of short sequences. Lastly, we discuss the implications of systematic sequencing errors in downstream analyses of dRNA-seq, with a focus on transcript isoform identification and base modification detection.

2 Methods

2.1 Data

We collected public dRNA-seq datasets covering a diverse range of organisms, as well as *in vitro* transcribed RNAs. Sequencing datasets in raw *fast5* files from the following studies are included in the analysis of sequencing accuracy and errors:
1) native and *in vitro* transcribed human data [Jenjaroenpun et al., 2021],
2) the house mouse *Mus musculus* [Bilska et al., 2020],
3) yeast wildtype strain SK1 [Liu et al., 2019],
4) *Escherichia coli* [Grünberger et al., 2022a],
5) in-cell and *in vitro* transcribed SARS-CoV-2 [Kim et al., 2020],
6) *Caenorhabditis elegans* [Roach et al., 2020],
7) the zebrafish *Danio rerio* [Begik et al., 2022a],
8) *Arabidopsis thaliana* [Parker et al., 2020],
9) *Haloferax volcanii* [Grünberger et al., 2022b],
10) *in vitro* transcribed short RNAs [Begik et al., 2021], from a) *Bacillus subtilis* guanine riboswitch, b) *B. subtilis* lysine riboswitch, and c) *Tetrahymena* ribozyme (reference length 202, 274 and 460 bases, respectively).
The reference sequences for organisms 1), 2), 4), 6), 7), 8), 9) are obtained from the NCBI reference genome database and are respectively Human Build 38 GRCh38.p14, Mouse Build 39 GRCm39, *E. coli* ASM584v2, *C. elegans* WBcel235, Zebrafish Build 11 GRCz11, *A. thaliana* TAIR10.1 and *H. volcanii* DS2 ASM2568v1, while the references for 3), 5) and 10) are provided by the original authors on the respective GitHub repositories.

2.2 Basecalling, mapping and filtering

Each dataset consists of raw *fast5* files and were basecalled with the ONT basecaller Guppy (version 6.1.7) using the RNA model `rna_r9.4.1_70bps_hac`, with the option `--disable_qscore_filtering` to switch off the default read Q-score filtering threshold of 7.
The basecalled reads in *fastq* format are aligned to the respective reference genomes with minimap2 [Li, 2018] version 2.24, with the following options `-ax splice -k14 -uf -secondary=no --eqx --sam-hit-only`, which allows for split read alignment and disables secondary alignments. Aligned reads are further filtered to exclude supplementary alignments and reads with a mapping quality score ("MapQ") lower than 20 or shorter than 100 bases.
To compare Guppy with the community-published basecaller RODAN [Neumann et al., 2022], we used

the pretrained model provided at its GitHub repository (`https://github.com/biodlab/RODAN`). As RODAN failed to process some of the older versions of single read *fast5* files for some datasets, we converted them into multi read *fast5* files using the `single_to_multi_fast5` function provided in the `ont_fast5_api` library (`https://github.com/nanoporetech/ont_fast5_api`).

2.3 Evaluation metrics

After alignment and filtering, the alignment errors (mismatches, insertions and deletions) in the reads are counted by parsing through the CIGAR string in the `.sam` file format. For the accuracy of a single read, we used the following definition

$$\text{Read Accuracy} = \frac{N_{match}}{N_{match} + N_{mis} + N_{del} + N_{ins}}$$

where N_{match}, N_{mis}, N_{del} and N_{ins} refer to the number of bases on a read that are reported as match, mismatch, deletion and insertion by the aligner. Similarly, the mismatch, insertion and deletion rate per read is defined as

$$\text{Mis/Ins/Del} = \frac{N_{mis/ins/del}}{N_{match} + N_{mis} + N_{del} + N_{ins}}$$

2.4 Systematic and context-dependent errors

To evaluate systematic errors at single nucleotide and motif level, we parsed the CIGAR string of the aligned reads from minimap2 and constructed the error profiles, which are deletions and mismatches for single nucleotides, and additionally insertions that occur within 2-mer motifs. Larger datasets are subsampled to 200,000 reads each and reads aligned to soft–masked positions (represented as lower case letters) and ambiguous positions (represented as N/n) in the reference genomes are excluded.

For the analysis of homopolymers, the accuracy of a homopolymer motif is defined as the frequency of the homopolymer being correctly basecalled (as the correct nucleotide type of the correct length). The basecalled length of a homopolymer is counted only when the basecalled sequence is a homopolymer of the correct nucleotide type.

For the analysis of context-dependent errors, the "center error rate" was computed as the frequency of the incorrect basecalls of the center nucleotide within a 3-mer context where the first and last bases are correctly sequenced, in order to limit the effect of potential long-range errors propagating through the neighboring positions.

2.5 Read quality score

In addition to the per base quality scores typically found in *fastq* files, Guppy also outputs a quality score per read, "read Q-score", which is reported as `mean_qscore_template` in the `sequencing_summary.txt` file. Q-scores are typically used as a filter for poorly sequenced and basecalled reads. Guppy by default classifies reads with a Q-score lower than 7 into the `failed` folder, which is usually excluded in downstream analysis.

As the quality scores are calibrated to follow the expected error distribution of Phred quality scores [Cock et al., 2010], the expected error probability of a read is then

$$\text{P} = 10^{\frac{-\text{Q-score}}{10}} \tag{1}$$

where P is essentially 1 - Expected Read Accuracy. Consequently, Guppy's default read filter at a Q-score of 7 corresponds to an expected error rate of 20%, or equivalently, a read accuracy of 80%.

Alternatively, the read Q-score can be obtained from *fastq* files by calculating the mean per base error rate, followed by a back conversion to Phred Q-score:

$$\text{Read Q-score} = -10 \log_{10} \left[\frac{1}{N} \sum 10^{-q_i/10} \right] \tag{2}$$

where q_i is the individual base quality for base i in a read of length N.

2.6 Adapter detection analysis

In dRNA-seq, RNA molecules are ligated to a 3' DNA sequencing adapter. As a result, the start of the signals of a read contains observations of the DNA adapter. Guppy has built-in capabilities to detect the end of the DNA adapter and begin basecalling only at the start of the RNA molecule. The start of basecalling is reported as `first_sample_template` in the *fast5* files.

3 Results

3.1 Deletions outnumber mismatches and insertions

We downloaded publicly available dRNA-seq datasets in raw *fast5* format consisting of both native and *in vitro* transcribed (IVT) samples of diverse organisms. The raw signal data were then basecalled into reads with the ONT basecaller Guppy, without the default read quality filtering. The basecalled reads were aligned to the respective references and further filtered by alignment quality and a minimum length threshold of 100. The statistics regarding each dataset included in the evaluation are given in Supplementary Table 1. The datasets range from 22,315 aligned reads in the archaeon *H. volcanii* to over 2.3 million aligned reads in the *in vitro* transcribed SARS-CoV-2 sample. For most organisms, the median read accuracy is within 88%–92%, with the mouse and the zebrafish datasets being the least accurate at 87.8% and 86.7%, respectively. The median read accuracy of 90% is in general agreement with previous reports for dRNA-seq transcriptomes of single organisms [Rousseau-Gueutin et al., 2020, Jain et al., 2022, Grünberger et al., 2022a].

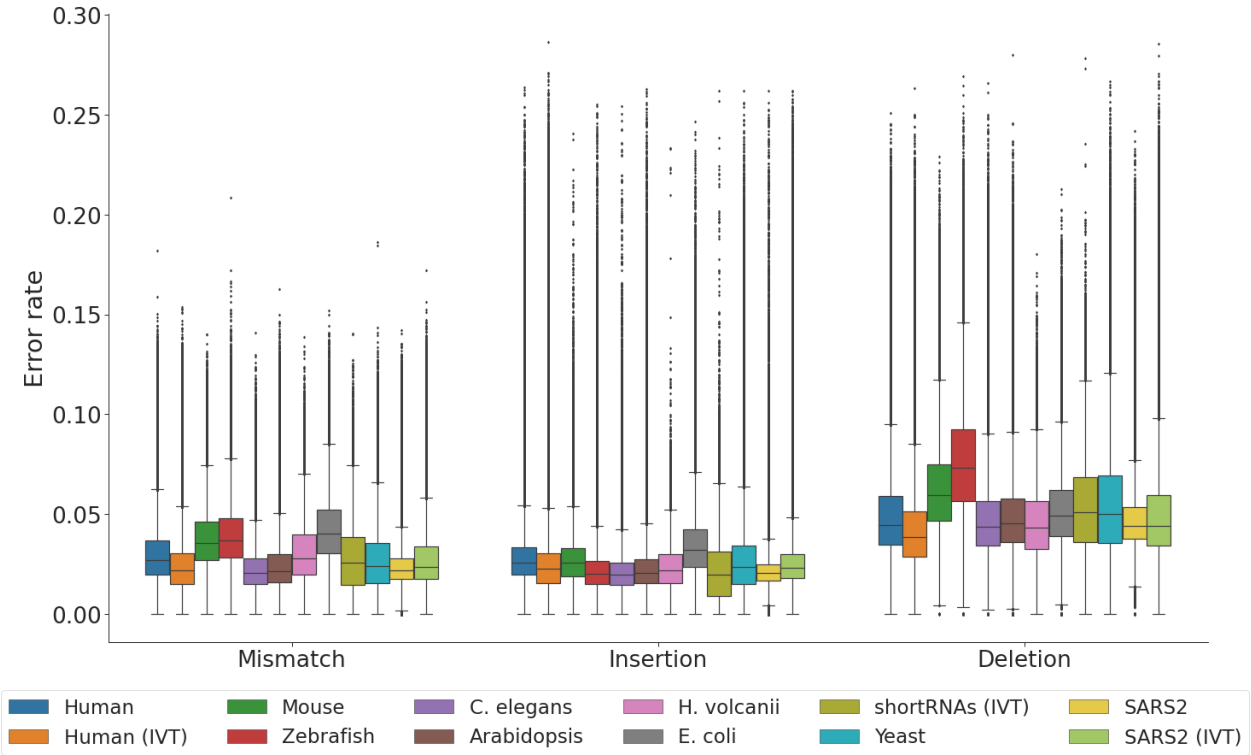

Figure 1: **Sequencing accuracy of nanopore dRNA-seq.** An overview of sequencing accuracy across diverse organisms, in terms of mismatch, insertion and deletion rate per read. Reads with accuracy lower than 70% were filtered for visualization. IVT – *in vitro* transcribed RNA.

The distributions of mismatches, insertions and deletions are similar across most organisms, where deletions represent the most frequent type of error at around 5% per read, while mismatches and insertions each account for 2–3% (Figure 1). The reads from the IVT human sample have overall 1.5% fewer errors than the native human sample, possibly due to the absence of RNA base modifications. Still, such a difference is not seen between the two SARS-CoV-2 samples.

To investigate whether the lengths of reads are related to their accuracy, we grouped reads by length, and found that the median read accuracy remains roughly constant as read lengths increase (Supplementary Figure 1). However, the distribution of read accuracy tends to have a larger variance in shorter reads.

3.2 Systematic errors in dRNA-seq

To explore how the sequencing errors are distributed at single nucleotide level, we computed the frequency of (in)correct basecalls across the four RNA bases (Figure 2a, Supplementary Figure 2). Across all datasets, guanines have consistently the highest accuracy, whereas cytosines and uracils tend to be the least accurate. Similar to the accuracy at read level, deletions remain the most frequent type of error for each base, but are much more prevalent in Cs and Us. In terms of mismatches, C bases are more likely to be basecalled as U while G bases are more often confused with A. The mismatch error patterns are consistent with the similarity of the current signals between the bases (Figure 2b), where the four bases ranked by mean signal levels are G > A > U > C (and thus C is most similar to U). Mismatch errors between neighboring nucleotides in the signal space are generally higher, especially between U and C, than between nucleotides that are further apart in signal.

Similarly, we found large differences between the errors of motifs of length 2, especially for deletions and mismatches (Figure 2c). Among the 2-mers, GA/AG are the most accurate while UC/CU have consistently the lowest accuracy. Interestingly, although most 2-mers are close to their mirrored counterparts ("XY" and "YX"), there exist clear differences between certain pairs. In particular, GC motifs have consistently higher accuracy than CGs. Whilst in general deletions were more common than mismatches and insertions, for some 2-mers mismatch errors were very close to the deletion rate, or even higher as seen in the "CG" motif. One possible explanation for the lower deletion error in some 2-mers is that these motifs are more different in signal space, and as a consequence more distinguishable by the basecaller. Insertion errors were similar for all 2-mers, suggesting that insertions likely arise from signal noise that is not motif-specific. The inserted bases generally consist of repeats of the second base in the motif, but can also occasionally involve other bases, for example, GA $\longrightarrow$ GUA (Figure 2d, Supplementary Figure 3).

Accurately sequencing and basecalling homopolymers

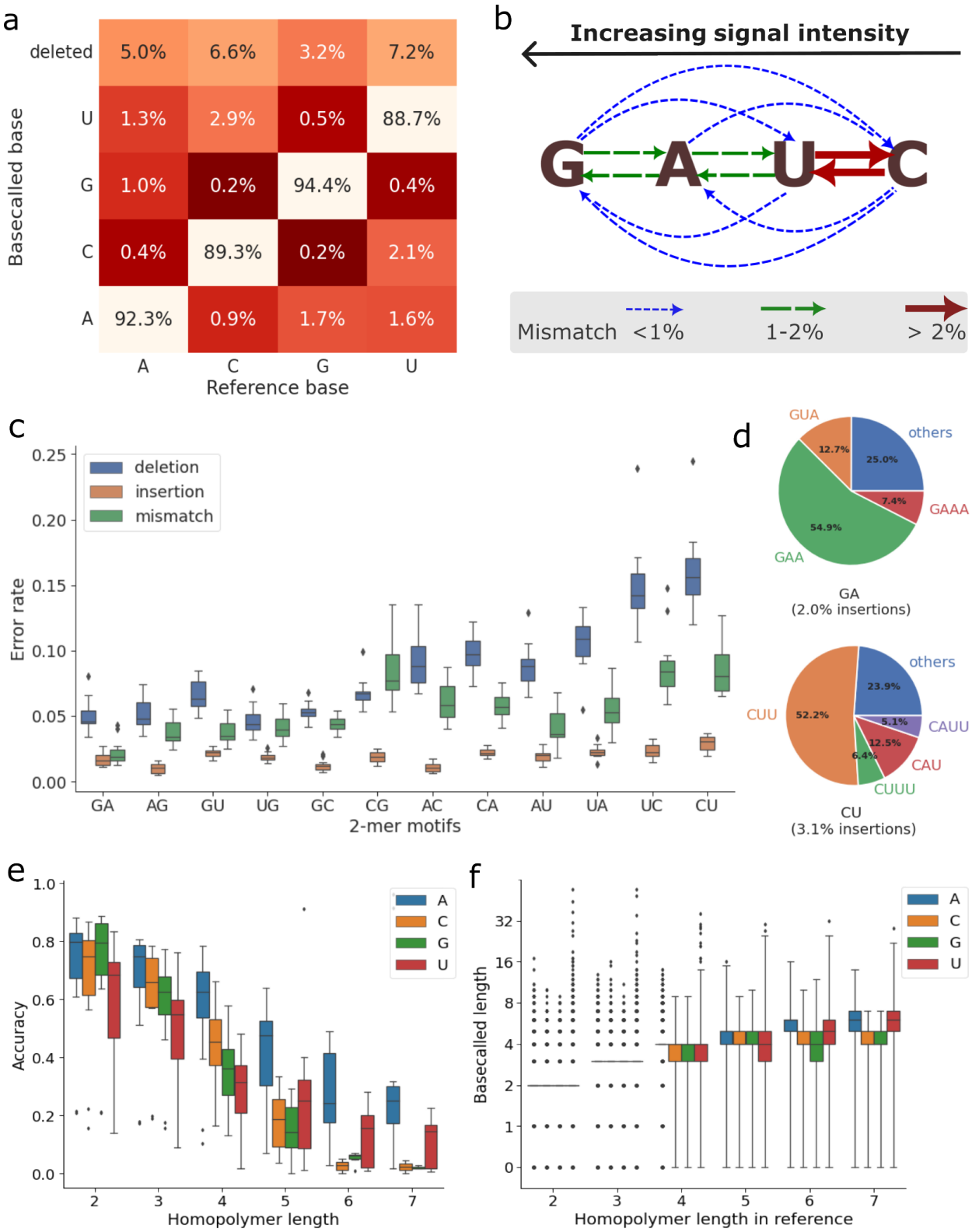

Figure 2: **Systematic errors in nanopore dRNA-seq.** a) Confusion matrix showing the frequencies of each base being correctly basecalled, miscalled or deleted, computed by taking the mean of all samples. b) The relationship between signal similarity of nucleotides and their mismatch profiles. The arrow types indicate the mismatch rate from one nucleotide to another. c) The error rate of 2-mer motifs across the samples, grouped by error type. Each data point corresponds to the accuracy of the motif in one dataset. d) The insertion error profiles of GA and CU motifs, based on the native human dataset. Inserted motifs fewer than 5% of the total number of insertions are grouped into "others". e) Basecalling accuracy of homopolymers with reference length between 2 and 7, grouped by nucleotides. f) Basecalled lengths of homopolymers with reference length between 2 and 7, when the basecalled motif is a homopolymer of the correct nucleotide type.

(motifs consisting of repeats of one nucleotide) is a well-known challenge in nanopore DNA sequencing [Delahaye and Nicolas, 2021, Sereika et al., 2022]. Unsurprisingly, we observed similar declines in the accuracy of homopolymer motifs in the dRNA-seq datasets as the motif length increases, especially at those longer than four bases (Figure 2e). In addition to the decreasing accuracy, the basecalled lengths of homopolymers are also more likely to be underestimated with increasing lengths (Figure 2f). While the lengths of most homopolymers of length 2 and 3 are basecalled correctly, homopolymers longer than 4 bases are consistently underestimated. Looking at the specific nucleotides, A and U homopolymers are more accurate than Cs and Gs. Homopolymers of A appear to be the most robust against increasing lengths, being the only nucleotide with over 50% accuracy at length of 5. Interestingly, the higher accuracies at A and T (the equivalent of U in DNA) homopolymer motifs have also been observed in nanopore DNA sequencing datasets [Delahaye and Nicolas, 2021]. This is likely due to the larger presence of A/U homopolymers in the training data for basecallers arising from well known mechanisms of RNA adenylation and uridylation [Lim et al., 2014, Morgan et al., 2017].

3.3 Context-dependent sequencing errors

At each time step of nanopore sequencing, around five consecutive RNA bases should reside in the pore simultaneously and produce a signal event, which, if distinct enough from the surrounding signals, will lead to a correct basecall. The surrounding sequence neighborhoods of a particular position, or the "sequence contexts", are known to have a effect on the accuracy of the center.

To examine the impact of sequence contexts on the accuracy in dRNA-seq data, we extended the analysis of single nucleotide errors by including the 3-mer sequence contexts – the errors of the center nucleotide, conditioned on that the neighboring 2 bases are correctly sequenced. We found that there are strong biases in the error profiles of each nucleotide towards specific sequence contexts with a strong correlation between organisms (Figure 3). Among the four RNA nucleotides, guanine is the most robust center base over its sequence contexts, with overall lower deletion than mismatch rate, indicating that the signal of G is more distinct than others. The accuracy of cytosine is the most impacted by the sequence contexts: the presence of uridines in its neighborhood can drastically change its errors, with "UCU" having both the highest deletion and mismatch rates. Except for guanines, we did not observe a general correlation between mismatch and deletion errors, suggesting that the two errors likely arise from different signal-level causes.

3.4 Estimating read errors from read and base quality scores

Quality scores are a useful estimate for read accuracy without the presence of reference genomes. There are two types of quality scores provided by ONT basecallers such as Guppy, 1) read-level quality scores (Q-scores) in the `sequence_summary.txt` file, and 2) base-level quality scores encoded in ASCII characters typically found in *fastq* files for each base on each read. Both types of quality scores are Phred quality scores, e.g. Q10 is equivalent to an error rate of 10% and Q20 equivalent to 1% (Equation 1). For dRNA-seq basecalling with Guppy, the default read Q-score filter is set to 7 (corresponding to an expected read accuracy of 80%). Reads with a lower Q-score than 7 are stored in the *failed* folder and are typically not included in downstream analysis.

To evaluate the estimation of error rate without reference, we computed the estimated read error rates based on the read Q-scores and compared them with the empirical read error rates measured after alignment. For most organisms, the measured read error rates can be reasonably estimated at read Q-scores higher than 10, while at lower Q-scores the read error is often overestimated (Figure 4a). This suggests that the default Q-score filter of 7 may be overly stringent. One important note here is that the measured read errors are a result of both basecalling and alignment. That is, aligners such as minimap2 employ "soft clipping" and do not always align reads at full length. Therefore, the empirical read error rates only reflect the errors at the aligned part of the reads. Indeed, reads with lower Q-scores tend to have more soft-clipped bases (Figure 4b), indicating a larger number of errors at the start or end of the reads.

In addition to the read Q-scores, the base-level quality scores for each basecalled read can be useful in predicting per base errors in the basecalls (mismatches and insertions). We found that correctly sequenced bases tend to have higher quality scores than mismatched and inserted positions (Figure 4c). Moreoever, except for uracil, the per base error rate can be well estimated when base-level quality scores are higher than 15 (Figure 4d).

Theoretically, the read Q-scores are obtained by the log transformation of the mean per base error estimate, based on the relationship described in Equation 2. However, we observed a systematic deviation between the empirically computed Q-scores and the original read Q-scores reported by Guppy, especially at higher values (Figure 4e). This deviation is found across all datasets and could potentially be a consequence of the loss of numerical precision in the discretisation of per base quality scores into ASCII characters. Nevertheless, the deviation remains highly linear (fitted with a ordinary least square model, p-value ≈ 0).

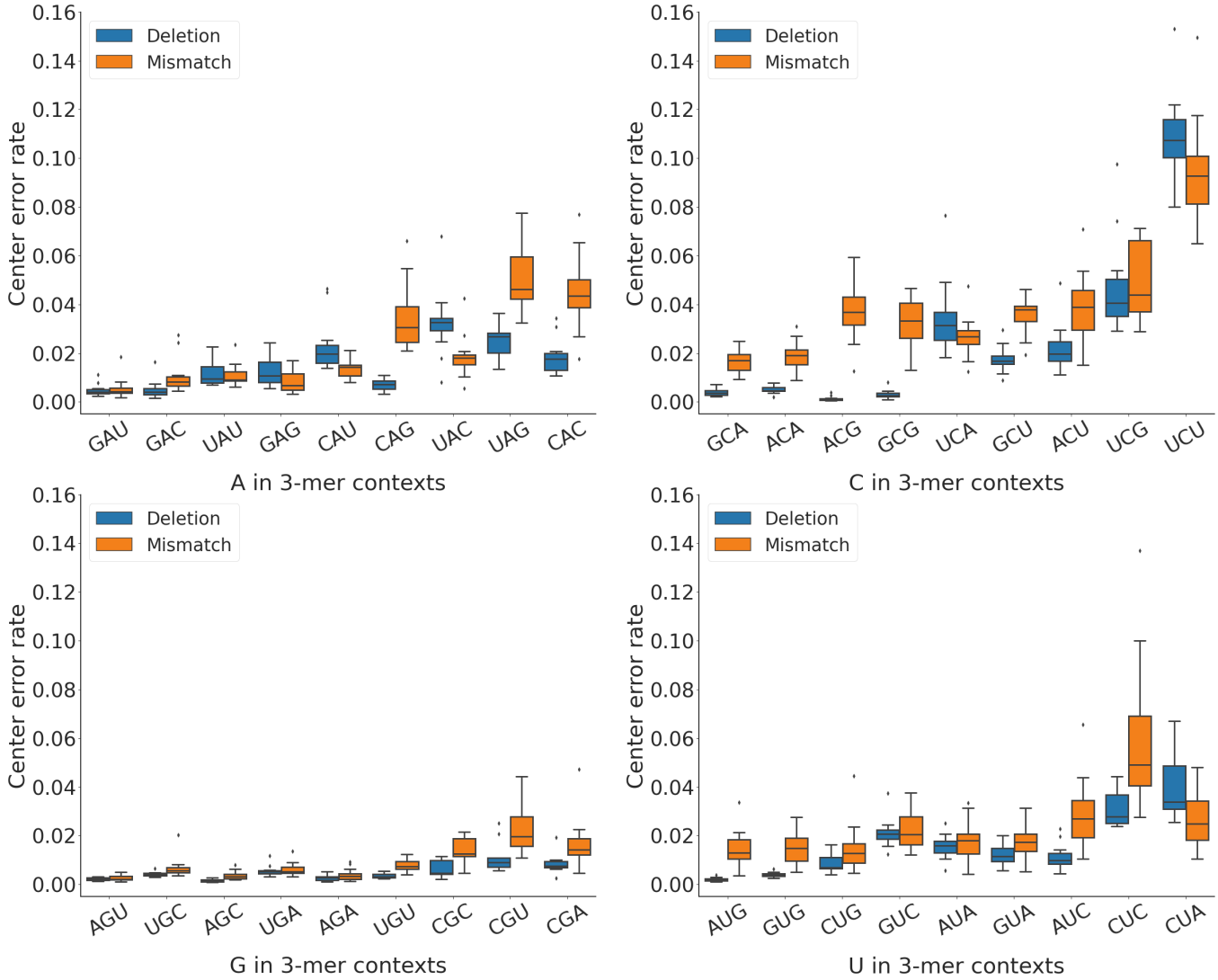

Figure 3: **Context-dependent errors.** The error profiles of the center nucleotides across 3-mer contexts, conditional on that the neighboring two bases are correctly basecalled. Each data point represents the mean error rate of all such 3-mer motifs in one dataset.

3.5 Adapter detection failure

Whilst read lengths are not found to impact read accuracy (Supplementary Figure 1), we observed that for shorter reads, the discrepancy between the basecalled read length and the aligned read length can be larger (Figure 5a), as a consequence of more soft-clipped bases at the start and end of the reads.

We then counted the soft-clipped bases per read and found that in contrast to the 5' end, the length distribution of the soft-clipped bases at the 3' end showed a clear bimodal pattern, with a second mode locating roughly between 50 and 130 bases and accouting for 30% of the reads (Figure 5b, native human). The unique bimodality of 3' soft-clips are found across datasets, with the second mode accounting for up to 41.5% of the reads in the zebrafish sample (Supplementary Figure 4), indicating that this is a common pattern in dRNA-seq data.

Upon inspecting the signal data of these reads, we discovered that the soft–clipped bases at the 3' end were actually errorneous basecalls originating from the DNA adapter. For these reads, Guppy began basecalling at the beginning of the adapter, instead of at the start of the RNA segment. The beginning of reads with this issue typically contains a jump in signal intensity that is comparable to the one typically seen at the end of the adapter, marking the DNA/RNA boundary (Figure 5c).

To evaluate the impact of adapter detection failure on the quality of downstream RNA basecalls, we compared the alignment error rates and basecall quality scores of reads with and without successful adapter detection. We found that basecalls in the DNA adapter are of low quality (Supplementary Figure 5a), but we did not observe that having these spurious basecalls at the start of the read negatively impacts the error rate of the subsequent RNA basecalls (Supplementary Figure 5b). However, for shorter reads (less than 400 bases), we observed that the spurious DNA basecalls signifi-

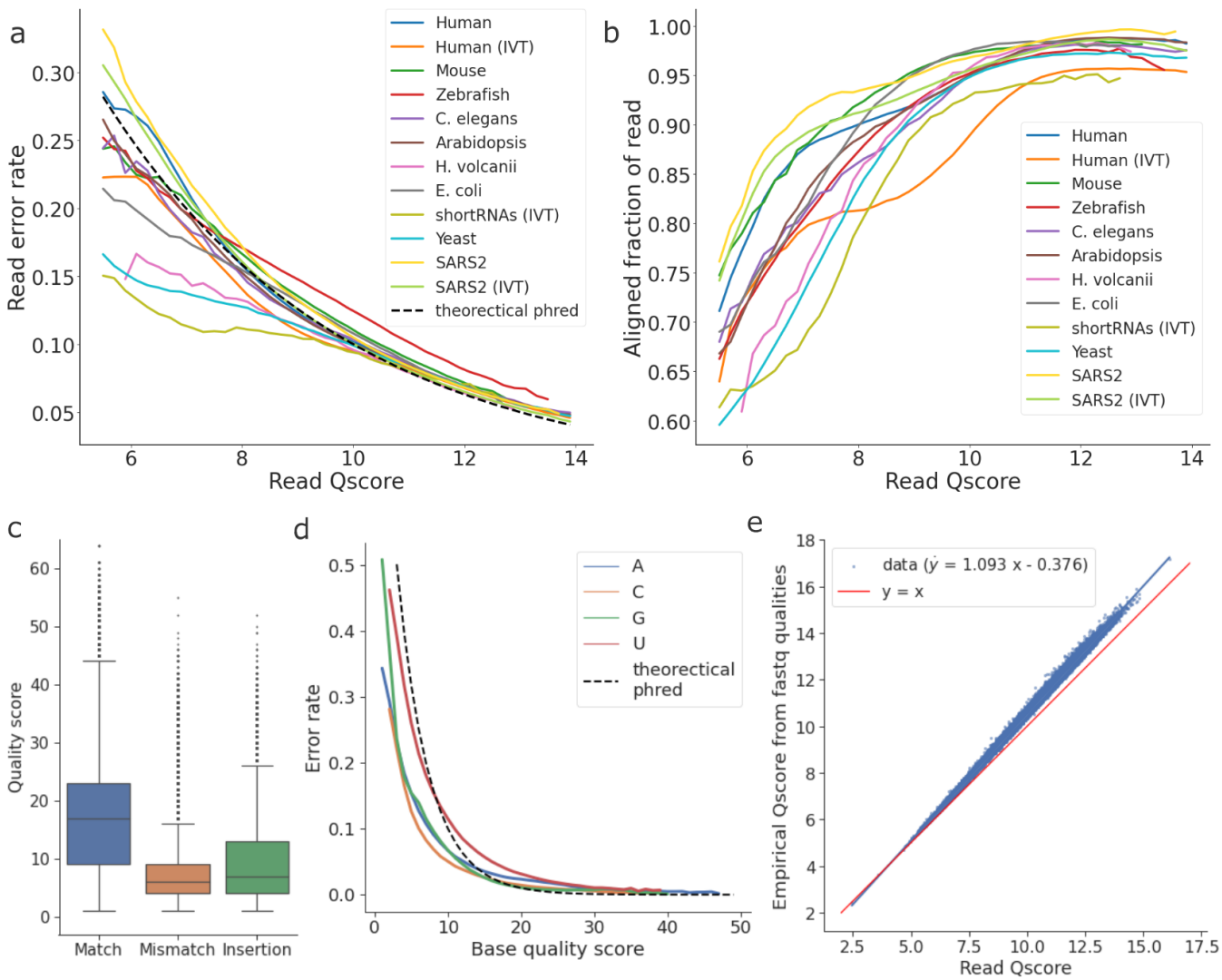

Figure 4: **Read and base quality scores in dRNA-seq.** (a) The relationship between read Q-scores and read error rates. The dashed line reflects the theoretical error rate based on the Phred quality scores. (b) The relationship between read Q-scores and the aligned fraction of reads after alignment for each dataset. (c) The distribution of per base quality scores, grouped by the per base error type. (d) The relationship between per base quality scores and per base error rate (mismatch rate + insertion rate), grouped by nucleotide. (e) The relationship between read Q-scores and "empirical" Q-scores computed from per base quality scores of reads in *fastq* files, based on the native human sample. The linear model is fitted by the ordinary least square method.

cantly impact the read Q-score, dragging it below the default Q-score cutoff of 7 and leading to unnecessary data loss (Figure 5d, Supplementary Figure 5c).

3.6 Comparison with the community basecaller RODAN

For dRNA-seq data, RODAN [Neumann et al., 2022] is currently the only non-ONT basecaller published by the research community. To examine whether RODAN has significantly improved over Guppy in basecalling performance, we re-basecalled all the datasets in this study with RODAN and compared its performance with Guppy (Supplementary Table 2). Among the datasets, four organisms (native human, *C. elegans*, *E. coli*, and *Arabidopsis*) were included in the training data of RODAN. RODAN performed notably higher than Guppy in three of these four datasets (except *Arabidopsis*). In the species outside its training set, RODAN performs better than Guppy on the mouse, *H. volcanii*, the short RNAs and the IVT SARS2 datasets with 1–2% improvement in median read accuracy, and worse than Guppy on the zebrafish dataset (a decrease of 4.8% in median read accuracy).

While the read accuracies are comparable on most datasets, the reads from RODAN have shorter median aligned read lengths in 10 out of 12 datasets (Figure 6a, Supplementary Table 2). The largest drop in the aligned read lengths is observed in the native SARS2 dataset, where the median read length from RODAN is 708 nucleotides shorter, i.e. a 34% decrease. To explore whether RODAN's performance is related to

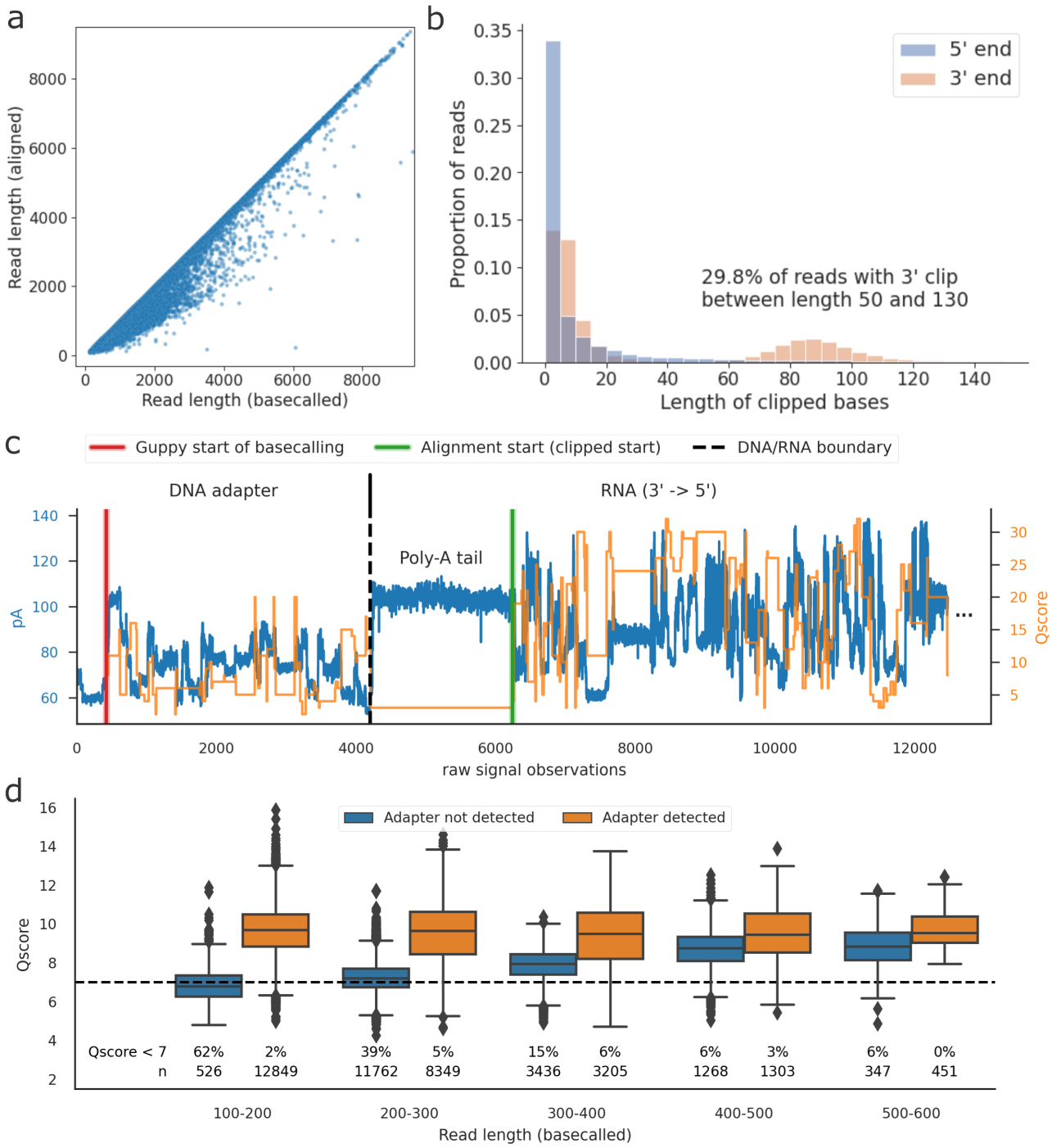

Figure 5: **Adapter detection failure in Guppy.** a) The relationship between the basecalled read length and the aligned read length after soft-clipping, based on the native human sample. b) The length distribution of the soft-clipped bases at the 5' and 3' ends of reads, the latter of which shows a clear bimodal distribution, indicative of adapter detection failure, based on the native human sample. c) A signal visualization of a read whose adapter failed to be detected by Guppy. The left y-axis shows the signal intensity value (in pA) and the right y-axis shows the quality scores per base. d) The distribution of read Q-scores depending on whether the adapter was detected successfully, grouped by read length, based on the short RNA dataset. Adapter detection failure can reduce Q-score for short reads, leading to them being filtered at the conventional Q-score threshold of 7 (dashed line).

read length, we grouped reads by length and found that RODAN performs overall better than Guppy in shorter reads, but the improvement diminishes as read length increases, with the exception of organisms in RODAN's training dataset (Figure 6b).

Lastly, the systematic sequencing errors at single nucleotide and motif levels are also prevalent in the reads basecalled by RODAN (Supplementary Figure 7), sug-

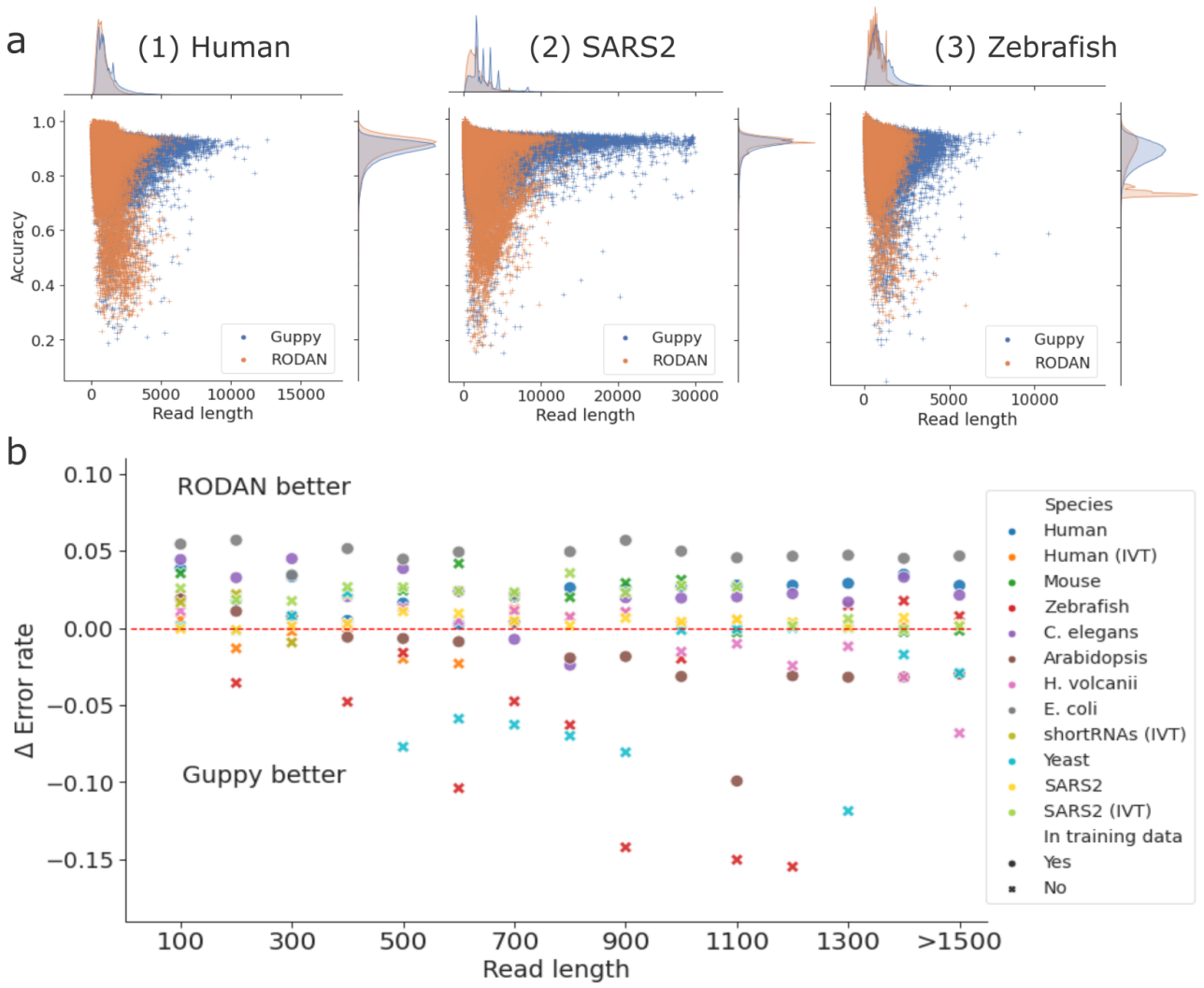

Figure 6: **Comparison of basecalling performance between Guppy and the community basecaller RODAN**. (a) The joint and marginal distributions of read length and read accuracy in 1) the native human, 2) SARS2 and 3) zebrafish samples. Each dataset were subsampled to 200,000 reads due to the runtime for kernel density estimation. (b) The relationship between aligned read length and the performance difference of RODAN versus Guppy. The Δ Error rate is $\text{Error}_{\text{Guppy}}$ - $\text{Error}_{\text{RODAN}}$, thus a positive value means RODAN is more accurate over Guppy. The symbols represent whether the dataset was present (•) or absent (×) in the training set of RODAN.

gesting that the fundamental causes of these errors reside in the raw signal data.

4 Discussion

Since the release of the first MinION device by ONT, the accuracy of nanopore DNA sequencing has seen significant improvements, thanks to developments in both the pore chemistry and the basecalling algorithm [Rang et al., 2018, Sereika et al., 2022]. At the same time, dRNA-seq has gained considerable popularity due to its reduced bias and the ability to sequence RNA base modifications and characterise the epitranscriptome. However, improvements in the sequencing accuracy of dRNA-seq has so far been limited [Jain et al., 2022]. Here, we systematically evaluated the accuracy and systematic error patterns of dRNA-seq in datasets covering diverse organisms across the tree of life. We found that the read accuracy is around 90% across organisms in both native and IVT samples, and that deletions account for the majority of errors.

The high error rate of dRNA-seq has several important implications in downstream analyses. Firstly, as each native RNA read represents a single polyadenylated RNA transcript that are potentially protein coding, one promising application of dRNA-seq is to identify and quantify the expression of novel transcript isoforms and open reading frames (ORFs). However, the high presence of insertion and deletion (indel) errors can often lead to frameshift errors in gene prediction, a well-known issue also in long-read DNA-seq [Watson and Warr, 2019]. To address this, Depledge et al. [2019] attempted to reduce errors by correcting dRNA-

seq reads with Illumina short reads, yet the remaining indel errors still precluded ORF prediction in more than 80% of the reads. Moreover, despite being much longer than Illumina reads, alignments of dRNA-seq reads to transcriptomes often produce a considerable amount of secondary alignments with similar mapping scores to primary alignments, thus making the identification of true transcript origin difficult [Soneson et al., 2019, Price et al., 2020], which we suspect to also be a result of the high error rate.

In addition to the overall error rates, we also observed systematic errors across our datasets that show sequence-specific patterns and biases. The systematic errors at single nucleotide and motif levels are particularly relevant to the application of dRNA-seq in characterising the epitranscriptome by detecting native RNA base modifications. Current computational methods for detecting RNA modifications attempt to identify potentially modified sites through either higher basecalling error rates or deviations in the signal at a particular position, by comparing with control samples that have fewer modifications (in the case of writer gene knockouts) or no modifications (in the case of IVT samples) [Begik et al., 2022b]. Because there are inherent background errors and signal changes even in unmodified RNA, the inclusion of control samples is necessary to reduce the false positive rate of modification detection [Kim et al., 2020], thus limiting our ability to detect rare modifications *de novo* when negative controls are hard to obtain.

It is worth noting that we did not investigate the role of native base modifications on the observed error patterns. While the modification-free IVT human dataset has less errors than its native counterpart, the same pattern is not seen in the two SARS2 datasets. Certain types of modifications, such as m^6A, m^5C and pseudouridine, have been found to lead to an overall increase in basecalling errors, but the error patterns are also dependent on the local sequence contexts [Parker et al., 2020, Leger et al., 2021]. Moreover, one would expect that commonly modified motifs, such as the DRACH motif of m^6A, are already present in the training data and can be learned by the basecaller, but less so for the rare or species-specific types of modifications. Indeed, a careful review into the differences between the native and the unmodified IVT samples is warranted to uncover the likely complex relationship between basecalling errors and specific types of modifications.

In general, there are two sources of sequencing errors in nanopore sequencing data: firstly, the raw signals produced by the nanopore sequencer and secondly, the basecalling algorithm translating the signal data into nucleotide sequences [Rang et al., 2018]. In addition to the constant pore updates for nanopore DNA-seq, there is also a constant influx of new basecallers released by both ONT and the broader research community [Wick et al., 2019]. To facilitate such efforts, ONT provided softwares such as the *Taiyaki* tool (`https://github.com/nanoporetech/taiyaki`) to support the development of research basecallers. However, for dRNA-seq data, RODAN [Neumann et al., 2022] remains currently the only published community basecaller outside of ONT. From our evaluation, the performance of RODAN holds up well against Guppy in terms of read accuracy, especially for organisms in its training data. As species-specific training data are known to improve performance of nanopore basecallers in those species [Wick et al., 2019, Ferguson et al., 2022], the improvements of RODAN suggest a promising direction for training species-specific basecallers also for dRNA-seq data. Lastly, the presence of the same systematic error patterns in RODAN points to more fundamental causes of errors in the raw signal data, necessitating further development of better pore chemistry to produce higher quality dRNA-seq data.

In conclusion, while dRNA-seq offers exciting opportunities for studying RNA transcription and modifications, our evaluation highlights the need for continued improvements in data quality and accuracy. Clearly, further development and optimization of dRNA-seq protocols, pore chemistry and basecalling algorithms are desirable. At the same time, appropriate quality control and error correction methods are needed to mitigate the effects of high error rates and systematic biases in downstream analyses, especially in the area of *de novo* transcript identification and modification detection.

Funding

W.L.W was funded by European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF).

References

O. Begik, M. C. Lucas, L. P. Pryszcz, J. M. Ramirez, R. Medina, I. Milenkovic, S. Cruciani, H. Liu, H. G. S. Vieira, A. Sas-Chen, et al. Quantitative profiling of pseudouridylation dynamics in native rnas with nanopore sequencing. *Nature biotechnology*, 39 (10):1278–1291, 2021.

O. Begik, G. Diensthuber, H. Liu, A. Delgado-Tejedor, C. Kontur, A. M. Niazi, E. Valen, A. J. Giraldez, J.-D. Beaudoin, J. S. Mattick, et al. Nano3p-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cdna sequencing. *Nature Methods*, pages 1–11, 2022a.

O. Begik, J. S. Mattick, and E. M. Novoa. Exploring the epitranscriptome by native rna sequencing. *RNA*, 28(11):1430–1439, 2022b.

A. Bilska, M. Kusio-Kobiałka, P. S. Krawczyk, O. Gewartowska, B. Tarkowski, K. Kobyłecki, D. Nowis, J. Golab, J. Gruchota, E. Borsuk, et al.

Immunoglobulin expression and the humoral immune response is regulated by the non-canonical poly (a) polymerase tent5c. *Nature Communications*, 11 (1):2032, 2020.

Y. Chen, N. M. Davidson, Y. K. Wan, H. Patel, F. Yao, H. M. Low, C. Hendra, L. Watten, A. Sim, C. Sawyer, V. Iakovleva, P. L. Lee, L. Xin, H. E. V. Ng, J. M. Loo, X. Ong, H. Q. A. Ng, J. Wang, W. Q. C. Koh, S. Y. P. Poon, D. Stanojevic, H.-D. Tran, K. H. E. Lim, S. Y. Toh, P. A. Ewels, H.-H. Ng, N. Iyer, A. Thiery, W. J. Chng, L. Chen, R. DasGupta, M. Sikic, Y.-S. Chan, B. O. P. Tan, Y. Wan, W. L. Tam, Q. Yu, C. C. Khor, T. Wüstefeld, P. N. Pratanwanich, M. I. Love, W. S. S. Goh, S. B. Ng, A. Oshlack, J. Göke, and . A systematic benchmark of nanopore long read rna sequencing for transcript level analysis in human cell lines. *bioRxiv*, 2021. doi: 10.1101/2021.04.21.440736. URL `https://www.biorxiv.org/content/early/2021/04/22/2021.04.21.440736`.

P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

D. Deamer, M. Akeson, and D. Branton. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5): 518–524, 2016.

C. Delahaye and J. Nicolas. Sequencing dna with nanopores: Troubles and biases. *PloS one*, 16(10): e0257521, 2021.

D. P. Depledge, K. P. Srinivas, T. Sadaoka, D. Bready, Y. Mori, D. G. Placantonakis, I. Mohr, and A. C. Wilson. Direct rna sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature communications*, 10(1):754, 2019.

S. Ferguson, T. McLay, R. L. Andrew, J. J. Bruhl, B. Schwessinger, J. Borevitz, and A. Jones. Species-specific basecallers improve actual accuracy of nanopore sequencing in plants. *Plant Methods*, 18 (1):1–11, 2022.

Y. Gao, X. Liu, B. Wu, H. Wang, F. Xi, M. V. Kohnen, A. S. Reddy, and L. Gu. Quantitative profiling of n 6-methyladenosine at single-base resolution in stem-differentiating xylem of populus trichocarpa using nanopore direct rna sequencing. *Genome biology*, 22:1–17, 2021a.

Y. Gao, X. Liu, B. Wu, H. Wang, F. Xi, M. V. Kohnen, A. S. Reddy, and L. Gu. Quantitative profiling of n 6-methyladenosine at single-base resolution in stem-differentiating xylem of populus trichocarpa using nanopore direct rna sequencing. *Genome biology*, 22:1–17, 2021b.

D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201–206, 2018.

A. V. Grozhik and S. R. Jaffrey. Distinguishing rna modifications from noise in epitranscriptome maps. *Nature Chemical Biology*, 14(3):215–225, 2018.

F. Grünberger, S. Ferreira-Cerca, and D. Grohmann. Nanopore sequencing of rna and cdna molecules in escherichia coli. *RNA*, 28(3):400–417, 2022a.

F. Grünberger, M. Jüttner, R. Knüppel, S. Ferreira-Cerca, and D. Grohmann. Insights into rrna processing and modification mapping in archaea using nanopore-based rna sequencing. *bioRxiv*, 2022b. doi: 10.1101/2021.06.14.448286. URL `https://www.biorxiv.org/content/early/2022/11/21/2021.06.14.448286`.

K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38 (12):e131–e131, 2010.

C. Hendra, P. N. Pratanwanich, Y. K. Wan, W. S. Goh, A. Thiery, and J. Göke. Detection of m6a from direct rna sequencing using a multiple instance learning framework. *Nature Methods*, pages 1–9, 2022.

M. Jain, R. Abu-Shumays, H. E. Olsen, and M. Akeson. Advances in nanopore direct rna sequencing. *Nature Methods*, 19(10):1160–1164, 2022.

P. Jenjaroenpun, T. Wongsurawat, T. D. Wadley, T. M. Wassenaar, J. Liu, Q. Dai, V. Wanchai, N. S. Akel, A. Jamshidi-Parsian, A. T. Franco, et al. Decoding the epitranscriptional landscape from native rna sequences. *Nucleic acids research*, 49(2):e7–e7, 2021.

D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, and H. Chang. The architecture of sars-cov-2 transcriptome. *Cell*, 181(4):914–921, 2020.

A. Leger, P. P. Amaral, L. Pandolfini, C. Capitanchik, F. Capraro, V. Miano, V. Migliori, P. Toolan-Kerr, T. Sideri, A. J. Enright, et al. Rna modifications detection by comparative nanopore direct rna sequencing. *Nature communications*, 12(1):7198, 2021.

H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

J. Lim, M. Ha, H. Chang, S. C. Kwon, D. K. Simanshu, D. J. Patel, and V. N. Kim. Uridylation by tut4 and tut7 marks mrna for degradation. *Cell*, 159(6):1365–1376, 2014.

H. Liu, O. Begik, M. C. Lucas, J. M. Ramirez, C. E. Mason, D. Wiener, S. Schwartz, J. S. Mattick, M. A. Smith, and E. M. Novoa. Accurate detection of m6a rna modifications in native rna sequences. *Nature communications*, 10(1):1–9, 2019.

V. Marx. Method of the year: long-read sequencing. *Nature Methods*, 20(1):6–11, 2023.

J. Mattick and P. Amaral. *RNA, the epicenter of genetic information*. Taylor & Francis, 2023.

M. Morgan, C. Much, M. DiGiacomo, C. Azzi, I. Ivanova, D. M. Vitsios, J. Pistolic, P. Collier, P. N. Moreira, V. Benes, et al. mrna 3 uridylation and poly (a) tail length sculpt the mammalian maternal transcriptome. *Nature*, 548(7667):347–351, 2017.

D. Neumann, A. S. Reddy, and A. Ben-Hur. Rodan: a fully convolutional architecture for basecalling nanopore rna sequencing data. *BMC bioinformatics*, 23(1):1–9, 2022.

T. A. Nguyen, J. W. J. Heng, P. Kaewsapsak, E. P. L. Kok, D. Stanojević, H. Liu, A. Cardilla, A. Praditya, Z. Yi, M. Lin, et al. Direct identification of a-to-i editing sites with nanopore native rna sequencing. *Nature Methods*, 19(7):833–844, 2022.

M. T. Parker, K. Knop, A. V. Sherwood, N. J. Schurch, K. Mackinnon, P. D. Gould, A. J. Hall, G. J. Barton, and G. G. Simpson. Nanopore direct rna sequencing maps the complexity of arabidopsis mrna processing and m6a modification. *Elife*, 9:e49658, 2020.

P. N. Pratanwanich, F. Yao, Y. Chen, C. W. Koh, Y. K. Wan, C. Hendra, P. Poon, Y. T. Goh, P. M. Yap, J. Y. Chooi, et al. Identification of differential rna modifications from nanopore direct rna sequencing with xpore. *Nature biotechnology*, 39(11):1394–1402, 2021.

A. M. Price, K. E. Hayer, A. B. McIntyre, N. S. Gokhale, J. S. Abebe, A. N. Della Fera, C. E. Mason, S. M. Horner, A. C. Wilson, D. P. Depledge, et al. Direct rna sequencing reveals m6a modifications on adenovirus rna are necessary for efficient splicing. *Nature communications*, 11(1):6016, 2020.

F. J. Rang, W. P. Kloosterman, and J. de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, 19(1):90, 2018.

N. P. Roach, N. Sadowski, A. F. Alessi, W. Timp, J. Taylor, and J. K. Kim. The full-length transcriptome of c. elegans using direct rna sequencing. *Genome research*, 30(2):299–312, 2020.

M. Rousseau-Gueutin, C. Belser, C. Da Silva, G. Richard, B. Istace, C. Cruaud, C. Falentin, F. Boideau, J. Boutte, R. Delourme, et al. Long-read assembly of the brassica napus reference genome darmor-bzh. *GigaScience*, 9(12):giaa137, 2020.

L. Schulz, M. Torres-Diz, M. Cortés-López, K. E. Hayer, M. Asnani, S. K. Tasian, Y. Barash, E. Sotillo, K. Zarnack, J. König, et al. Direct long-read rna sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biology*, 22(1):1–12, 2021.

M. Sereika, R. H. Kirkegaard, S. M. Karst, T. Y. Michaelsen, E. A. Sørensen, R. D. Wollenberg, and M. Albertsen. Oxford nanopore r10. 4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature methods*, 19(7):823–826, 2022.

C. Soneson, Y. Yao, A. Bratus-Neuenschwander, A. Patrignani, M. D. Robinson, and S. Hussain. A comprehensive examination of nanopore native rna sequencing for characterization of complex transcriptomes. *Nature communications*, 10(1):3359, 2019.

A. Viehweger, S. Krautwurst, K. Lamkiewicz, R. Madhugiri, J. Ziebuhr, M. Hölzer, and M. Marz. Direct rna nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome research*, 29(9):1545–1554, 2019.

M. Watson and A. Warr. Errors in long-read assemblies can critically affect protein prediction. *Nature biotechnology*, 37(2):124–126, 2019.

R. R. Wick, L. M. Judd, and K. E. Holt. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome biology*, 20:1–10, 2019.

R. E. Workman, A. D. Tang, P. S. Tang, M. Jain, J. R. Tyson, R. Razaghi, P. C. Zuzarte, T. Gilpatrick, A. Payne, J. Quick, et al. Nanopore native rna sequencing of a human poly (a) transcriptome. *Nature methods*, 16(12):1297–1305, 2019.

5 Supplementary materials

Dataset	Reads aligned	med. Q-score	med. Length	Acc	Mis	Ins	Del
Human	1483755	10.2	876	0.901	0.027	0.026	0.045
Human (IVT)	2015950	10.7	486	0.916	0.022	0.022	0.039
Mouse	181977	9.6	616	0.878	0.036	0.025	0.059
Zebrafish	553604	9.6	839	0.867	0.037	0.020	0.073
C. elegans	227372	11.0	688	0.915	0.020	0.019	0.044
Arabidopsis	1010943	10.7	878	0.911	0.022	0.021	0.045
H. volcanii	22315	9.9	505	0.905	0.028	0.022	0.043
E. coli	198448	9.3	662	0.876	0.040	0.032	0.049
shortRNAs (IVT)	45028	8.5	160	0.899	0.026	0.019	0.051
Yeast	324889	9.4	314	0.898	0.024	0.024	0.050
SARS2	575673	11.0	2078	0.914	0.022	0.020	0.044
SARS2 (IVT)	2316635	10.5	1572	0.908	0.024	0.023	0.044

Table 1: **Details of the datasets.** Statistics regarding the aligned reads (counts, quality scores and lengths) and their accuracies of each dataset included in this study.

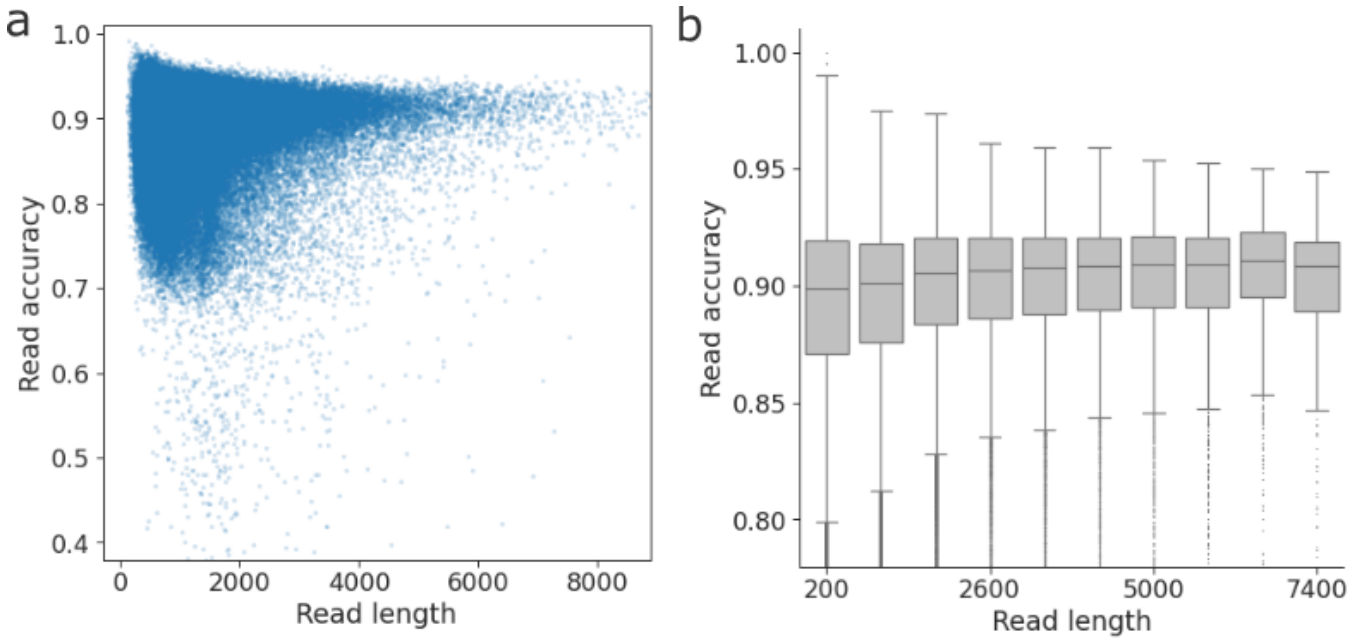

Figure 1: **The relationship between read length and read accuracy**. a) Scatter plot of read length versus read accuracy, based on the native human dataset. b) A boxplot of the same data, grouped by read length.

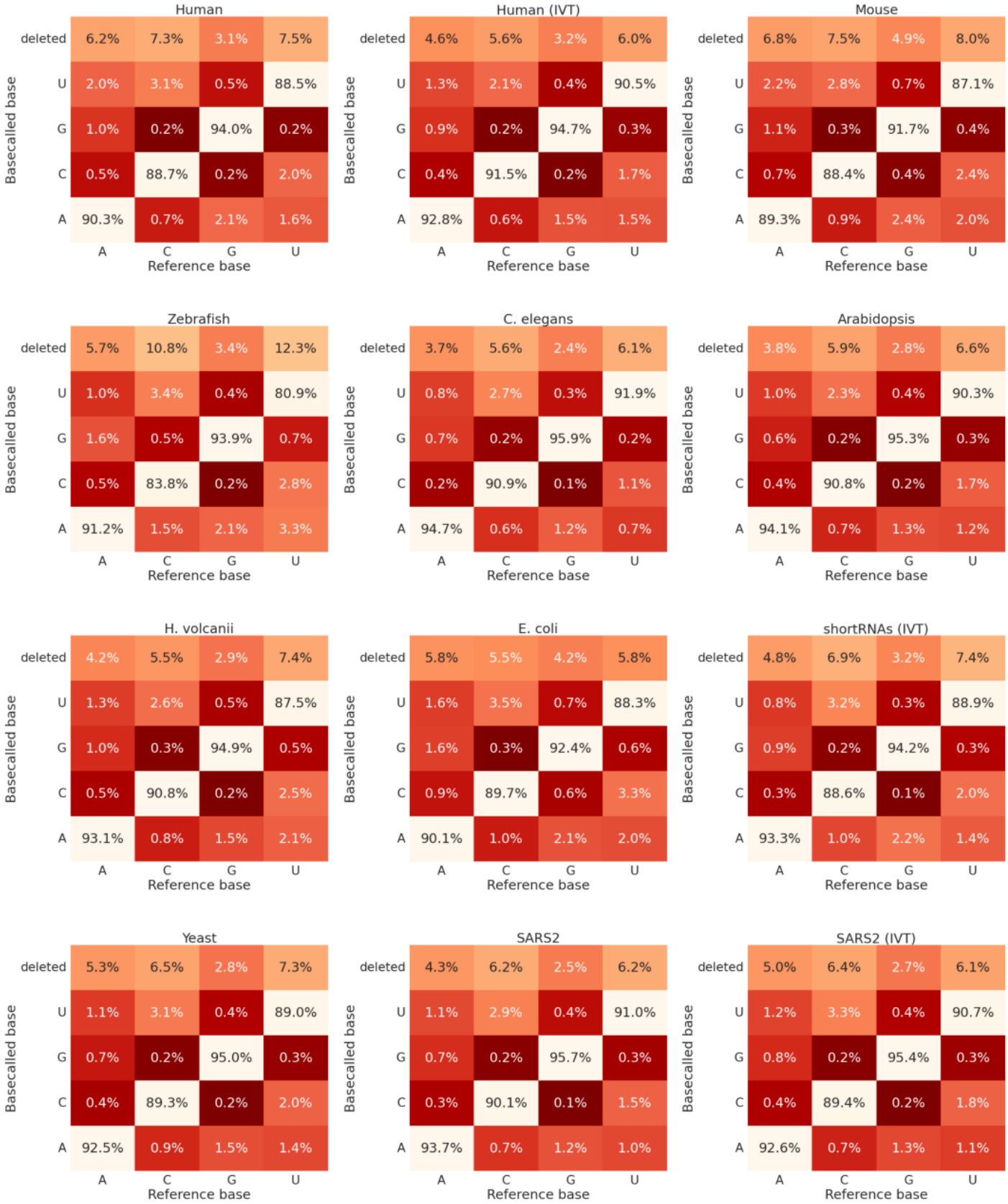

Figure 2: **Single nucleotide error profiles across organisms.** Confusion matrix showing the proportion of each base in reference being correctly basecalled, miscalled or deleted, for each organism included in the study.

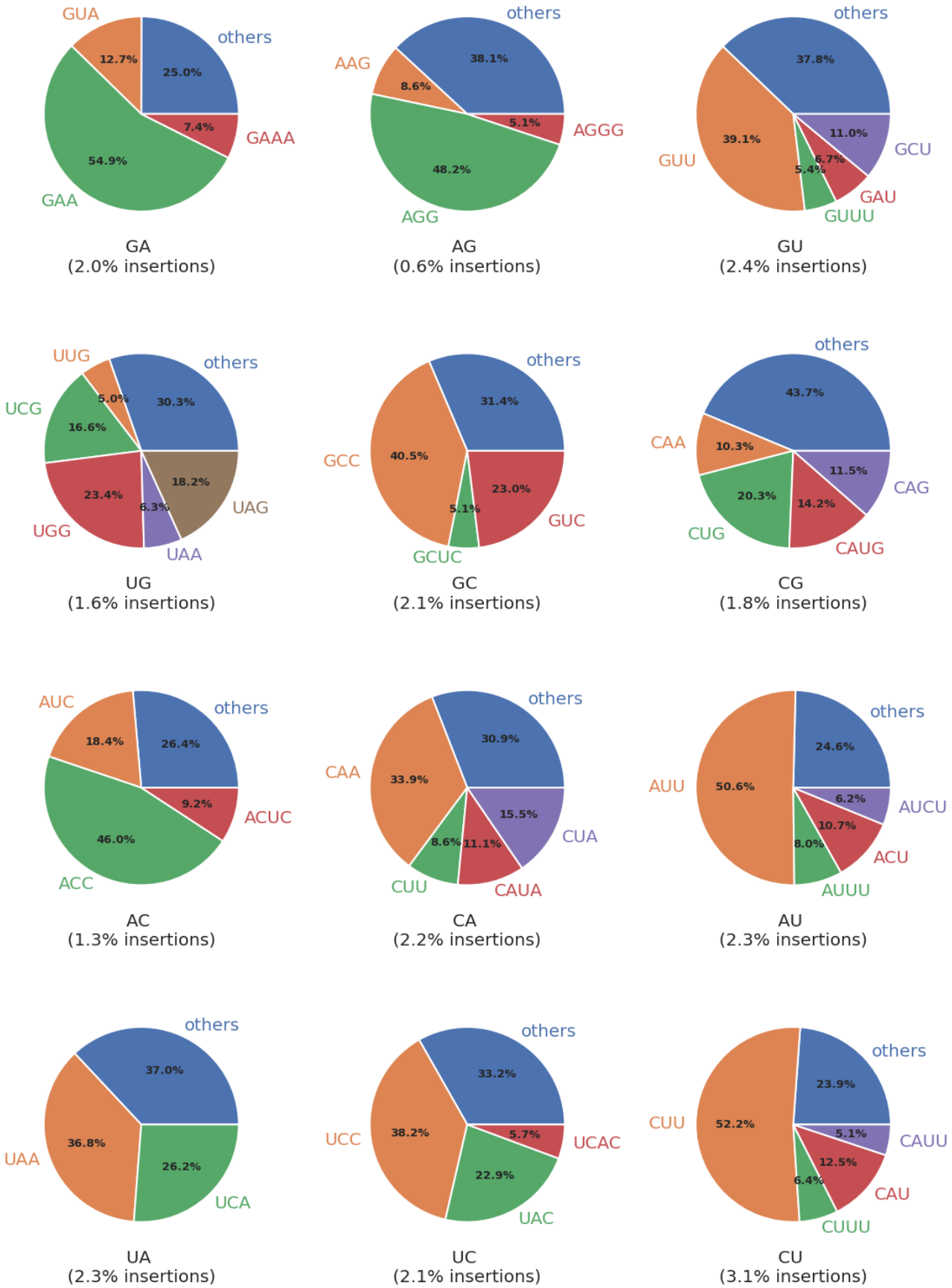

Figure 3: **Insertion profiles of 2-mer motifs.** Pie charts showing the proportion of differnt insertion errors, based on the native human sample. Inserted motifs fewer than 5% of the total number of insertions are grouped into "others".

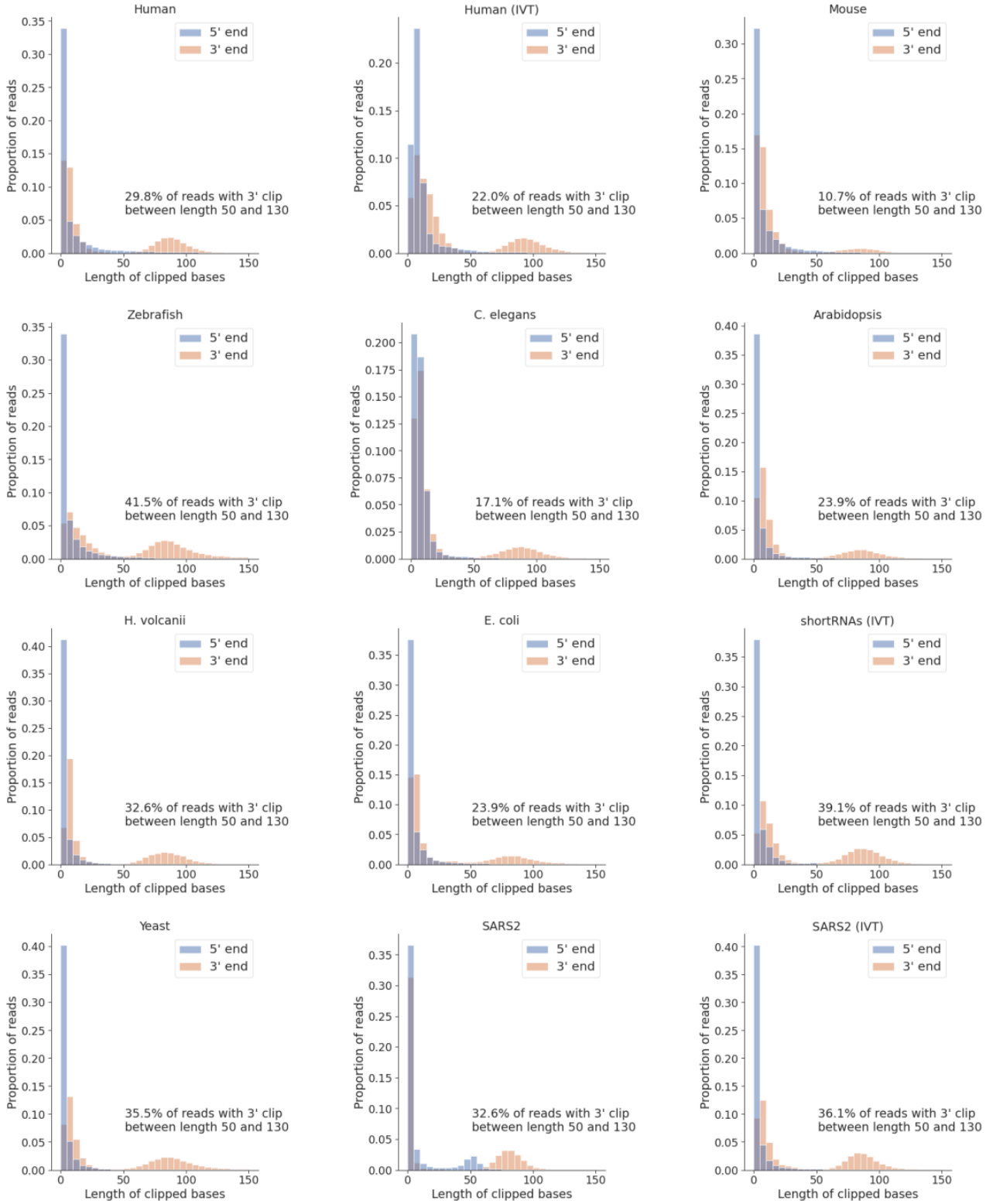

Figure 4: **Bias in soft-clipped bases at the 3' end.** The length distribution of the soft-clipped bases at the 5' and 3' ends of reads in each dataset in this study.

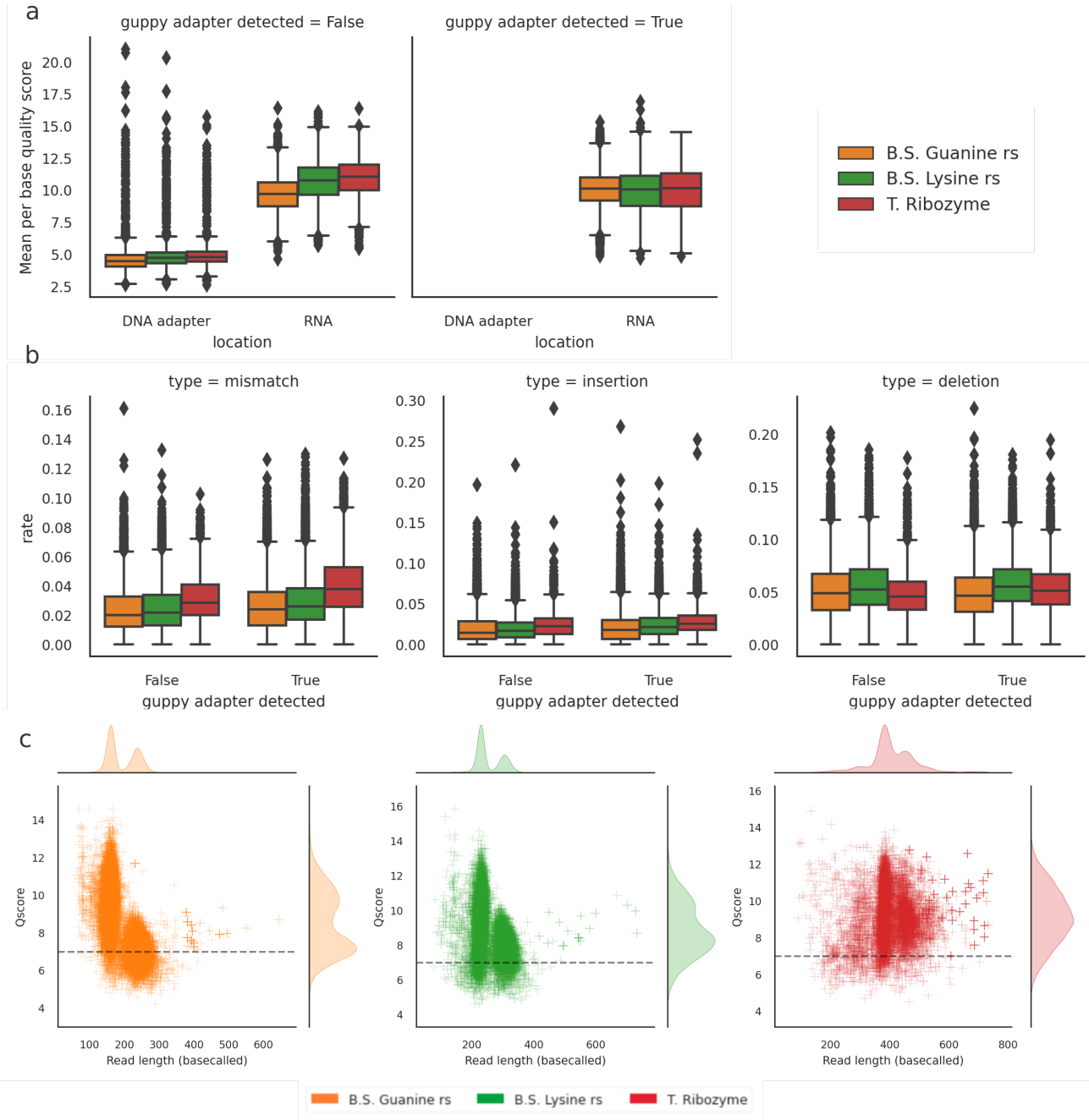

Figure 5: **Comparison of reads with and without successful adapter detection.** a) Basecalls in the DNA adapter are of low-quality, but adapter detection failure does not appear to affect the quality scores of the subsequent RNA basecalls. b) Failure to detect the adapter does not affect the error rates of the RNA basecalls after alignment. c) The impact of adapter detection failure on basecalled read length and read Q-score in short reads. Adapter detection failure causes bimodality in basecalled read length. The spurious DNA basecalls can significantly impact the read Q-score, potentially leading to data loss. The dashed line represents the conventional filter value of 7. Plots are based on the IVT short RNA dataset.

Dataset	In training data	Reads aligned	med. Length	Acc	Mis	Ins	Del
Human	**yes**	1648051 (+164296)	725 (-151)	0.923 (+0.022)	0.019 (-0.008)	0.021 (-0.004)	0.035 (-0.009)
Human (IVT)	no	1729393 (-286557)	375 (-111)	0.913 (-0.003)	0.023 (+0.001)	0.027 (+0.005)	0.037 (-0.002)
Mouse	no	85286 (-96691)	591 (-25)	0.903 (+0.025)	0.018 (-0.017)	0.015 (-0.01)	0.059 (-0.0)
Zebrafish	no	155285 (-398319)	659 (-180)	0.818 (-0.048)	0.051 (+0.014)	0.018 (-0.002)	0.114 (+0.041)
C. elegans	**yes**	222931 (-4441)	558 (-130)	0.941 (+0.026)	0.012 (-0.008)	0.014 (-0.005)	0.031 (-0.013)
Arabidopsis	**yes**	1097675 (+86732)	471 (-407)	0.903 (-0.008)	0.032 (+0.011)	0.028 (+0.007)	0.033 (-0.012)
H. volcanii	no	14258 (-8057)	959 (+454)	0.916 (+0.011)	0.028 (-0.0)	0.035 (+0.013)	0.021 (-0.022)
E. coli	**yes**	506596 (+308148)	578 (-84)	0.921 (+0.045)	0.025 (-0.015)	0.023 (-0.01)	0.026 (-0.023)
shortRNAs (IVT)	no	41329 (-3699)	170 (+10)	0.919 (+0.02)	0.020 (-0.005)	0.026 (+0.007)	0.027 (-0.024)
Yeast	no	372364 (+47475)	297 (-17)	0.892 (-0.006)	0.028 (+0.004)	0.026 (+0.002)	0.060 (+0.01)
SARS2	no	858189 (+282516)	1370 (-708)	0.916 (+0.002)	0.019 (-0.003)	0.017 (-0.003)	0.041 (-0.003)
SARS2 (IVT)	no	2452303 (+135668)	1103 (-469)	0.920 (+0.012)	0.018 (-0.005)	0.021 (-0.002)	0.033 (-0.011)

Table 2: **Basecalling performance of RODAN versus Guppy across organisms.** The differences in each metric from Guppy are indicated in brackets.

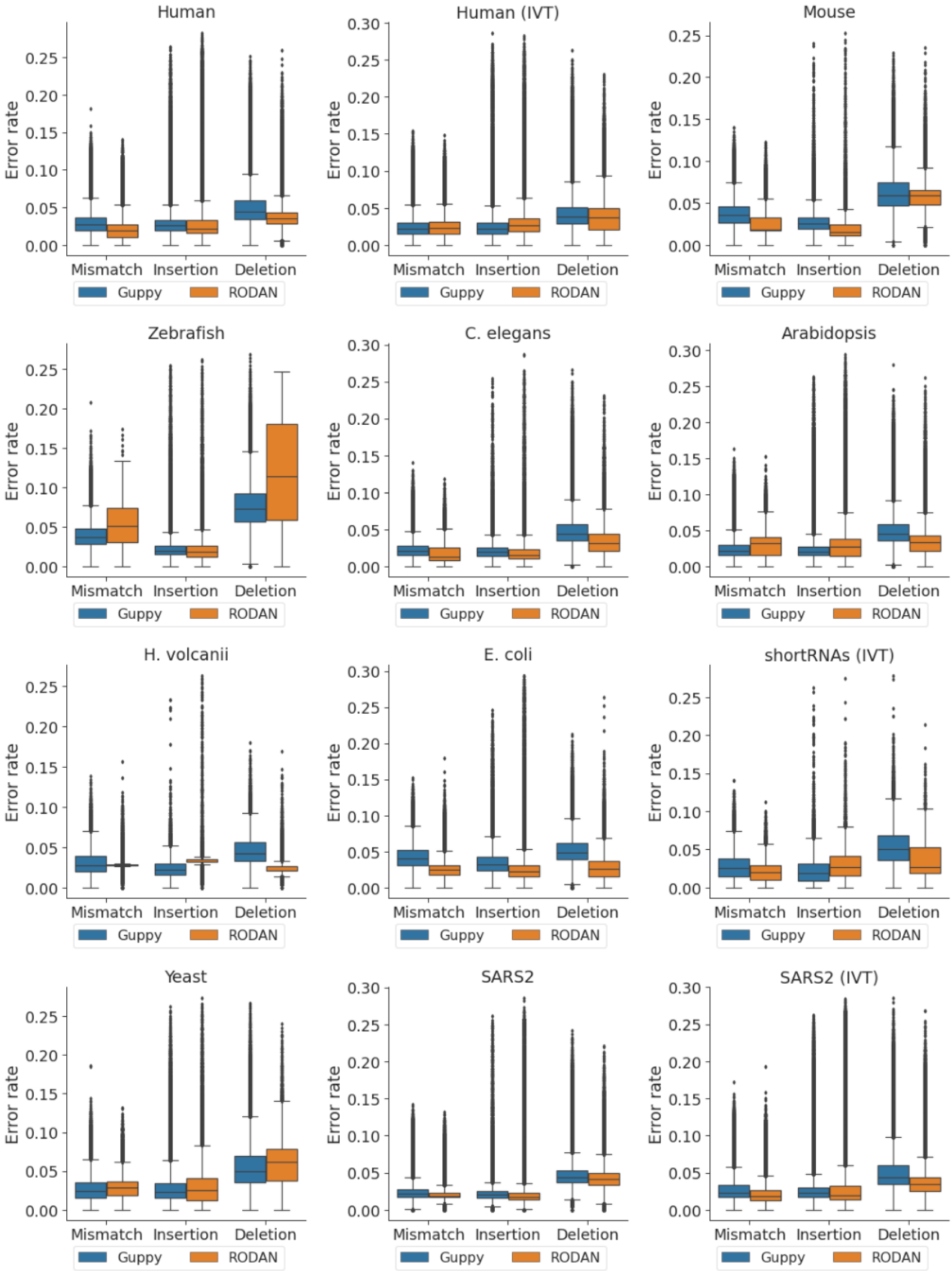

Figure 6: **Comparison of basecalling errors between RODAN and Guppy.** The mismatch, insertion and deletion rates for each dataset. Reads with accuracy lower than 70% (or equivalently, more than 30% errors), are filtered for visualization.

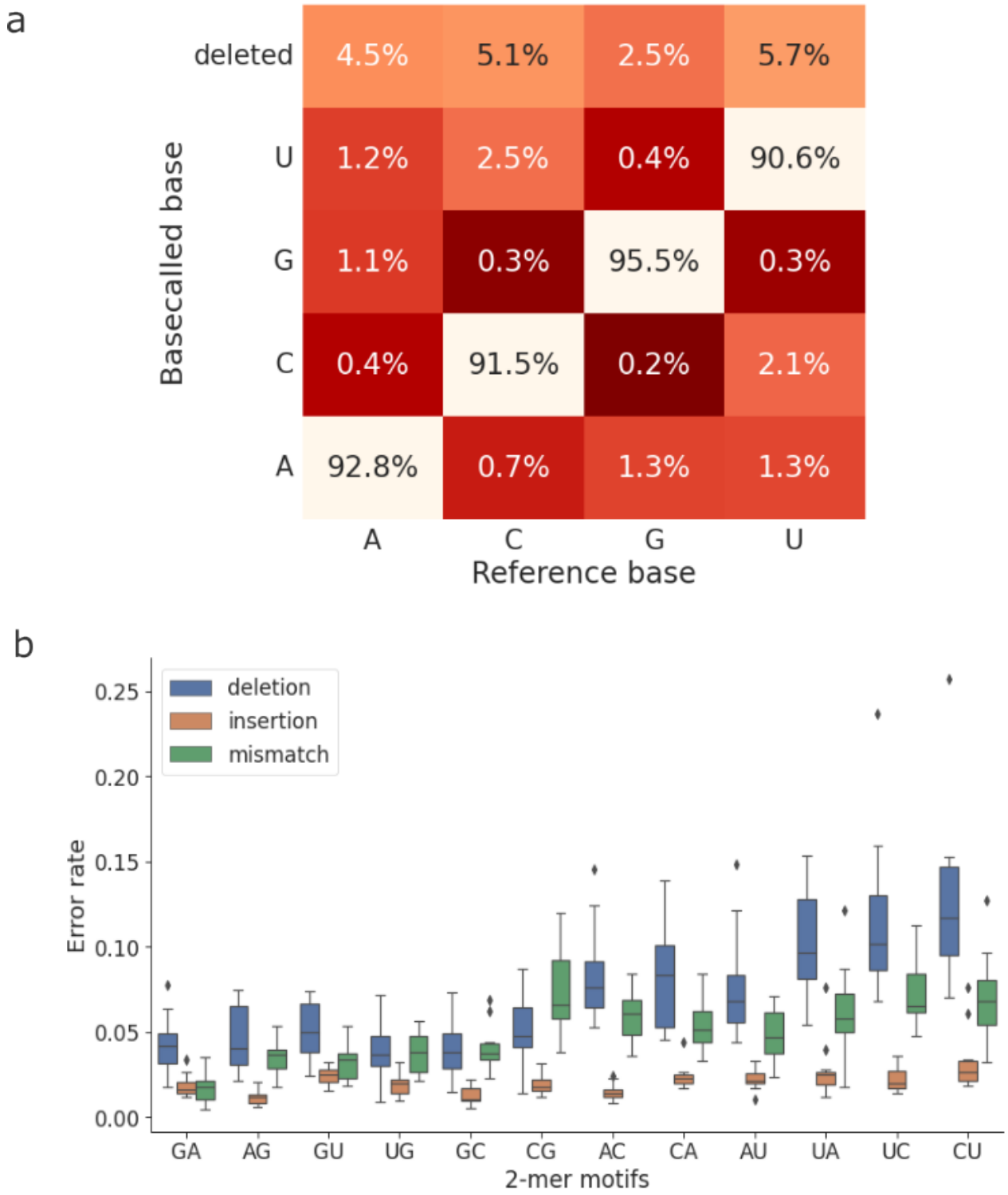

Figure 7: **Systematic errors in RODAN.** a) Confusion matrix showing the frequencies of each base being correctly basecalled, miscalled or deleted, computed by taking the mean of all samples basecalled by RODAN. b) The error rate of 2-mer motifs across the samples basecalled by RODAN, grouped by error type. Each data point corresponds to the mean error rate of the motif in one dataset.