Clair3-RNA: A deep learning-based small variant caller for long-read RNA sequencing data

Zhenxian Zheng[1,#], Xian Yu[1,#], Lei Chen[1], Yan-Lam Lee[1], Cheng Xin[1], Angel On Ki Wong[1], Miten Jain[2,3], Rupesh K Kesharwani[4], Fritz J Sedlazeck[4,5,6,*], Ruibang Luo[1,*]

[1] Department of Computer Science, School of Computing and Data Science, University of Hong Kong, Hong Kong, China

[2] Department of Bioengineering, Northeastern University, Boston, MA, USA

[3] Department of Physics, Northeastern University, Boston, MA, USA

[4] Human Genome Sequencing Center Baylor College of Medicine, Houston, TX, USA

[5] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

[6] Department of Computer Science, Rice University, Houston, TX, USA

[#] These authors contributed equally to this work

[*] To whom correspondence should be addressed. Email: Fritz.Sedlazeck@bcm.edu, rbluo@cs.hku.hk

Abstract

Variant calling using long-read RNA sequencing (lrRNA-seq) can be applied to diverse tasks, such as capturing full-length isoforms and gene expression profiling. It poses challenges, however, due to higher error rates than DNA data, the complexities of transcript diversity, RNA editing events, etc. In this paper, we propose Clair3-RNA, the first deep learning-based variant caller tailored for lrRNA-seq data. Clair3-RNA leverages the strengths of the Clair series' pipelines and incorporates several techniques optimized for lrRNA-seq data, such as uneven coverage normalization, refinement of training materials, editing site discovery, and the incorporation of phasing haplotype to enhance variant-calling performance. Clair3-RNA is available for various platforms, including PacBio and ONT complementary DNA sequencing (cDNA), and ONT direct RNA sequencing (dRNA). Our results demonstrated that Clair3-RNA achieved a ~91% SNP F1-score on the ONT platform using the latest ONT SQK-RNA004 kit (dRNA004) and a ~92% SNP F1-score in PacBio Iso-Seq and MAS-Seq for variants supported by at least four reads. The performance reached a ~95% and ~96% F1-score for ONT and PacBio, respectively, with at least ten supporting reads and disregarding the zygosity. With read phased, the performance reached ~97% for ONT and ~98% for PacBio. Extensive evaluation of various GIAB samples demonstrated that Clair3-RNA consistently outperformed existing callers and is capable of distinguishing ~67% and ~93% RNA high-quality editing sites on ONT dRNA004 and PacBio Iso-Seq

34 datasets, respectively. Clair3-RNA is open-source and available at (https://github.com/HKU-
35 BAL/Clair3-RNA).
36
37 **Introduction**
38 RNA sequencing (RNA-seq) is extensively utilized for gene expression quantification, isoform
39 identification and quantification, and gene fusion analysis [1, 2]. Recent advancements highlight the
40 potential of variant calling [3-6]in RNA-seq, showcasing its numerous advantages for this application.
41 First, it is more cost-effective than DNA sequencing (DNA-seq). Second, the presence of multiple
42 isoforms of some genes increases sequencing depth, enhancing the efficacy of variant detection. Third,
43 RNA-seq allows the identification of post-transcriptional modifications [7], which are not detectable
44 using DNA-seq.
45
46 Long-read RNA sequencing (lrRNA-seq) has become popular for its capability to capture full-length
47 isoforms [8] and novel transcript discovery [9]. Pacific Biosciences (PacBio) and Oxford Nanopore
48 Technology (ONT) are two leading platforms in this field. PacBio introduced Iso-Seq for full-length
49 cDNA sequencing using PacBio SMRT sequencing technology [10]. PacBio also provides the advanced
50 Multiplexed Arrays Sequencing (MAS-Seq) method, which increases throughput by combining cDNA
51 molecules into longer concatenated fragments. ONT also introduced direct RNA (dRNA) sequencing,
52 which is a nanopore-based technique that directly sequences RNA without enzymatic conversions [11].
53 ONT dRNA sequencing enables the observation of full-length transcripts, RNA modifications, and RNA
54 epigenetics analysis [12], which are challenging to detect using cDNA sequencing. Recently, deep-
55 learning methods have improved the capabilities of germline and somatic DNA variant calling [13-15]
56 and short-read RNA variant calling [3]. However, there exists no deep learning-based caller designed
57 specifically for variant calling in lrRNA-seq yet.
58
59 RNA-seq offers advantages for variant calling and interpretation, such as that the identified allele is
60 guaranteed to be expressed in contrast to DNA variant calling. Nevertheless, several disadvantages
61 need to be considered such as a higher error rate than DNA-seq, which averages a 1-5% error rate.
62 This necessitates robust variant-calling systems for distinguishing true variants from sequencing
63 artifacts. Second, unlike the uniform read coverage typically observed in DNA-seq data [4], the
64 coverage is uneven across genomic regions in RNA-seq, and the variable coverage poses challenges
65 for accurate variant calling, particularly in regions characterized by inadequate read support or
66 excessive read coverage that is not accounted for in neural network training. Additionally, RNA-seq
67 faces RNA editing events, especially A-to-I editing catalyzed by adenosine deaminases acting on RNA

(ADAR) [16], leading to A-to-G or T-to-C substitutions that can resemble genuine variants, thus introducing false positives in RNA-seq. Moreover, while phasing has been demonstrated to be advantageous for lrDNA-seq analyses [13, 14], no studies have yet investigated the impact of phasing on variant calling performance in lrRNA-seq, particularly when reads are shorter and zygosity switching can happen.

In this study, we introduce Clair3-RNA, the first deep learning-based small-variant caller designed for long-read RNA sequencing data. Clair3-RNA utilizes a Bi-LSTM pileup model with multi-task output for variant prediction. The model was trained on lrRNA-seq datasets using various techniques, such as coverage normalization, variant zygosity switching, training sample augmentation, and RNA editing-site tagging to optimize variant-calling accuracy. We further developed a workflow that incorporates additional phasing information into the neural network, thereby enhancing variant calling performance and demonstrating the efficacy of phasing for lrRNA-seq. We also designed a benchmarking pipeline that defines callable regions and benchmarking metrics, enabling a fair evaluation of variant-calling performance. Our benchmarking results demonstrate that Clair3-RNA outperformed other variant callers and can achieve ~93%, ~93%, and ~91% F1-score in PacBio Iso-Seq, PacBio MAS-Seq, and the ONT dRNA004 platform, respectively, for variants with at least four supporting reads. With the incorporation of additional pileup phasing features, Clair3-RNA reached 98.59% SNP F1-score for PacBio MAS-Seq, 98.24% for Iso-Seq, and 97.16% for ONT dRNA004 when excluding zygosity and requiring a minimum of 10 supporting reads. Furthermore, in our analysis of GENCODE-annotated genes, we identified 1,085 and 507 genes that could be completely phased, as well as 1,165 and 655 genes without SNP, Indel or switch errors in the PacBio MAS-Seq HG004 and ONT dRNA004 HG004 datasets, respectively. Furthermore, we established a high-quality benchmark set of RNA editing sites by cross-evaluating RNA samples with corresponding DNA samples. Our results show that Clair3-RNA can identify 93.27% and 67.07% of editing sites in PacBio and ONT HG004 datasets, respectively. Clair3-RNA also demonstrated outperforming runtime and memory usage compared to other DNA variant callers, requiring ~20 minutes for PacBio Iso-Seq and ~30 minutes for ONT dRNA004 variant calling using 16 CPU cores, with a memory footprint of about 7 GB.

Results

Overview of Clair3-RNA

Clair3-RNA is a multi-task bidirectional long short-term memory (Bi-LSTM) neural network designed for lrRNA-seq variant calling. Clair3-RNA takes a BAM/CRAM file as input to identify candidate variants with any minor allelic fraction exceeding a specific threshold (typically 0.08, configurable by option)

and then computes a summarized pileup tensor of the candidate and employs a neural network for prediction. Clair3-RNA includes two probabilistic output tasks - 21-genotype and zygosity - facilitating the prediction of variants and artifacts. The **Methods** section details techniques for training data augmentation and optimizations specific to lrRNA-seq, addressing challenges like uneven coverage, variant zygosity inconsistency, and RNA editing events. Clair3-RNA was trained with the Genome in a Bottle (GIAB) HG002 sample and supports variant calling on multiple PacBio and ONT instruments and chemistry configurations.

In this section, we assess the performance of Clair3-RNA on PacBio cDNA, ONT cDNA, and ONT dRNA datasets. Our evaluation involved comparing results with four other variant callers, including LongcallR [17] (a long-read SNP caller for RNA data), Clair3 [13] (our deep learning-based germline variant caller for DNA data), and DeepVariant [14] (the first deep learning-based variant caller, developed by Google). For Clair3-RNA and LongcallR, we used the corresponding specific RNA model and parameter settings for RNA variant calling. Although DeepVariant supports RNA variant calling [3], it is optimized for the Illumina platform, not for lrRNA-seq. Clair3 and the other tools in the series [15, 18-21] before Clair3-RNA were used for germline and somatic small variant calling, but none of them were designed for lrRNA-seq. The performance of Clair3 and DeepVariant on lrRNA-seq is not optimized and is provided solely for reference purposes. Therefore, for performance comparison on long-read data, the best available DNA models of Clair3 and DeepVariant were employed for benchmarking. We used two GIAB samples, HG004 and HG005, for benchmarking. Since LongcallR is limited to SNP variant detection, the comparisons with Clair3-RNA were conducted on SNP performance only. For other methods, we benchmarked both SNP and Indel performance.

Generating callable high-confidence benchmarking regions for evaluation

RNA transcripts represent only subsets of the corresponding DNA sequence. Certain regions may have insufficient coverage, resulting in undetectable variants in those regions. To address this, we identified callable high-confidence benchmarking regions by intersecting callable regions with adequate read coverage, with those high-confidence regions defined by GIAB. We also considered the zygosity imbalance in the RNA data. In certain cases, a heterozygous variant in DNA could be either a homozygous variant or a homozygous reference in RNA. This discrepancy arises from the dominant expression of one of the haplotypes. By disregarding zygosity, we aim to understand the impact of zygosity during evaluation better. More details are listed in the **Methods** - **Benchmarking metrics** section.

136 We further introduced two key benchmarking factors for performance assessment: read coverage (DP)
137 and allele depth (AD). DP signifies the read coverage of the candidate site meeting minimum mapping
138 quality support and excluding all supplementary and secondary alignments. AD represents the read
139 coverage that supports the true alternate allele at the site. We also considered RNA editing events
140 and created a reliable benchmark set of editing sites for comprehensive evaluation. More details are
141 provided in the **Methods** - **Orthogonal high-quality RNA editing benchmark set generation** section.
142
143 **Performance on ONT**
144 The performance on the ONT platform was carried out in six datasets. Details of the benchmarking
145 datasets are provided in **Supplementary Table 1**. These ONT datasets include ONT cDNA sequencing,
146 dRNA sequencing using the SQK-RNA002 kit (hereafter referred to as dRNA002), and dRNA sequencing
147 employing the ONT latest SQK-RNA004 kit (hereafter dRNA004). The ONT cDNA and dRNA002 data
148 were made available by GIAB. The ONT cDNA data was sequenced using R9.4.1 chemistry and
149 basecalled using Guppy [22] version 6.4.6. The ONT dRNA002 data was sequenced using the dRNA
150 SQK-RNA002 sequencing kit and basecalled with Guppy 6.4.6. In 2023, ONT released the dRNA SQK-
151 RNA004 kit [23] with an enhanced throughput and a reduced error rates compared to the SQK-RNA002
152 kit. We sequenced GIAB HG002, HG004, and HG005 using the dRNA SQK-RNA004 kit at the University
153 of Hong Kong. The total RNA of three samples was acquired from the Coriell Institute, sequenced using
154 the SQK-RNA004 kit, and basecalled using the Dorado basecaller [24] version 0.7.2. All basecalled
155 alignments were mapped to the GRCh38 reference genome using the minimap2 aligner in splicing
156 alignment mode. More details can be found in the **Methods - ONT SQK-RNA004 kit library**
157 **preparation and sequencing** section. As shown in **Figure 2a**, the overall throughput increased from an
158 average of 3 GB to 18 GB per flowcell using dRNA004 compared with dRNA002 and the average error
159 rate decreased from 9.7% to 1.8%.
160
161 As shown in **Figure 3b** and **Supplementary Table 2**, Clair3-RNA consistently outperformed other
162 variant callers in the ONT dRNA004 and cDNA datasets for both the HG004 and HG005 samples.
163 Specifically, Clair3-RNA achieved 91.00% and 91.73% F1-score in HG004 and HG005, respectively,
164 using dRNA004 sequencing when DP⩾4 and AD⩾2. In comparison, LongcallR had HG004/61.07% and
165 HG005/65.20% F1-score, respectively; Clair3 had HG004/74.85% and HG005/77.21%, and
166 DeepVariant had HG004/57.94% and HG005/56.96% F1-score. LongcallR employs a Bayesian
167 statistical model for variant calling, demanding higher sequencing coverage for accuracy, as suggested
168 by the authors [17]. When we increased the minimal requirement of DP to 10 and AD to 4, we
169 observed that LongcallR performance improved; the F1-score reached HG004/72.42% and

170 HG005/74.42%, resulting in an average ~10% improvement against the default setting. Using the same
171 setting, while the performance gap narrowed, Clair3-RNA still performed better than longcallR,
172 reaching HG004/94.75% and HG005/95.17%, respectively.
173
174 In the cDNA dataset, Clair3-RNA yielded HG004/68.60% and HG005/79.83% F1-score and achieved
175 better performance than longcallR (HG004/+23.72%, HG005/+23.72%), Clair3 (HG004/+17.85%,
176 HG005/+19.19%) and DeepVariant (HG004/+18.60%, HG005/+30.99%). The lower performance across
177 all callers might be because of a higher base error rate of the cDNA dataset generated with R9.4.1
178 chemistry. We expect the performance to improve notably with cDNA sequencing using R10 chemistry.
179
180 **Boosting the performance across ONT RNA-specific flow cells**. In our evaluation, we compared the
181 performance of different callers in the GIAB HG004 and HG005 dRNA002 and dRNA004 datasets. As
182 illustrated in **Figures 2b and 2c,** the data throughput increased ~ six times per flowcell, from ~3GB in
183 dRNA002 to ~18GB in dRNA004. The read N50 also increased from ~1100bp in dRNA002 to ~1500bp
184 in dRNA004. Across all callers, we observed improvements in precision, recall and F1-score, signifying
185 improved read accuracy in the dRNA004 data. Notably, Clair3-RNA exhibited an average F1-score
186 improvement of 22.07% (rising from 68.61% to 91.00% in HG004 and from 69.98% to 91.73% in
187 HG005), and LongcallR improvement by 16.35%, Clair3 by 48.69%, and DeepVariant by 4.36%.
188
189 Among the three ONT benchmarking datasets, we noted that Clair3 had lower precision and
190 DeepVariant had lower recall. In the dRNA004 dataset, Clair3 had only HG004/66.93% and
191 HG005/70.33% for precision, but HG004/84.89% and HG005/85.58% for recall. DeepVariant had
192 HG004 /86.53% and HG005/87.14% for precision, but almost half the recall with HG004/43.55% and
193 HG005/42.31%. LongcallR also had better precision than its recall, with HG004/94.49% and
194 HG005/94.89% precision, but HG004/45.11% and HG005/49.67% recall. Compared with other callers,
195 Clair3-RNA demonstrated balanced precision and recall, underscoring the necessity of a variant caller
196 specific for lrRNA-seq.
197
198 **Performance on PacBio**
199 We benchmarked the performance on the PacBio platform using two sequencing technologies, Iso-
200 Seq and MAS-Seq, both developed for PacBio cDNA sequencing. We collected the HG004 and HG005
201 Iso-Seq RNA sequencing datasets from two sources, Baylor College of Medicine (BCM) and Google,
202 and the HG004 MAS-Seq RNA sequencing dataset from BCM. More details of sequencing data are
203 provided in **Supplementary Table 1**. Compared with ONT sequencing, PacBio MAS-Seq and Iso-Seq

204 have a higher N50 read length, with ~3000bp for PacBio Iso-Seq and ~1700bp for MAS-Seq, versus
205 ~1200bp for ONT sequencing.
206
207 As indicated in **Figure 3** and **Supplementary Table 3**, Clair3-RNA consistently achieved a SNP F1-score
208 exceeding 92% across all Iso-Seq and MAS-Seq datasets for HG004 and HG005, showcasing better
209 performance than the second-best LongcallR, with an F1-score ranging from 71% to 80%. Clair3-RNA
210 also demonstrated consistent performance across all PacBio sequencing sources, achieving
211 HG004/94.93% and HG005/95.50% F1-score for BCM Iso-Seq, HG004/91.84%, an HG005/92.13% F1-
212 score for Google Iso-Seq, and an HG004/93.38% F1-score for BCM MAS-Seq. In terms of precision and
213 recall balance, Clair3-RNA exhibited a balanced performance with 95.25% precision and 91.93% recall.
214 Notably, Clair3 maintained an average recall rate of 95.08%, but had a low precision (63.76%).
215 Conversely, LongcallR had 93.94% precision but 65.31% recall and DeepVariant showed lower
216 precision (73.25%) and recall (27.15%).
217
218 The PacBio datasets were originally aligned with the GRCh38 reference genome using pbmm2, a
219 minimap2 aligner wrapper specifically designed for the PacBio platform. During our experimental
220 analysis focusing on alignment methods, we observed that the pbmm2 aligner tended to introduce
221 more false positives and false negatives compared to the minimap2 aligner when using Clair3-RNA for
222 small variant calling on PacBio. We aligned five datasets using pbmm2 and minimap2 splice mode to
223 assess the impact of different alignment strategies. Further details regarding the alignment procedure
224 are included in the **Supplementary Notes** - **Command line used** section.
225
226 As shown in **Figure 8** and **Supplementary Table 5**, when utilizing minimap2 in place of pbmm2, Clair3-
227 RNA's performance improved across all datasets. The average number of false positives and false
228 negatives decreased by 25.42% and 0.79%, respectively. The averaged F1-score of Clair3-RNA
229 improved from 94.45% to 92.21% (+0.77%) for BCM Iso-Seq, from 91.37% to 91.98% (+0.61%) for
230 Google Iso-Seq, and from 92.76% to 93.38% (+0.62%) for BCM MAS-Seq. Similar trends were observed
231 in other callers; longcallR, Clair3, and DeepVariant demonstrated average F1-score improvements of
232 0.46%, 14.40%, and 2.17%, respectively. These improvements suggest that using minimap2 had better
233 performance compared to using pbmm2 when using the default splice mode setting in PacBio.
234
235 **Sequencing coverage and allele depth analysis**
236 Insufficient coverage and inadequate read support for alternate bases may lead to false negatives in
237 evaluations. **Figure 4a** shows the benchmarking performance across different DP and AD thresholds.

238 In the ONT dRNA004 HG004 dataset, increasing the depth from 4 to 8 resulted in a 1.54% increase in
239 precision and a 2.67% improvement in recall. Raising the depth to 10 showed marginal gains of +0.33%
240 in precision and +0.58% in recall, highlighting the difficulty in capturing low-coverage variants. Similar
241 trends were observed in the PacBio datasets. In the BCM Iso-Seq HG004 dataset, raising the depth
242 benchmark from 4 to 8 led to precision and recall improvements of 2.15% and 1.90%, respectively.
243 Increasing to a depth of 10 showed modest gains of 0.38% in precision and 0.18% in recall. AD is
244 another crucial factor for variant evaluation. When the AD was increased from 2 to 4, the ONT
245 dRNA004 HG004 dataset showed a 0.28% improvement in precision and a 1.91% increase in recall.
246 Similarly, in the BCM Iso-Seq HG004 dataset, there was an improvement of 0.51% in precision and
247 1.51% in recall. This highlights the importance of allele depth in accurately calling missing variants.
248
249 **Zygosity switch analysis in RNA dataset**
250 Zygosity switch is also common in RNA data due to factors like differential gene expression and RNA
251 splicing [3]. To evaluate this, we used the Clair3-RNA's "calculate_overall_metrics" function to
252 consider the zygosity switch by disregarding zygosity in benchmarking. As shown in **Figure 4a**,
253 **Supplementary Table 2, and Supplementary Table 3**, disregarding zygosity led to an average 2.43%
254 and 2.88% improvement in the F1-score on the ONT and PacBio, respectively. Benchmarking on the
255 ONT dRNA004 HG004 dataset, disregarding zygosity resulted in a reduction of 2,699 FPs and 2,698
256 FNs, enhancing the F1-score by 2.76%. Similarly, in the BCM Iso-Seq HG004 dataset, the number of
257 FPs decreased by 4334 and FNs by 4331, leading to an improvement of 2.69% in the F1-score.
258 Moreover, the impact of disregarding zygosity lessened as DP increased. For instance, in the ONT dRNA
259 HG004 dataset, setting DP⩾8 while considering zygosity boosted the F1-score by 4.19% over DP⩾4,
260 whereas disregarding genotyping resulted in a 2.45% F1-score improvement. These results emphasize
261 the importance of read coverage in variant detection, especially when incorrect genotypes are
262 introduced without adequate reads to determine variant zygosities.
263
264 **Performance analysis by genomic context**
265 We also carried out benchmarking in challenging genomic regions. Specifically, we compared the SNP
266 performance of Clair3-RNA and LongcallR in areas defined as difficult by GIAB Stratification V3 [25].
267 We categorized the analysis into five groups: “Low complexity”, “Segmental duplications”, “Low
268 mappability”, “Functional regions”, and “Other difficult regions” – and conducted it on ONT dRNA004
269 HG004, PacBio BCM Iso-Seq HG004, and PacBio MAS-Seq HG004 datasets. As shown in **Figure 4b** and
270 **Supplementary Table 6**, Clair3-RNA consistently outperformed LongcallR in SNP calling accuracy.
271

272 Benchmarking on the ONT dRNA004 dataset, Clair3-RNA demonstrated varying performance,
273 achieving an average 80% F1-score. Notably, Clair3-RNA excelled in coding regions, attaining a 94.58%
274 F1-score in the Coding Sequence (CDS) regions. In contrast, its performance dropped to 61.80% in low
275 complexity regions characterized by homopolymer regions greater than 11 bp. In comparison,
276 LongcallR exhibited more significant variability on the same dataset, peaking at 82.51% in the CDS
277 regions, while dropping to 29.85% in homopolymer regions greater than 11 bp. Similar trends were
278 observed in the PacBio Iso-Seq and MAS-Seq datasets, Clair3-RNA's F1-score averaged ~86%, with the
279 highest at 95.22% in the CDS regions and the lowest at 74.76% in the homopolymer regions. In this
280 dataset, LongcallR averaged about 66%, with a peak of 91.74% in the CDS regions and a low of 37.07%
281 in the homopolymer regions. Among all categories, "Low complexity" and "Segmental duplications"
282 were the most challenging for variant detection, with the lowest average F1-score of all categories.
283
284 **Indel performance**
285 In our evaluation, we compared the performance of various callers, excluding LongcallR, which
286 identifies only SNP variants. For the ONT platform, as shown in **Supplementary Table 2**, using DP⩾10
287 and AD⩾4 coverage settings, Clair3-RNA achieved HG004/66.50% and HG005/72.37% F1-score for
288 the dRNA004 dataset, and HG004/50.63% and HG005/60.88% for the cDNA dataset. It consistently
289 outperformed both Clair3 (dRNA004 HG004/52.65% and HG005/60.64%, cDNA HG004/45.63% and
290 HG005/50.04%) and DeepVariant (dRNA HG004/40.03% and HG005 44.94%, cDNA HG004/35.95% and
291 HG005 36.82%) in both the cDNA and dRNA004 datasets. In a comparison between the dRNA004 and
292 dRNA002 datasets, Clair3-RNA demonstrated a significant performance improvement of 43.11%, with
293 the average F1-score increasing from 26.33% to 69.44%. Even in the most precise dRNA004 dataset,
294 the averaged indel recall was lower than 60%, attributed primarily to the presence of indel signals
295 with varying lengths clustered at the same site, confusing the model from calling Indel precisely.
296
297 On the PacBio platform, as shown in **Supplementary Table 3,** using the DP⩾10 and AD⩾4 coverage
298 setting, Clair3-RNA achieved an Indel F1-score of HG004/89.42% and HG005/92.90% for BCM Iso-Seq,
299 HG004/82.33% and HG005/82.82% for Google Iso-Seq, and HG004/80.05% for BCM MAS-Seq. With
300 an average indel performance of 85.50%, Clair3-RNA significantly outperformed Clair3 and
301 DeepVariant, where the average F1-score was 63.25% (+22.25%) and 40.31% (+45.19%), respectively.
302
303 **RNA editing tagging using the REDIportal database**. The REDIportal has compiled a total of 36,234,771
304 RNA editing sites, categorized as follows: 1) 13,101,296 editing sites collected only in ATLAS (A); 2)
305 22,836,582 editing sites collected by ATLAS and RADAR (A&R); 3) 4,708 editing sites collected by ATLAS

306 and DARNED (A&D), and 4) 292,185 editing sites reported by all three sources (A&R&D). We tagged
307 the variant as an RNA editing site that was validated by at least two independent sources within the
308 REDIportal, specifically in ATLAS, DARNED, and RADAR. As shown in **Supplementary Table 4**, excluding
309 the variants tagged by the REDIportal database, there was enhanced SNP performance across all
310 platforms. Specifically, Clair3-RNA achieved a 0.82% improvement in the ONT dRNA004 HG004 dataset,
311 with a total of 1,898 editing sites tagged. For the PacBio datasets, the improvement was less significant,
312 with 1,033 and 912 editing sites tagged in the BCM Iso-Seq HG004 and MAS-Seq HG004 datasets,
313 respectively, resulting in a 0.28% and 0.13% F1-score improvement.
314
315 **A-to-G and T-to-C false positives analysis**
316 A-to-I RNA editing poses a significant challenge for RNA variant calling algorithms because they must
317 distinguish between A-to-T and T-to-C SNPs induced by editing events and genuine A-T and T-to-C
318 SNPs. To address this issue, we analyzed A-to-G and T-C SNPs identified by Clair3-RNA and Clair3 that
319 were not present in the GIAB ground truth.
320
321 The high-quality benchmark set of the editing sites were established and analyzed using the PacBio
322 HG004 BCM Iso-Seq dataset and the ONT HG004 dRNA004 dataset. The dRNA004 dataset was aligned
323 with the HG004 DNA dataset, sequenced by ONT R10.4.1 Q20+ chemistries, and basecalled using
324 Dorado, provided by Nanopore EPI2ME Labs [26]. For the BCM HG004 dataset, we compared the RNA
325 dataset with the PacBio HiFi HG004 DNA dataset generated by Google [27]. We then identified
326 potential A-to-I editing sites by looking for evident A-to-T or T-to-C substitutions in the RNA data that
327 were absent in the DNA data. More details of the benchmarking editing site generation are described
328 in the **Methods** - **Orthogonal high-quality RNA editing benchmark set generation** section. Finally, we
329 identified 3,800 and 6,086 editing sites on the ONT dRNA004 and PacBio Iso-Seq datasets, respectively.
330 46.21% (1,756 over 3,800) and 55.71% (3,391 over 6,086) of the benchmark set of editing sites were
331 found in the REDIportal database.
332
333 In the Venn diagrams presented in **Figure 5b**, the A-to-G and T-to-C false positives identified by Clair3-
334 RNA and Clair3 were extracted for comparison. Clair3-RNA reported 3,940 false positives in ONT and
335 3,979 in PacBio, significantly fewer than Clair3, which identified 30,411 false positives in ONT and
336 94,476 in PacBio. Among these false positives, 1,251 (539+712) and 409 (81+328) variants intersected
337 with the benchmark set of editing sites for ONT and PacBio, respectively, indicating that Clair3-RNA
338 can classify 67.07% (1,505+1,044 over 3,800) and 93.27% (5,677 over 6,086) high-quality RNA editing
339 sites as non-variants through network prediction. In comparison, Clair3 detected only 1,592 and 1,011

340 RNA editing sites for ONT and PacBio, respectively. The results indicate that Clair3-RNA demonstrates
341 a capability in rejecting RNA editing sites during variant detection. As shown in **Figure 5c**, for the
342 remaining false positives in Clair3-RNA, 29.87% (1,177 over 3,940) in ONT and 26.94% (1,072 over
343 3,979) in PacBio were annotated by the REDIportal database. After network prediction and REDIportal
344 database tagging, 62.77% of the false positives in the PacBio dataset and 38.37% of the false positives
345 in ONT were identified. After network prediction and editing site tagging by the REDIportal database,
346 the remaining A-to-G and T-to-C false positives can be reduced from 27,175 to 2,051 on ONT and from
347 54,942 to 2,579 on the PacBio platform. The results suggest the feasibility of utilizing the REDIportal
348 database to annotate challenging editing sites that are missed by network prediction.
349
350 **Editing sites comparison and analysis**
351 We identified the high-quality editing sites across additional datasets and analyzed their intersections,
352 as depicted in **Figure 5d**. In the Iso-Seq (Google) dataset, we identified 8,049 sites, while the MAS-Seq
353 dataset yielded 15,659 sites. It is noteworthy that the ONT dRNA004 dataset contained fewer sites
354 compared to the PacBio datasets, primarily due to the absence of a reverse transcription step in the
355 ONT protocol. This is in contrast to PacBio's cDNA preparation, which includes the reverse
356 transcription step **Supplementary Figure 2**. Additionally, the MAS-Seq dataset exhibited a significantly
357 higher number of sites than the other datasets, which can be attributed to its broader genomic region
358 coverage. This resulted in a greater number of detected variants on the MAS-Seq dataset. Among the
359 four datasets, the PacBio datasets demonstrated a higher degree of overlap with each other than with
360 the ONT dataset, highlighting the distinct methodological differences between ONT dRNA and PacBio
361 cDNA sequencing approaches.
362
363 We used the vcf2maf tool [28] to annotate the variants identified by Clair3-RNA and editing sites in
364 the PacBio Iso-Seq HG004 and ONT dRNA004 HG004 datasets. As shown in **Figure 5e and**
365 **Supplementary Table 7**, both the variants and editing sites were found to be predominantly located
366 in intronic regions. About half of the PacBio editing sites, ONT variants, and ONT editing sites were in
367 introns, while over three-quarters of PacBio variants were in intron. Editing sites, however, occur more
368 often in splice sites and splice regions. Specifically, 27.43% of the editing sites were in these regions,
369 compared to just 0.19% of the variants. In the ONT dRNA004 dataset, 5.50% of the editing sites but
370 0.53% of the variants, were found in splice sites and splice regions. These findings align with research
371 on the A-to-I editing survey [16, 29].
372

373 **Enhancing variant calling performance with phasing information**

374 A heterozygous variant identified in a single haplotype is considered more reliable than sequencing
375 errors that affect both haplotypes [13]. Based on this, we attempted to incorporate phasing haplotype
376 information into the neural network, with more details provided in the **Methods** - **Integrating phasing**
377 **information into neural network** section. In the variant phasing phase, as illustrated in **Figure 6d**, we
378 analyzed phased heterozygous variants and found that at least 63.34% of heterozygous variants can
379 be phased. The percentage rises to 78.05% and 79.07% when using the latest PacBio MAS-Seq HG004
380 and ONT dRNA004 HG004 datasets, respectively. These datasets have average read lengths of 1,731
381 and 1,404 for variant phasing.

382

383 For read haplotagging, we computed the percentage of haplotagged reads against all sequencing
384 reads, while excluding supplementary reads, secondary reads, and reads with a mapping quality lower
385 than 5, as recommended by WhatsHap[30]. **Figure 6b** and **Figure 6c** display the haplotagged
386 percentages of all datasets and the number of haplotagged reads - for PacBio MAS-Seq HG004 and
387 ONT dRNA004 HG004 datasets. The ONT dRNA004 achieved an average haplotagging rate of 26.17%,
388 with PacBio MAS-Seq 35.83%, and Iso-Seq 38.22%. The haplotagging rates of lrRNA-seq are
389 significantly higher than those of the short-read datasets, which is typically less than 10%. This
390 suggests the potential for enhancing variant detection by incorporating phasing information.

391

392 We trained the Clair3-RNA pileup model with phasing information from both the PacBio and ONT
393 platforms using the GIAB HG002 dataset. Subsequently, we evaluated the trained models on HG004
394 and HG005 datasets. As shown in **Figure 6e** and **Supplementary Table 8**, both SNP and Indel precision
395 and recall improved across all datasets. Benchmarking results for the ONT and PacBio datasets indicate
396 an overall improvement of 1.52% in the F1-score for SNP. Specifically, using the PacBio MAS-Seq
397 HG004 dataset, the F1-score increased by 1.23% (94.61% versus 93.38%). The improvement was more
398 significant in the ONT dRNA004 dataset, with an average 2.01% F1-score increase. For HG004, the F1-
399 score increased from 91.73% to 93.56%, and for HG005, it increased from 91.00% to 93.19%. With
400 phasing information, the F1-score reached 98.59% for PacBio MAS-Seq, 98.24% for Iso-Seq, and 97.16%
401 for ONT dRNA004, with at least 10 supporting reads and disregarding zygosity switches.

402

403 The improvement in Indel detection is more significant compared to SNP, achieving an average 5.86%
404 F1-score improvement across all benchmarking datasets. PacBio Iso-Seq demonstrates greater
405 improvement than MAS-Seq (6.54% versus 5.01%) and ONT dRNA004 (6.54% versus 4.92%). The Indel
406 performance reached 85.61% F1-score in the PacBio Iso-Seq HG005 dataset, representing a 4.21%

407 improvement. ONT dRNA004 HG005 achieved a 70.63% F1-score with a 4.58% improvement. By
408 increasing the minimum thresholds to DP 10, AD 4, and disregarding zygosity switches, the average
409 Indel F1-score can reach 89.42% for PacBio and 72.56% for ONT. Notably, the highest Indel F1-score
410 achieved was 95.17% for PacBio Iso-Seq HG005 and 75.51% for ONT dRNA004 HG005.
411
412 **GENCODE annotated gene analysis**
413 We evaluated the Clair3-RNA variant calls across gene regions to assess its applicability for functional
414 studies in lrRNA-seq. We used the annotated genes from the GENCODE (version 46) database [31]. For
415 gene region analysis, we selected those genes that overlapped the defined callable regions specified
416 in the **Methods** - **Benchmarking metrics** section.
417
418 We first categorized all callable genes into three distinct groups based on their overlap with callable
419 high-confidence regions. This overlap ratio *H* was computed how many percent of a gene is in callable
420 high-confidence regions, defined as the intersected regions of the callable regions and GIAB high-
421 confidence regions. More details are given in the **Methods** - **Analysis of overlap and error statistics**
422 **for GENCODE annotated gene** section. Genes were classified as wholly (H=100%), partially
423 (75%<H<100%), or not (H=0%). We then analyzed how many genes are with variants covered and has
424 no calling mistake in each dataset. Furthermore, we compared these genes to the phasesets computed
425 by WhatsHap to determine how many of them are wholly phased.
426
427 **Figure 7a** presents the number of genes covered by callable high-confidence regions for both the
428 PacBio MAS-Seq HG004 and ONT dRNA004 HG004 datasets. For PacBio, a total of 3,859 genes were
429 fully covered, 3,589 were partially covered, and 1,135 were not included in defined callable high-
430 confidence regions. In contrast, the ONT platform showed 2,199 fully covered genes, 1,870 partially
431 covered, and 876 not included in callable high-confidence regions. Further analysis revealed the
432 number of genes without SNP error. In the PacBio dataset, 1,279 genes (33.1% of fully covered) and
433 2,404 genes (67.0% of partially covered) were without SNP error. For the ONT platform, 735 genes
434 (33.4% of fully covered) and 1,167 genes (62.4% of partially covered) were without SNP error. Note
435 that the SNP errors in the "not in callable high-confidence region" category could not be determined,
436 because they fall outside the GIAB high-confidence region.
437
438 We also summarized the numbers by looking into only those genes annotated as “protein_coding” in
439 the GENCODE database, which are more important for studying protein synthesis and biological
440 functions. As shown in **Figure 7b**, we identified a total of 449 and 375 protein-coding genes in PacBio

441 and ONT, respectively. Among these, 242 and 186 genes were detected without SNP error, while 85
442 and 49 genes were wholly phased in PacBio and ONT. When examining only the coding sequence (CDS)
443 regions, we found that 263 genes (58.6% of the protein-coding genes) in PacBio and 227 genes (58.7%)
444 in ONT were free of SNP errors. Considering only the exon regions, 242 genes (53.9% of the protein-
445 coding genes) in PacBio and 220 genes (58.7%) in ONT had no SNP errors. We further restricted our
446 analysis to genes without SNPs, Indels, or switch errors. In this context, we observed that 261 genes
447 (58.1%) in CDS and 251 genes (55.9%) in exons for PacBio, as well as 227 genes (60.0%) in CDS and 214
448 genes (57.1%) in exons for ONT, met this criterion. These results indicate a promising variant calling
449 accuracy in the protein-coding regions.
450
451 **Figure 7c** presents the number of wholly phased genes across all benchmarking datasets. The MAS-
452 Seq HG004 identified 1,085 wholly phased genes, exceeding 386 and 304 genes reported by the Iso-
453 Seq HG004 and HG005, respectively. Similarly, ONT dRNA004 identified 504 and 479 wholly phased
454 genes in HG004 and HG005, respectively, compared to 211 and 317 in the cDNA datasets and 58 and
455 52 in the dRNA002 datasets. **Figure 7d** illustrates the number of genes free from SNPs, Indels, or switch
456 errors, with MAS-Seq leading at 1,165 genes, followed by dRNA004 with 655 in HG004 and 653 in
457 HG005. The substantial numbers of wholly phased and error-free genes highlight the potential for
458 reliable downstream gene analysis, particularly using the PacBio MAS-Seq and ONT dRNA004
459 technologies.
460
461 **Benchmarking runtime and memory usage**
462 As shown in **Figure 3e** and **Supplementary Table 9**, we evaluated the runtime and memory usage of
463 the Clair3-RNA algorithm and other variant callers. Clair3-RNA demonstrated an outperforming
464 runtime against other variant callers. Specifically, using 16 CPU cores, Clair3-RNA required ~18 minutes
465 and ~50 minutes, on HG004 PacBio Iso-Seq and MAS-Seq, and ~34 minutes and ~29 minutes on HG004
466 ONT cDNA and dRNA004 dataset, respectively. This is generally faster than Clair3, LongcallR, and
467 DeepVariant, in which the average runtime ranges from 42 to 80 minutes. Clair3-RNA also showcased
468 uniform memory consumption across all datasets, with a footprint of ~7 GB. In contrast, DeepVariant
469 required about 26 GB and 24 GB of memory for the PacBio and ONT, respectively. Clair3 displayed a
470 slightly higher memory footprint, on average 49 GB, due to its variant phasing process. In comparison,
471 LongcallR recorded the highest memory consumption at PacBio MAS-Seq, with 83 GB, and averaged
472 47 GB.
473

Discussion

Variant calling in RNA-seq datasets offers more comprehensive insights into the transcriptome analysis. However, RNA-seq comes with higher error rates, making it challenging to differentiate between sequencing artifacts and true variants. RNA reads also display imbalanced coverage distribution across the reference genome, making it difficult to detect low-frequency variants. Additionally, A-to-I RNA editing events present a significant challenge in RNA-seq analysis, often resulting in higher false positives.

In this study, we introduced Clair3-RNA, the first deep learning-based variant caller designed for long-read RNA sequencing. Clair3-RNA utilizes a pileup neural network and incorporates various techniques, such as coverage normalization, variant zygosity switching, and editing site tagging, for accurate variant detection in lrRNA-seq. Across different platforms, including PacBio Iso-Seq, PacBio MAS-Seq, and ONT dRNA004, Clair3-RNA successfully identified ~90% of true RNA variants with a minimum four reads support, and further improved to 92% when the minimum read coverage was increased to 10 and zygosity information was disregarded, showing its capability for accurate variant detection in lrRNA-seq. By identifying known editing sites and analyzing A-to-G and T-to-C SNP false positives, we also found that Clair3-RNA was capable of distinguishing major A-to-I editing sites from genuine variants, which is challenging for other DNA variant callers. We further investigated the incorporation of phasing into the Clair3-RNA architecture and achieved a ~2% F1-score improvement in our benchmarking, demonstrating the advantages of haplotype phasing in lrRNA-seq.

Clair3-RNA demonstrates greater SNP accuracy than other variant callers across all platforms and sequencing scenarios. Nevertheless, the accuracy of indel calling remains challenging due to allele coverage fluctuations and increased error rates. Enhancing Indel accuracy could involve incorporating techniques such as local assembly for Indel realignment and local haplotype assembly to differentiate low-frequency Indels from sequencing artifacts, leading to more precise Indel detection. Moreover, although Clair3-RNA successfully identified more than 63% and 93% of the editing sites in the ONT dRNA004 and PacBio Iso-Seq, respectively. As the false positive distribution shown in **Supplementary Figure 3**, the presence of RNA false positives was linked primarily to excessively low allelic fraction RNA editing sites or high allelic fraction sequencing artifacts that the neural network failed to capture. These editing sites exhibited subtle differences against true variants at the base level, making them indistinguishable from genuine A-to-G or T-to-C variants without additional contextual information. Further attempts will involve integrating raw sequencing signal features for editing site detection, especially in dRNA sequencing, as the A-to-I RNA editing event is more distinct at the raw signal level.

508 Dorado also supports outputting modified bases in its latest release [24], which can be used to detect
509 editing sites. Furthermore, in our model architecture design, we utilized a recurrent neural network
510 to encode the pileup features. As more RNA data becomes accessible, a meticulously crafted
511 transformer-based architecture might further improve the performance. The attention module within
512 a transformer-based network exhibits the capacity to capture the connections in a distant flanking
513 window, thereby enhancing accuracy in challenging regions.
514
515 **Methods**
516 **Overview of Clair3-RNA**
517 **Figure 1** shows an overview of the Clair3-RNA workflow. Starting from the read alignments in
518 BAM/CRAM format of an input sample, Clair3-RNA identifies potential variants with adequate read
519 support and base alteration. Leveraging the pileup methodology inherited from Clair3, the candidates
520 and their corresponding flanking window sites are aggregated to create pileup tensors, which are then
521 processed through a Bi-LSTM-based pileup neural network for inference. The neural network produces
522 probabilistic multi-task output that differentiates variants from editing sites and sequencing artifacts.
523 Identified variants are further refined by the editing tagging module to tag known editing sites. Clair3-
524 RNA is compatible with various platforms, including PacBio Sequel with the Iso-Seq kit, PacBio Revio
525 with the MAS-Seq kit, ONT cDNA, ONT dRNA002, and ONT dRNA004.
526
527 **Clair3-RNA's input and output**
528 **Input**. Clair3-RNA takes the input, which includes two components: 1) a reference genome of the
529 sequencing sample, and 2) a sequencing alignment with the same reference material as 1). Clair3-RNA
530 automates the whole variant calling process to get the **VCF output**, as illustrated below. The pileup
531 tensor consists of 594 integers, covering 33 genome positions, which include the candidate position
532 and 16 flanking positions on both the left and right sides, with 18 features at each position. These 18
533 pileup features are designated as "A+", "C+", "G+", "T+", "I_S+", "I_{1S}+", "D_S+", "D_{1S}+", "D_R+", "A-", "C-",
534 "G-", "T-", "I_S-", "I_{1S}-", "D_S-", "D_{1S}-", and "D_R-". A, C, G, T, I, and D represent the counts of read support
535 for the four nucleotides, insertion and deletion, while the symbols "+" and "-" represent the read from
536 the forward and reverse strands, respectively. The superscripts "S" and "R" represent the start and
537 remaining positions of Indel. A detailed description of the 18 pileup features is provided in the
538 **Supplementary Notes** – **Description of RNA pileup input features** section. All the features above
539 worked with sites with the reads covered, and we assigned a background intensity of 0 for all padding
540 sites. This technique of zero-padding is also utilized in RNA sequencing to account for surrounding RNA
541 splicing regions.

542

543 **Clair3-RNA model architecture**. The Clair3-RNA pileup network incorporates two bidirectional long
544 short-term memory (Bi-LSTM) layers and three fully connected layers to encode sequential pileup
545 features. The output of the pileup model encompasses two probabilistic tasks: 21-genotype and
546 zygosity. Further details about the output tasks are provided in the **Supplementary Notes** –
547 **Description of pileup network outputs** section. The identified variants, encompassing both nucleotide
548 alterations and genotypes, were determined by evaluating the combined probabilities from two
549 output tasks. In the case of an Indel variant, the network specifies the variant type, and the length of
550 the Indel is inferred from the supporting read evidence.

551

552 **VCF output**. Three categories of variants are shown in Clair3-RNA's VCF output: 1) variants labelled as
553 "PASS" that passed quality score threshold, 2) labelled as "LowQual" that has low-quality scores (i.e.,
554 QUAL < 2 for PacBio and QUAL < 8 for ONT, configurable), and 3) tagged as "RNAEditing" corresponded
555 to known editing sites identified by the REDIportal database. Specifically, variants were tagged as
556 "RNAEditing" if they match both the reference and alternate alleles given in REDIportal. For each
557 variant entry, the reference, alternate variant, genotype, coverage of the reference allele, and
558 coverage of all alternate alleles are shown.

559

560 **Training techniques tailored for RNA data**

561 **Training candidate generation**. In DNA model training, the ground truths were established using the
562 golden GIAB truth variants. However, in RNA model training, owing to limited sequencing coverage in
563 specific regions, not all truth variants are suitable for training. Initially, we scanned the training dataset
564 to acquire the read coverage (DP) and the alternate allele (AD) supporting truth variants, focusing on
565 evident variants with adequate read support. For the artifacts, we scanned the alignment site-by-site
566 to identify candidates with DP⩾4, AD⩾2, and exceeding a minor allelic fraction (typically 0.08, which
567 was manually decided). Additionally, we accounted for zygosity switches and editing sites in the
568 training data, as elaborated in the **Methods** - **Variant zygosity switching** section.

569

570 **Variant zygosity switching.** Zygosity switching in RNA data can lead to misinterpretations of genetic
571 variants. This occurs when a heterozygous variant appears as a homozygous reference or a
572 homozygous variant due to allelic imbalance. Zygosity switching in RNA data might be due to
573 differential gene expression when one allele dominates in specific regions. Additionally, RNA splicing
574 might create isoforms, altering the variant zygosity across different isoforms [5]. Ignoring zygosity
575 switching in RNA data can have a significant impact on model training. Therefore, before training the

576 model, we computed the alternate allelic fraction of each candidate and adjusted the zygosity
577 obtained from GIAB based on DNA sequencing samples. By defining each observed candidate allelic
578 fraction as VAF_O and the truth zygosity as GT_{Truth}, where $GT \in \{0/1, 1/1, 1/2\}$, we can determine the
579 inferred zygosity $GT_{Inferred}$, as follows:

580
$$GT_{Inferred} \begin{cases} 0/1 & if\ VAF > VAF_{HET} - \epsilon\ and\ VAF < VAF_{HET} + \epsilon\ and\ GT_{Truth} \neq 0/1 \\ 1/1 & if\ VAF > VAF_{HOM} - \epsilon\ and\ GT_{Truth} \neq 1/1 \\ 0/0 & if\ VAF < VAF_{Artifact} \\ else & GT_{Truth} \end{cases}$$

581

582 where GT_{Truth} is the observed ground truth genotype in GIAB truths. VAF_{HET}, VAF_{HOM}, and $VAF_{Artifact}$
583 represent the expected allelic fraction for heterozygous variants, homozygous variants, and
584 sequencing artifacts were set as 0.5, 1.0, and 0.08, respectively. ϵ represents the tolerance epsilon
585 parameter and was manually set as 0.2. For multi-allelic 1/2 genotypes, we considered each allele and
586 inferred the genotype when any of the alternate alleles was missing. Finally, we used $GT_{Inferred}$ instead
587 of GT_{Truth} for constructing zygosity and 21-genotype labels during model training.

588

589 **Coverage normalization for extremely high-coverage data**

590 RNA data exhibits an uneven distribution of sequencing read coverage, stemming primarily from
591 variations in transcript abundance, non-uniform RNA expression, and RNA splicing events. These
592 factors result in uneven coverage compared to homogeneous DNA sequencing data. The presence of
593 high-coverage sites poses challenges for model inference, as the training data may not adequately
594 represent the full spectrum of coverage distribution. Although Clair3-RNA has implemented
595 techniques like Focal Loss [32] to address data imbalances during model training, excessively high
596 coverage levels can significantly impede the accuracy of model inference.

597

598 One potential strategy involves subsampling reads during inference to mitigate the gap; this approach,
599 however, may impact performance due to less alignment included in inferences. Building upon this,
600 we proposed a coverage normalization strategy. Initially, we computed the coverage distribution of
601 the training sample and excluded the top 10% of high-coverage sites, identifying the maximum
602 coverage confidence interval as C_T. In the inference phase, we determined the coverage of each
603 candidate tensor C_I by computing the reads supporting the nucleotide A, C, G, T, and deletions.
604 Subsequently, for each candidate coverage C_I exceeding C_T, we computed the normalized pileup
605 features F_{pileup_norm} using the formula:

606
$$F_{pileup_norm} = F_{pileup} * \frac{C_T}{C_I}$$

where F_{pileup} is the total 594 pileup integers used for prediction. In our experiment, using the normalized feature generally improved recall by ~1%, effectively mitigating the influence of extremely high coverage on the model's inference.

Orthogonal high-quality RNA editing benchmark set generation

RNA A-to-I editing is a post-transcriptional modification, where adenosine (A) residues are deaminated to inosine (I) by adenosine deaminases acting on RNA (ADAR) enzymes. The editing process can alter the RNA structure and function, impacting gene expression and protein diversity [16]. In sequence alignments, A-to-I during editing is interpreted as a G by reverse transcription. RNA A-to-I editing is often represented as an A-to-G change on the edited RNA strand and as a T-to-C change on the corresponding DNA template strand. dRNA sequencing detects A-to-I editing directly from RNA data, as it omits the reverse transcription step, offering a precise view of RNA modifications without potential artifacts.

While no definitive editing site reference is available from GIAB, we generated a high-quality benchmark set of editing sites by cross-validating potential editing sites with DNA sequencing data, as shown in **Figure 5a**. The rationale is that an editing site is determined if present in RNA data but absent in DNA data. Initially, we identified A-to-G and T-to-C candidates from RNA alignments and computed read support in the corresponding DNA sequencing data. If the sequencing coverage (DP) and variant allelic fraction (VAF) supporting either editing allele (G or C) in RNA (R) and DNA (D) is calculated as DP_R, VAF_R, DP_D, and VAF_D, respectively, then the editing site is decided as follows:

$$isEditing = \begin{cases} True & if\ DP_R \geq MDP\ and\ VAF_R \geq MAF_R\ and\ DP_D \geq MDP\ and\ VAF_D < MAF_D \\ False & else \end{cases}$$

where *MDP* represents minimal coverage depth, and VAF_D and VAF_R represent the VAF observed in DNA and RNA, respectively. MAF_D represents maximal allelic fraction supporting A-to-C or G-to-C in DNA, and MAF_R represents minimal allelic fraction supporting A-to-C or G-to-C in RNA. We set *MDP*, MAF_D, and MAF_R as 8, 0.08, and 0.75, respectively, for both the PacBio and ONT experiments. **Supplementary Figure 1** presents the detected allele frequency distribution of high-quality RNA editing sites across PacBio and ONT datasets.

We incorporated candidates from the GIAB high-confidence regions for model training. These candidates were categorized in RNA BAM files as RNA variants, RNA editing sites, or artifacts. Throughout our experiments, we explored the inclusion of a separate RNA editing site task to identify

641 RNA edits. Additionally, we tried incorporating the editing sites as an additional category within the
642 zygosity task. We found that both approaches failed to accurately identify true RNA variants compared
643 to directly integrating editing sites into the homozygous reference category. This indicates that the
644 model is able to classify the editing sites as references in model training.
645
646 **RNA editing site tagging using the REDIportal database**
647 REDIportal is a comprehensive database that consolidates a vast array of RNA editing sites sourced
648 from three main repositories: ATLAS, DARNED, and RADAR. Within REDIportal v2, approximately 40
649 million RNA editing sites are documented. In Clair3-RNA, variants are tagged as RNA editing sites if
650 they correspond to A-to-G or T-to-C alterations found in the REDIportal database, ensuring a matching
651 alternate representation. REDIportal prioritizes RNA editing sites validated by agreement of at least
652 two of its constituent sources (ATLAS, DARNED, RADAR), with adjustable configuration setting for
653 users' customs. Clair3-RNA supports REDIportal versions customized for both the GRCh37 and GRCh38
654 reference genomes, providing a reliable resource for exploring established RNA editing events.
655
656 **Incorporating phasing information into neural network**
657 Haplotype assembly, or phasing, is a fundamental technique for reconstructing an individual's
658 haplotypes [30, 33]. Long-read sequencing generates reads spanning tens of kilobases or more,
659 enabling the capture of haplotype information often lost in short-read sequencing data. This method
660 has been shown to significantly enhance variant calling performance in long-read sequencing data. In
661 comparison to lrDNA-seq, which typically has read lengths of 10-100k for phasing, lrRNA-seq can still
662 achieve read lengths of 1-3k, allowing for phasing in specific regions. However, to date, no methods
663 have investigated the impact of phasing on the performance of lrRNA-seq. In this study, we attempted
664 to incorporate additional phasing information into Clair3-RNA architecture to gain a deeper
665 understanding of its contribution to lrRNA-seq performance.
666
667 **Variant phasing and read haplotagging**. As illustrated in **Figure 6a**, during the training phase, we
668 selected only the heterozygous germline SNP variants identified in both Clair3-RNA and the GIAB
669 truths for variant phasing. This approach was chosen to mitigate the potential introduction of false
670 positives, which could adversely affect phasing accuracy and lead to cumulative phasing errors during
671 model training. Incorporating only GIAB truths might introduce variants that have insufficient read
672 coverage and discrepancies in zygosity. Alignments are then haplotagged using the phased variants
673 using the WhatsHap “haplotag” submodule, which assigns each phased read the "HP:x" tag, where x
674 indicates the haplotype assigned.

During model training, read subsampling is adopted. We first applied variant phasing and read haplotagging to the BAM file with a full-coverage spectrum. The phased full-coverage reads were then subsampled into varying coverage levels for data augmentation. For candidates with phased reads, we randomly switched the phased reads to unphased reads by discarding the “HP” tag. This switching process simulates phasing scenarios for low-coverage BAM files, where the read haplotagging percentage is lower, while avoiding introducing incorrect phasing that could occur when applying phasing to each subsampled coverage. All subsampled candidates and switched candidates were generated with additional phasing features for model training. In the inference phase, the heterozygous variants were acquired from Clair3-RNA’s first round called variants. We enabled the "--distrust-genotypes" option in WhatsHap phasing to allow for switching variant zygosities in the event of incorrectly called zygosities.

Pileup phasing feature. We incorporated 12 pileup phasing features for each candidate position, represented as X_{H1} and X_{H2}, where $X \in \{A, C, G, T, I, D\}$. The phasing features denote the total counts of nucleotides A, C, G, and T, as well as insertions (I) and deletions (D) for each candidate site in the phased paternal and maternal haplotypes, derived from the "HP" tag of the phased reads. The 12 additional pileup phasing features were concatenated with the original 18 pileup features, resulting in a total of 30 features that were input into the neural network for model training and inference.

Analysis of overlap and error statistics for GENCODE annotated gene

The gene list was obtained from the GENCODE database (version 46), focusing specifically on the "gene" category for analysis. We categorized the genes based on their overlap with callable high-confidence regions, defined as the intersection of the callable region with read support and the GIAB high-confidence regions. Giving each gene coordinate and all its overlapping high-confidence regions (*hc*), we calculate the intersection size using the formula:

$$L_O(g) = \sum_{i=1}^{n} \max\left(0, \min\left(g_{end}, hc_{end}^{i}\right) - \max\left(g_{start}, hc_{start}^{i}\right)\right)$$

Where *n* represents the number of overlapping regions. *g* and *hc* refer to gene and callable high-confidence region, respectively. The overlap ratio score $H(g)$ for each gene is then computed as follows:

$$H(g) = \frac{L_O(g)}{L(g)}$$

Where $L(g)$ denotes the gene size, a gene is categorized as "wholly covered" if $H(g)$ equals 1. Conversely, it is classified as "partially covered" if the total overlap length $H(g)$ is greater than or equal to α, where α is a threshold percentage for partial overlapping and is manually set to 0.75. If $H(g)$ equals 0, the gene is classified as "not covered".

A gene is considered to have no SNP or Indel errors if true positives are present within the specified gene coordinates, with no false positives or false negatives in the same region, based on benchmarking results obtained from hap.py. A gene is considered to have no switch errors if all heterozygous variants in the given phasesets are phased correctly. Switch errors [30] were assessed using the WhatsHap "compare" submodule within the given overlapping phasesets of lrDNA-seq and lrRNA-seq. The editing distance of the phased heterozygous variants in these two groups was computed as switch errors. The command used for this analysis is given in the **Supplementary Notes** – **Command line used** section.

Data preparation

PacBio mapping with pbmm2 and minimap2. PacBio Iso-Seq and MAS-Seq reads were aligned with pbmm2 or minimap2 with the configurations provided in the **Supplementary Notes**. ONT reads were aligned with minimap2 splice mode. As discussed in the **Results** – **Performance on PacBio** section, we found miniamp2 alignment output performed in RNA data alignment, even though pbmm2 was optimized for PacBio data.

ONT dRNA004 base-calling and mapping. The ONT dRNA002 and cDNA reads were basecalled using Guppy (version 6.4.6), and the ONT dRNA004 reads were basecalled using Dorado (version 3.2.4). Basecalled reads with low read quality scores ("qs" < 10) were labelled as "FAILED" and discarded, as recommended by the ONT protocol. Subsequently, the basecalled reads were mapped to the GRCh38 human reference genome using minimap2 for evaluation.

ONT SQK-RNA004 kit library preparation and sequencing. The RNA sequencing libraries were constructed following the ONT's SQK-RNA004 protocol. RNA sourced from GIAB HG002, HG004, and HG005 samples was obtained from the Coriell Institute. The total RNA underwent assessment for concentration, purity, and integrity using Nanodrop and Qubit analyses. Subsequently, the RNA fragments were ligated with a sequencing adapter utilizing ONT's latest SQK-RNA004 direct RNA

740 sequencing kit. Finally, the RNAs were sequenced on the PromethION flow cell using our in-house
741 PromethION 2 Solo sequencing device, running MinKNOW software version 1.18.02 for 96 hours.
742
743 **Model training and evaluation.** Model training was conducted using the known variants in the GIAB
744 HG002 dataset, encompassing the full spectrum of read coverage, along with four subsampled
745 coverage levels. For all pre-trained models described in the manuscript, we consistently used variants
746 from chromosomes 1 to 19, 21, and 22 for training, with a random selection of 90% of the data for
747 training and the remaining 10% for evaluation, while always holding out chromosome 20 in training.
748 Performance evaluation was conducted in callable regions defined in the **Methods** - **Benchmarking**
749 **metrics** section. Clair3-RNA trains the model for either 30 epochs or until the validation loss does not
750 decrease for five consecutive epochs, employing an exponentially decaying learning rate until
751 convergence. The epoch that performs best on the validation set will be selected for benchmarking
752 and as the production model. The training process utilizes the Rectified Adam optimizer [34] and
753 includes an initial weight warm-up. An optimized cross-entropy loss function, specifically Focal Loss
754 [32], is used during model training to enhance performance.
755
756 **Benchmarking metrics**
757 RNA sequences represent transcribed segments of the genome, encompassing mainly the exon
758 regions. Notably, the number of total variants in RNA was considerably lower than that in DNA, where
759 there were ~27,000 to ~318,000 variants, depending on the sequencing platform, RNA throughput,
760 and specific gene expression. In contrast, ~4 million true variants were available for DNA data. **Figure**
761 **3a** shows the benchmarking workflow for RNA variant evaluation. We utilized Clair3-RNA's
762 "get_rna_bed" submodule to identify callable regions with minimum read support (default set at 4,
763 adjustable through options). These callable regions were then cross-referenced with the GIAB high-
764 confidence regions to establish reliable benchmarking regions. Within the benchmark VCF from GIAB,
765 variants with less than two reads supporting the alternate allele were excluded. We used hap.py [35]
766 to compare the variant-called variants against the truth variants in callable high-confidence regions.
767 We used Clair3-RNA's "calculate_overall_metrics" submodule to generate three metrics, "Precision",
768 "Recall", and "F1-score", for five categories: "Overall", "SNP", "Indel", "Insertion", and "Deletion".
769
770 **Computational performance**
771 Clair3-RNA was implemented with Python and TensorFlow and leveraged PyPy for speedup. RNA
772 variant calling using Clair3-RNA requires only CPU resources. Using one replicate sequencing data,
773 Clair3-RNA takes ~30 minutes for ONT dRNA004 or PacBio MAS-Seq, using two 12-core Intel Xeon

774 Silver 4116 processors. The memory consumption was capped at 500 MB per process. Model training
775 requires a high-end graphics processing unit (GPU). We used TensorFlow to train models on an Nvidia
776 GeForce RTX™ 3090 Ti and 4090 Ti GPU. We used the GNU time command to benchmark the runtime
777 and memory usage.
778
779 **Code availability**
780 Clair3-RNA is open source and available at https://github.com/HKU-BAL/Clair3-RNA under the BSD 3-
781 Clause license. A public Docker image hkubal/clair3-rna:latest that can evaluate Clair3-RNA's
782 evaluation pipeline is also available for simplicity. Users can run the workflow seamlessly in one
783 command with the input RNA BAM file and reference file. Clair3-RNA provides user interface flexibility
784 that supports parallel computing, user-customized parameter setting, and workflow refinement.
785
786 **Data availability**
787 The links to the reference genomes, truth variants, benchmarking materials, and ONT & PacBio dataset,
788 and the benchmarked commands and parameters used in this study are available in the
789 **Supplementary Notes**. All analysis output, including the VCFs and running logs, are available at
790 http://www.bio8.cs.hku.hk/clair3_rna/analysis_result. The ONT dRNA004 GIAB HG002, HG004, and
791 HG005 sequencing data using the dRNA SQK-RNA004 kit generated in this study was deposited in the
792 NCBI archive with accession ID PRJNA1166426.
793
794 **Acknowledgements**
795 R.L. was supported by Hong Kong Research Grants Council grants GRF (17113721) and TRS (T21-
796 705/20-N), the Shenzhen Municipal Government General Program (JCYJ20210324134405015), the
797 URC fund at HKU and Oxford Nanopore Technologies.
798
799 **Author contributions**
800 R.L. and F.S. conceived the study. Z.Z., X.Y., and R.L. designed the algorithms, implemented Clair3-RNA,
801 and wrote the paper. F.S. and J.M. supported datasets for benchmarking and evaluation. Y.L., C.L., X.C.,
802 R.K.K., F.S., and J.M. evaluated the benchmarking results. All authors revised the manuscript.
803
804 **Competing interests**
805 R.L. receives research funding from ONT. F.S. receives research support from Illumina, PacBio and ONT.
806 J.M. has received reimbursement for travel, accommodation, and conference fees to speak at events
807 organized by ONT. The remaining authors declare no competing interests.

References

1. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC bioinformatics* 2011, **12**:1-16.
2. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nature methods* 2010, **7**(12):1009-1015.
3. Cook DE, Venkat A, Yelizarov D, Pouliot Y, Chang P-C, Carroll A, De La Vega FM: **A deep-learning-based RNA-seq germline variant caller**. *Bioinformatics Advances* 2023, **3**(1):vbad062.
4. Guo Y, Zhao S, Sheng Q, Samuels DC, Shyr Y: **The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data**. *Bmc Genomics* 2017, **18**:37-46.
5. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data**. *The American Journal of Human Genetics* 2013, **93**(4):641-651.
6. Xie S, Leung AW-S, Zheng Z, Zhang D, Xiao C, Luo R, Luo M, Zhang S: **Applications and potentials of nanopore sequencing in the (epi) genome and (epi) transcriptome era**. *The Innovation* 2021, **2**(4).
7. Roundtree IA, Evans ME, Pan T, He C: **Dynamic RNA modifications in gene expression regulation**. *Cell* 2017, **169**(7):1187-1200.
8. Jain M, Abu-Shumays R, Olsen HE, Akeson M: **Advances in nanopore direct RNA sequencing**. *Nature methods* 2022, **19**(10):1160-1164.
9. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G: **Systematic assessment of long-read RNA-seq methods for transcript identification and quantification**. *Nature methods* 2024:1-15.
10. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome**. *Nature biotechnology* 2019, **37**(10):1155-1162.
11. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A: **Highly parallel direct RNA sequencing on an array of nanopores**. *Nature methods* 2018, **15**(3):201-206.
12. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C: **Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells**. *Nature communications* 2017, **8**(1):16027.
13. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R: **Symphonizing pileup and full-alignment for deep learning-based long-read variant calling**. *Nature Computational Science* 2022, **2**(12):797-803.
14. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT: **A universal SNP and small-indel variant caller using deep neural networks**. *Nature biotechnology* 2018, **36**(10):983-987.
15. Zheng Z, Su J, Chen L, Lee Y-L, Lam T-W, Luo R: **ClairS: a deep-learning method for long-read somatic small variant calling**. *bioRxiv* 2023:2023.2008. 2017.553778.
16. Nishikura K: **Functions and regulation of RNA editing by ADAR deaminases**. *Annual review of biochemistry* 2010, **79**(1):321-349.
17. **LongcallR** [https://github.com/huangnengCSU/longcallR]
18. Luo R, Sedlazeck FJ, Lam T-W, Schatz MC: **A multi-task convolutional deep neural network for variant calling in single molecule sequencing**. *Nature communications* 2019, **10**(1):998.
19. Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W: **Exploring the limit of using a deep neural network on pileup data for germline variant calling**. *Nature Machine Intelligence* 2020, **2**(4):220-227.

857 20. Su J, Zheng Z, Ahmed SS, Lam T-W, Luo R: **Clair3-trio: high-performance Nanopore long-read**
858 **variant calling in family trios with trio-to-trio deep neural networks**. *Briefings in*
859 *Bioinformatics* 2022, **23**(5):bbac301.
860 21. Yu H, Zheng Z, Su J, Lam T-W, Luo R: **Boosting variant-calling performance with multi-**
861 **platform sequencing data using Clair3-MP**. *BMC bioinformatics* 2023, **24**(1):308.
862 22. **Guppy basecalling software** [https://community.nanoporetech.com/guppy-software-
863 overview]
864 23. **Nanopore SQK-RNA004 kit** [https://nanoporetech.com/document/direct-rna-sequencing-
865 sqk-rna004]
866 24. **Dorado basecaller** [https://github.com/nanoporetech/dorado]
867 25. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B,
868 Miller N: **Benchmarking challenging small variants with linked and long reads**. *Cell genomics*
869 2022, **2**(5).
870 26. **Nanopore EPI2ME Labs** [https://labs.epi2me.io/giab-2023.05]
871 27. Baid G, Nattestad M, Kolesnikov A, Goel S, Yang H, Chang P-C, Carroll A: **An extensive**
872 **sequence dataset of gold-standard samples for benchmarking and development**. *bioRxiv*
873 2020:2020.2012. 2011.422022.
874 28. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP: **Maftools: efficient and comprehensive**
875 **analysis of somatic variants in cancer**. *Genome research* 2018, **28**(11):1747-1756.
876 29. Nishikura K: **A-to-I editing of coding and non-coding RNAs by ADARs**. *Nature reviews*
877 *Molecular cell biology* 2016, **17**(2):83-96.
878 30. Patterson M, Marschall T, Pisanti N, Van Iersel L, Stougie L, Klau GW, Schönhuth A: **WhatsHap:**
879 **weighted haplotype assembly for future-generation sequencing reads**. *Journal of*
880 *Computational Biology* 2015, **22**(6):498-509.
881 31. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C,
882 Wright J, Armstrong J: **GENCODE reference annotation for the human and mouse genomes**.
883 *Nucleic acids research* 2019, **47**(D1):D766-D773.
884 32. Lin T: **Focal Loss for Dense Object Detection**. *arXiv preprint arXiv:170802002* 2017.
885 33. Lin J-H, Chen L-C, Yu S-C, Huang Y-T: **LongPhase: an ultra-fast chromosome-scale phasing**
886 **algorithm for small and large variants**. *Bioinformatics* 2022, **38**(7):1816-1822.
887 34. Liu L, Jiang H, He P, Chen W, Liu X, Gao J, Han J: **On the variance of the adaptive learning rate**
888 **and beyond**. *arXiv preprint arXiv:190803265* 2019.
889 35. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle
890 MA, Tezak Z, Lababidi S: **Best practices for benchmarking germline small-variant calls in**
891 **human genomes**. *Nature biotechnology* 2019, **37**(5):555-560.
892

Figures

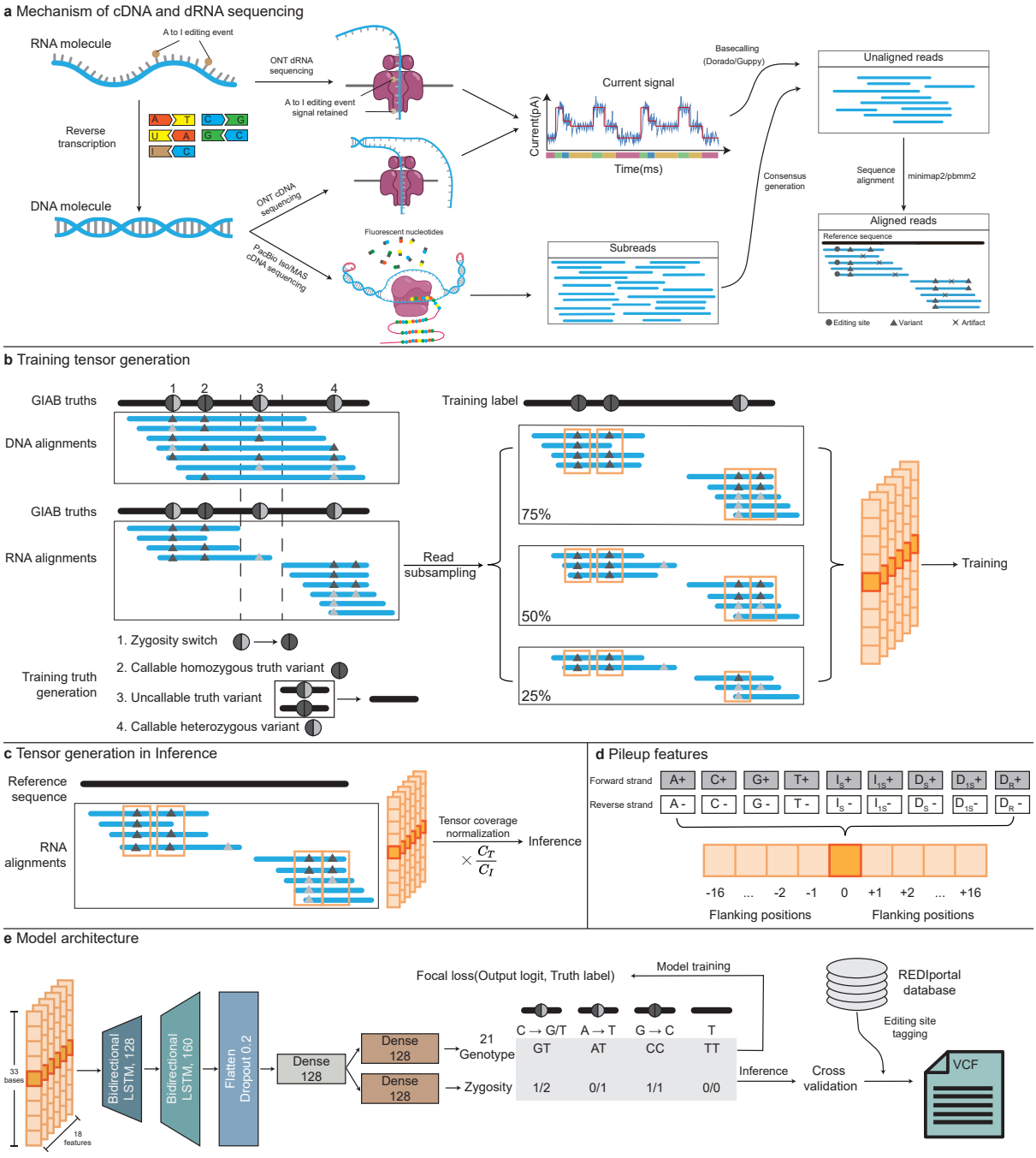
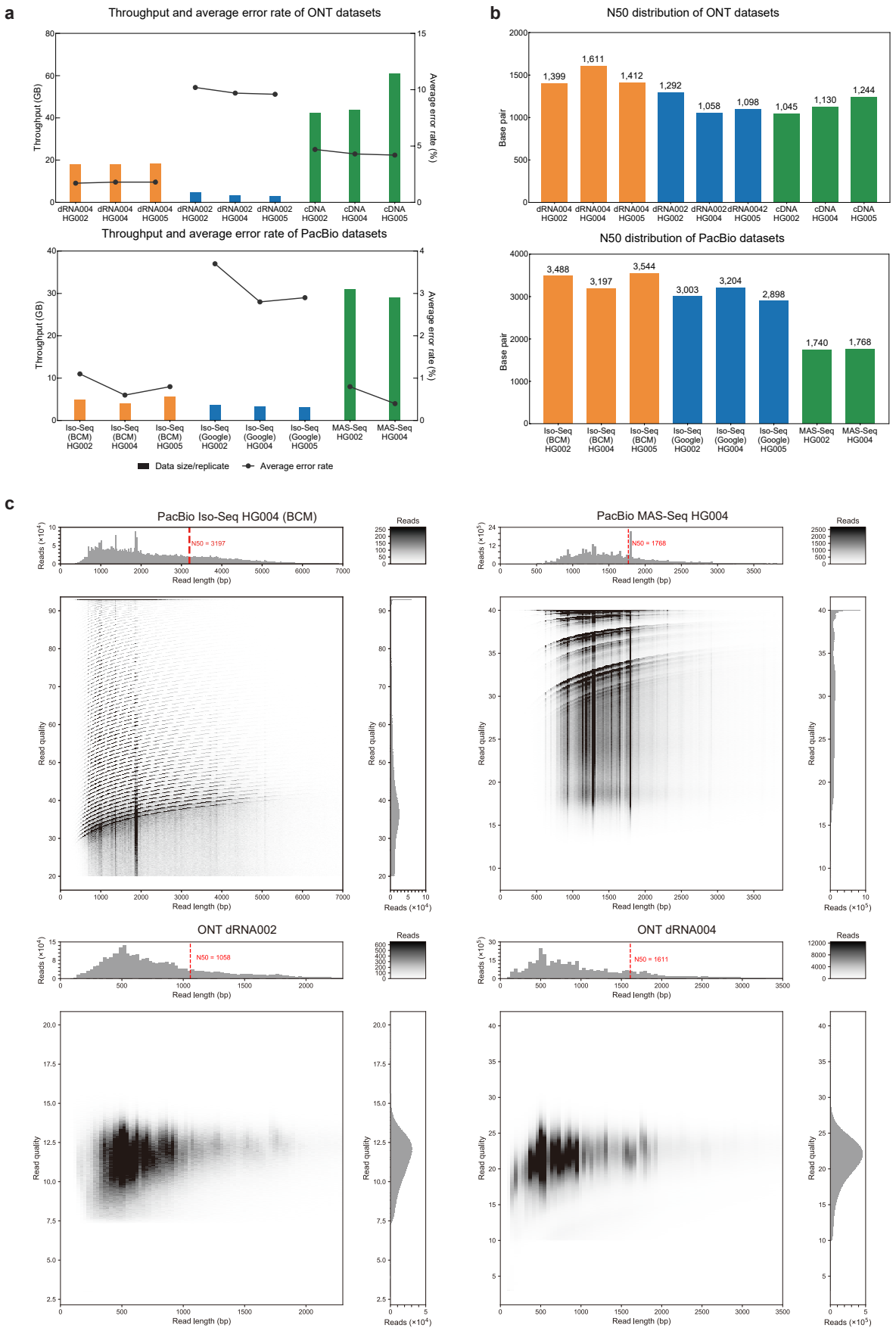


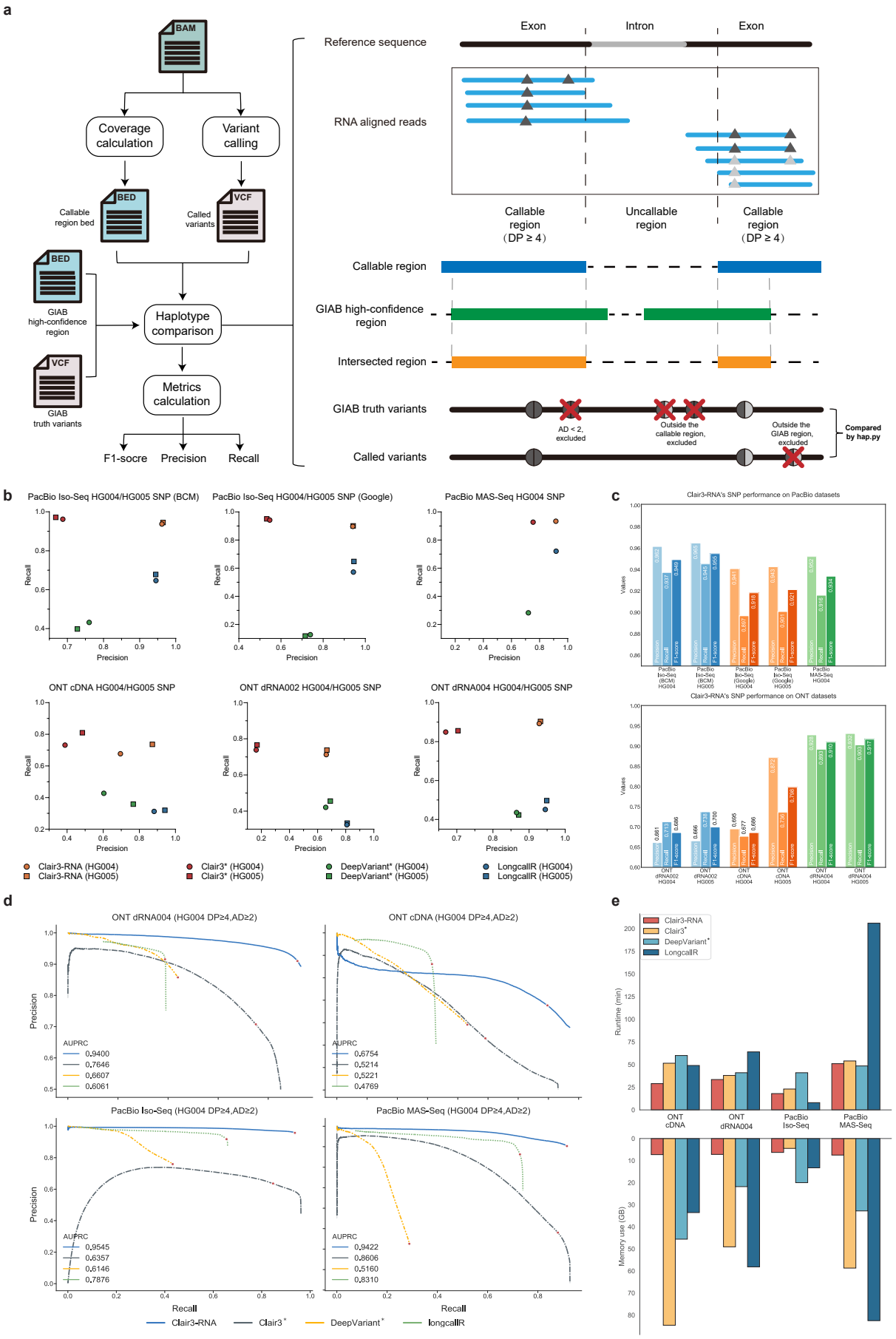
Figure 1. Overview of Clair3-RNA variant-calling workflow.
(a) The figure contrasts RNA sequencing techniques, specifically highlighting the mechanisms of direct RNA sequencing (dRNA) and complementary DNA (cDNA) sequencing with reverse transcription. In dRNA sequencing, Inosine editing is sequenced as is, while in cDNA sequencing, it is converted to Guanine nucleotide. (b) The illustration of training tensor generation outlines the process of deriving training tensors from DNA and RNA alignments, incorporating labels obtained from GIAB truths. Variants located outside the callable region are omitted, and the zygosity switch is applied, whereby heterozygous variants displaying allelic fractions indicative of homozygous variant or homozygous reference in RNA are reclassified. Read subsampling is employed for data augmentation, and pileup tensors are produced for every candidate site, along with flanking positions. (c) The illustration of tensor generation in inference shows the process of creating tensors in inference. Coverage normalization is adopted for exceeding coverage in inference. (d) The figure illustrates 18 pileup

features for each position in the forward and reverse strands. More details are provided in the **Supplementary methods – Description of RNA pileup input features** section. (e) In this illustration of the Clair3-RNA model architecture, the pileup network includes a two-layer LSTMs with two dense layers. The output supports two probabilistic tasks for classification: 21-genotype and zygosity. Variants identified by the model are tagged by the REDIportal database and recorded in VCF files.


a
Throughput and average error rate of ONT datasets
Throughput (GB)
Average error rate (%)
dRNA004 HG002
dRNA004 HG004
dRNA004 HG005
dRNA002 HG002
dRNA002 HG004
dRNA002 HG005
cDNA HG002
cDNA HG004
cDNA HG005
Throughput and average error rate of PacBio datasets
Iso-Seq (BCM) HG002
Iso-Seq (BCM) HG004
Iso-Seq (BCM) HG005
Iso-Seq (Google) HG002
Iso-Seq (Google) HG004
Iso-Seq (Google) HG005
MAS-Seq HG002
MAS-Seq HG004
Data size/replicate
Average error rate
b
N50 distribution of ONT datasets
Base pair
1,399
1,611
1,412
1,292
1,058
1,098
1,045
1,130
1,244
N50 distribution of PacBio datasets
3,488
3,197
3,544
3,003
3,204
2,898
1,740
1,768
c
PacBio Iso-Seq HG004 (BCM)
N50 = 3197
PacBio MAS-Seq HG004
N50 = 1768
ONT dRNA002
N50 = 1058
ONT dRNA004
N50 = 1611
Reads
Read length (bp)
Read quality

913 **Figure 2. Dataset distribution of different platforms.**

914 (a) The average sequencing throughput and sequencing error rate for all sequencing replicated across
915 different datasets are illustrated. (b) N50 distributions are compared among different datasets. (c) The
916 read length and read quality distributions of the HG004 sample are depicted across four platforms:
917 PacBio Iso-Seq, PacBio MAS-Seq, ONT dRNA002, and ONT dRNA004.f


a
BAM
Coverage calculation
Variant calling
BED
Callable region bed
VCF
Called variants
BED
GIAB high-confidence region
VCF
GIAB truth variants
Haplotype comparison
Metrics calculation
F1-socre
Precision
Recall
Reference sequence
Exon
Intron
Exon
RNA aligned reads
Callable region (DP ≥ 4)
Uncallable region
Callable region (DP ≥ 4)
Callable region
GIAB high-confidence region
Intersected region
GIAB truth variants
AD < 2, excluded
Outside the callable region, excluded
Outside the GIAB region, excluded
Compared by hap.py
Called variants
b
PacBio Iso-Seq HG004/HG005 SNP (BCM)
PacBio Iso-Seq HG004/HG005 SNP (Google)
PacBio MAS-Seq HG004 SNP
ONT cDNA HG004/HG005 SNP
ONT dRNA002 HG004/HG005 SNP
ONT dRNA004 HG004/HG005 SNP
Recall
Precision
Clair3-RNA (HG004)
Clair3-RNA (HG005)
Clair3* (HG004)
Clair3* (HG005)
DeepVariant* (HG004)
DeepVariant* (HG005)
LongcallR (HG004)
LongcallR (HG005)
c
Clair3-RNA's SNP performance on PacBio datasets
Clair3-RNA's SNP performance on ONT datasets
Values
Precision
Recall
F1-score
d
ONT dRNA004 (HG004 DP≥4,AD≥2)
ONT cDNA (HG004 DP≥4,AD≥2)
PacBio Iso-Seq (HG004 DP≥4,AD≥2)
PacBio MAS-Seq (HG004 DP≥4,AD≥2)
Precision
Recall
AUPRC
0.9400
0.7646
0.6607
0.6061
0.6754
0.5214
0.5221
0.4769
0.9545
0.6357
0.6146
0.7876
0.9422
0.8606
0.5160
0.8310
Clair3-RNA
Clair3*
DeepVariant*
longcallR
e
Clair3-RNA
Clair3*
DeepVariant*
LongcallR
Runtime (min)
Memory use (GB)
ONT cDNA
ONT dRNA004
PacBio Iso-Seq
PacBio MAS-Seq

919 **Figure 3. PacBio and ONT benchmarking results.**

920 (a) The RNA variant benchmarking workflow is depicted, starting with alignments utilized for coverage
921 calculation to determine the callable region. The callable region was then refined by intersecting it
922 with GIAB high-confidence regions. Variants with coverage depth (DP) of less than 4 or allele depth
923 (AD) of less than 2 were excluded from benchmarking. Evaluation metrics such as precision, recall, and
924 F1-score for SNP and Indel were obtained by comparing called variants with truth variants using hap.py.
925 (b) This shows the precision and recall performance on different sequencing datasets. The star
926 indicates that Clair3 and DeepVariant are two DNA variant caller. (c) This shows the SNP performance
927 of Clair3-RNA on PacBio and ONT datasets. (d) These are the precision-recall curves of four variant
928 callers on different datasets; the dot indicates the best F1-score. AUPRC is the area under the
929 precision-recall curve. (e) This shows the runtime and memory usage of four variant callers on
930 different datasets.

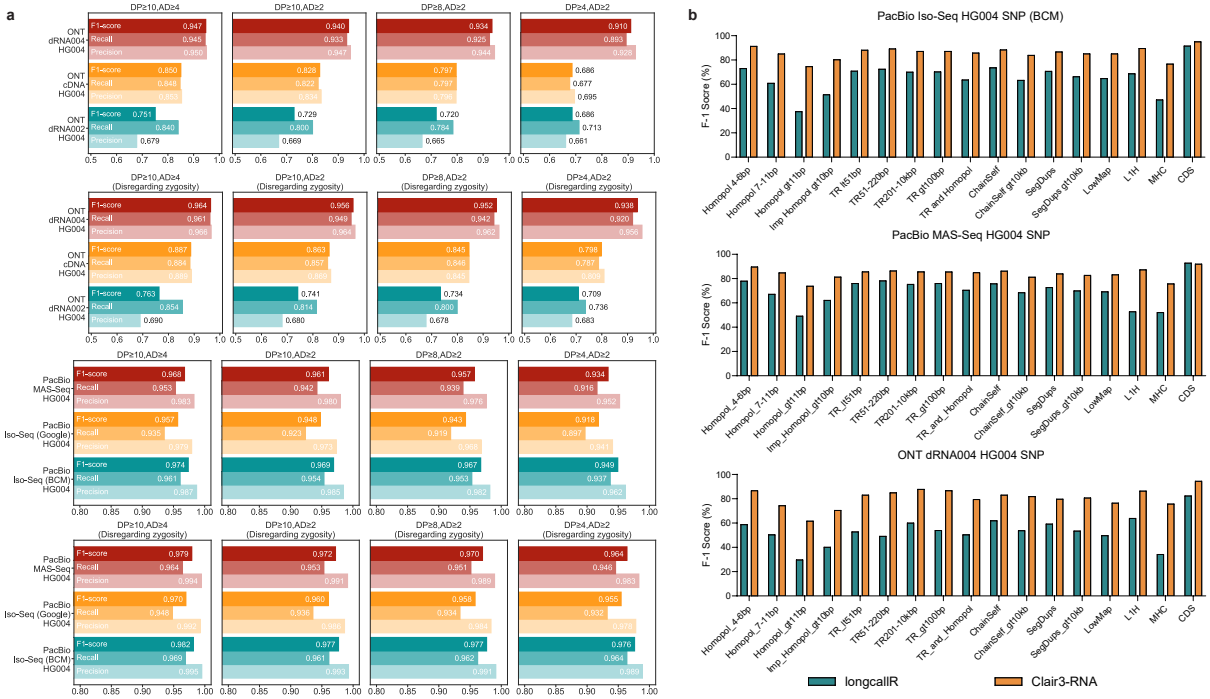


Figure 4. Performance evaluation under different settings.

(a) This shows Clair3-RNA's SNP performance on different datasets for different coverage depth (DP) and allele depth (AD) settings. The performance of "Disregarding zygosity" was determined by enabling the "--skip_genotyping" option when using the "get_overall_metrics" for metrics calculation. (b) This shows the SNP performance comparison between Clair3-RNA and LongcallR in different genomic contexts. The figure was created using Genome in a Bottle stratification v3 and hap.py. Detailed descriptions of each sub-category are provided in the **Supplementary Table 6**.

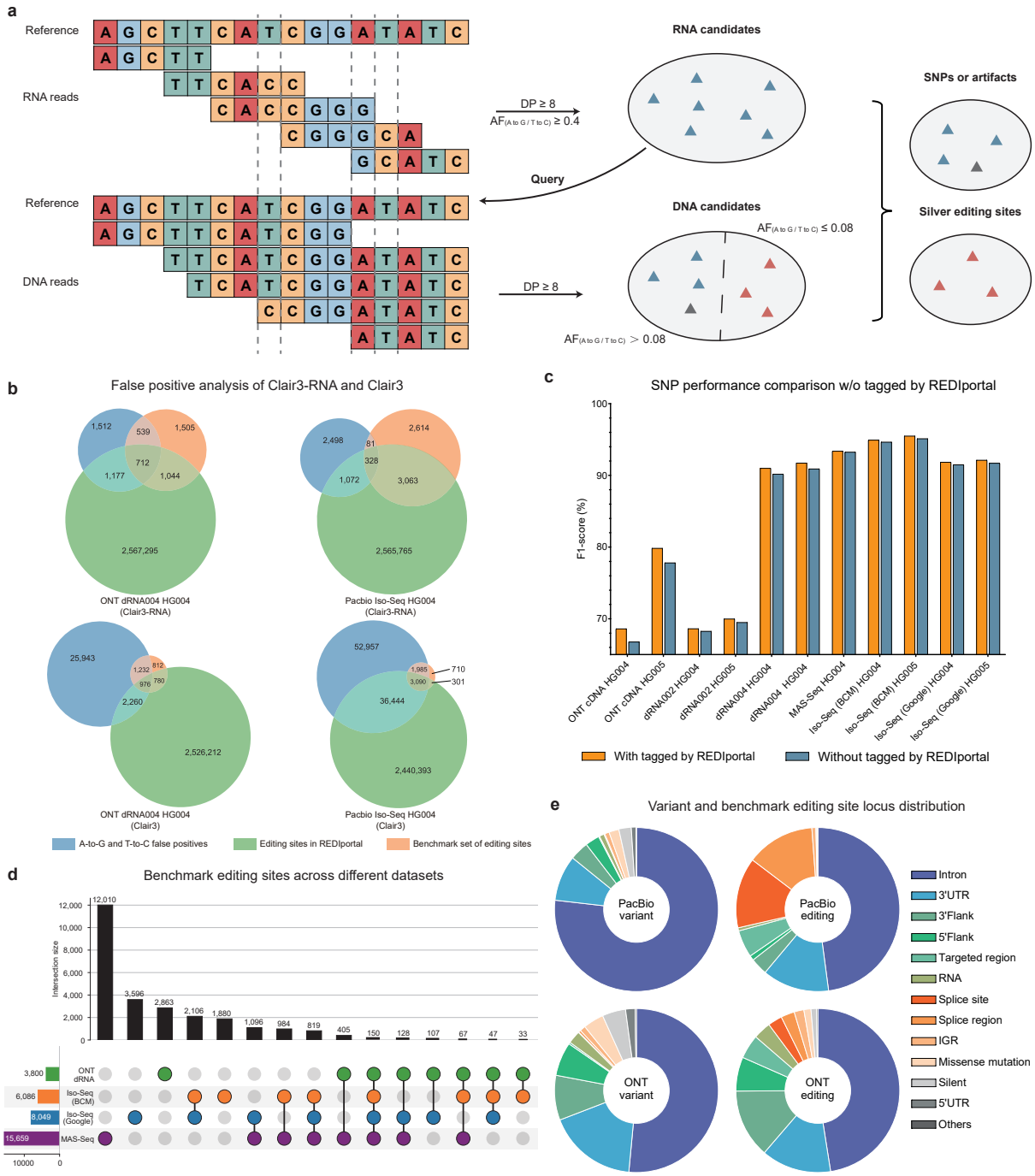


Figure 5. RNA editing event analysis.

(a) This illustrates the workflow of acquiring high-quality benchmark set of editing sites. First, candidate sites are assessed based on coverage depth and allelic fraction at various positions in RNA alignments. A site qualifies as a high-quality editing site candidate if the coverage depth exceeds 8 and the allelic fraction from A to G or T to C is above 0.75. These candidates are then cross-referenced with DNA alignments. If a position in DNA has a depth exceeding 8 and the allelic fraction from A to G or T to C is below 0.08, it is deemed a high-quality editing site; otherwise, it is classified as an SNP or an artifact. (b) Venn diagrams are presented for three sets: editing sites in REDIportal, A-to-G and T-to-C false positives, and the benchmark set of editing sites. A smaller overlap between A-to-G and T-to-C false positives and detected editing sites indicates Clair3-RNA's capability of distinguishing genuine SNPs and A-to-I editing sites. (c) This provides an SNP performance

comparison across different datasets when with (w) and without (o) variants tagged by the REDIportal database. (d) The upset plot illustrates the distribution of benchmark set of editing sites across different platforms. (e) This presents the distribution of benchmark editing sites and variants within the PacBio Iso-Seq (BCM) HG004 dataset and the ONT dRNA004 HG004 dataset.

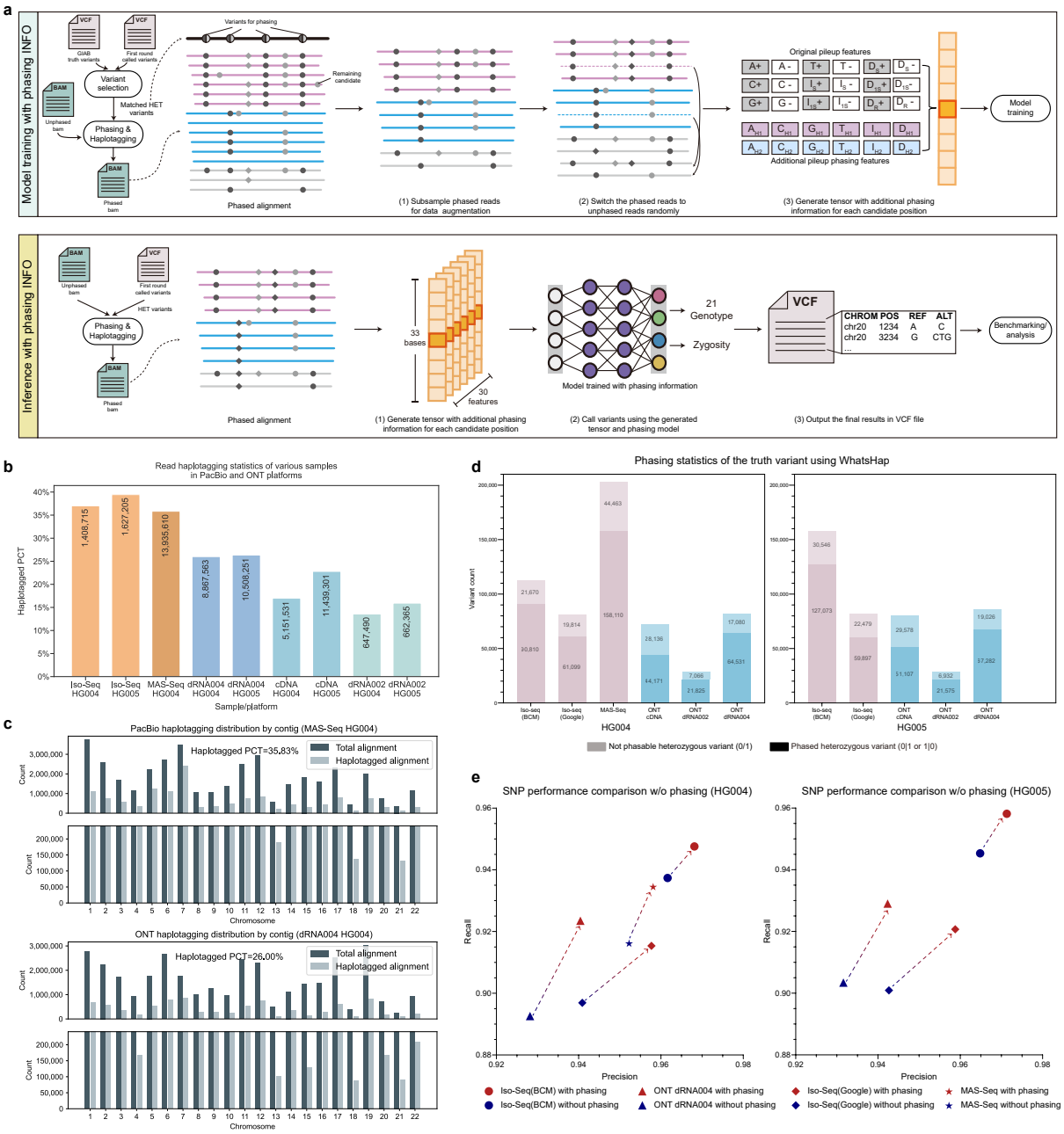


Figure 6. Phasing result analysis in lrRNA-seq.

(a) This figure illustrates the workflow of Clair3-RNA for model training and inference incorporating phasing information. During the training phase, variants were phased using the first round of called variants in conjunction with GIAB truths, and the alignments were haplotagged based on these phased variants. The alignments are color-coded, with pink representing haplotype 1, blue for haplotype 2, and grey indicating unknown haplotypes. The circle and rhombus in the alignments represent SNPs and indels, respectively. Twelve phasing features were concatenated with the original eighteen features and subsequently input into the neural network for model training. In the inference phase, heterozygous variants from the first round of called variants of Clair3-RNA were utilized for variant phasing. (b) This panel presents the read haplotagging statistics for various samples across PacBio and ONT platforms. The value on y axis refers to the percentage of haplotagged reads, and the numbers displayed in each histogram represent the total count of haplotagged reads within the respective subset. (c) This panel displays read haplotagging statistics,

969 stratified by chromosomes, for PacBio MAS-Seq HG004 and ONT dRNA004 HG004 samples. (d) This
970 panel presents variant phasing statistics for truth variants analyzed using WhatsHap across various
971 PacBio and ONT samples. (e) The precision-recall curve of the SNP performance across various
972 samples, comparing the performance with and without phasing information in the neural network.
973

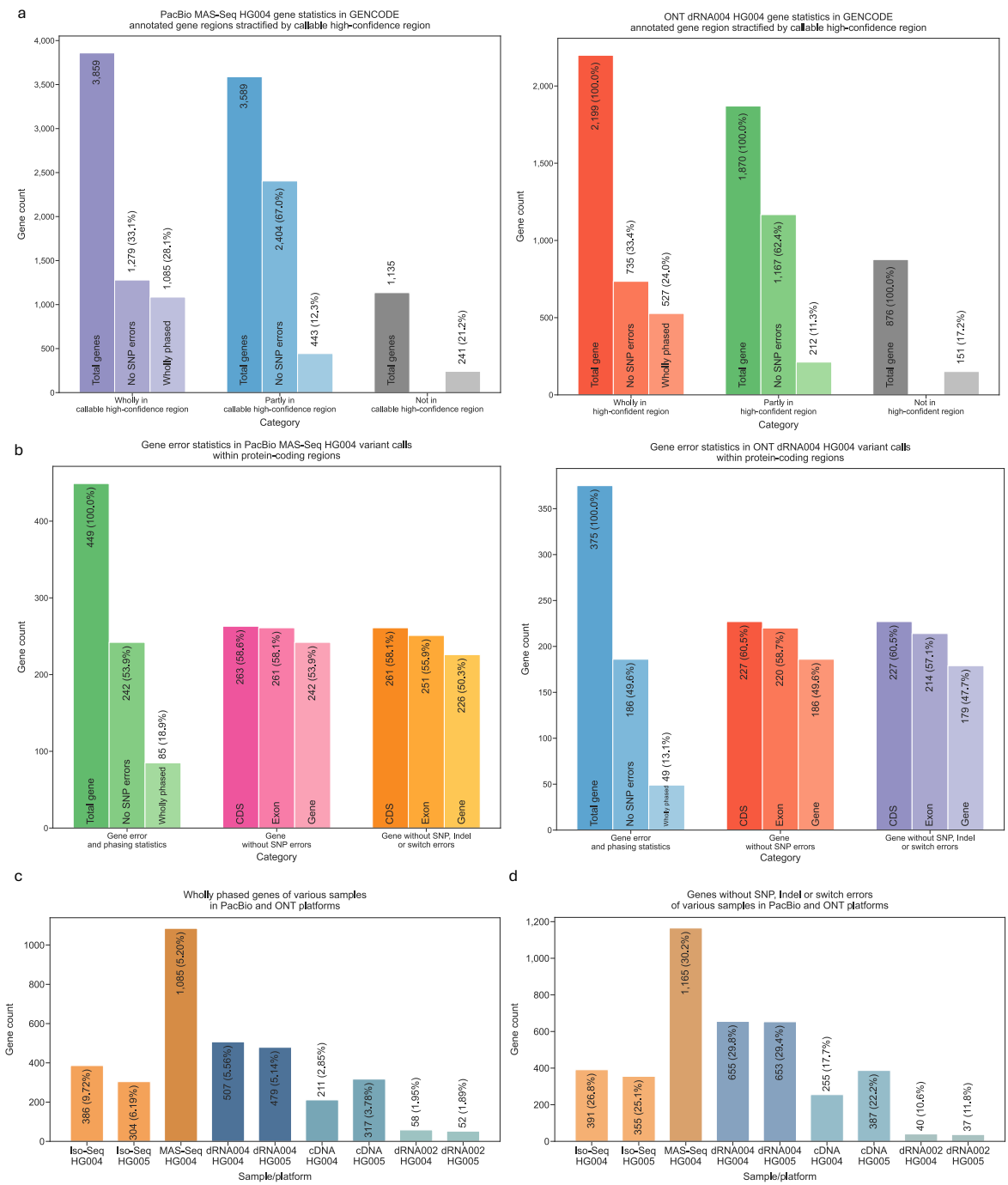


Figure 7. GENCODE annotated gene analysis.

(a) Gene statistics for PacBio MAS-Seq HG004 and ONT dRNA004 HG004 are presented, stratified by GENCODE annotated gene regions. The data is categorized into wholly phased genes and those with no SNP errors, based on their overlap with callable high-confidence regions, with percentages reflecting the proportion of each category relative to its predecessor. (b) This panel shows gene error statistics for variant calls within protein-coding regions for both PacBio MAS-Seq HG004 and ONT dRNA004 HG004 datasets. The percentages are calculated relative to the total number of genes in the subset. (d) The number of wholly phased genes across various samples from PacBio and ONT platforms is illustrated, with percentages relative to the total number of annotated genes within the callable high-confidence regions. (e) This panel depicts genes without SNP, Indel, or switch errors

985 from various samples in PacBio and ONT platforms, with percentages again relative to the total
986 annotated genes in the callable high-confidence regions.

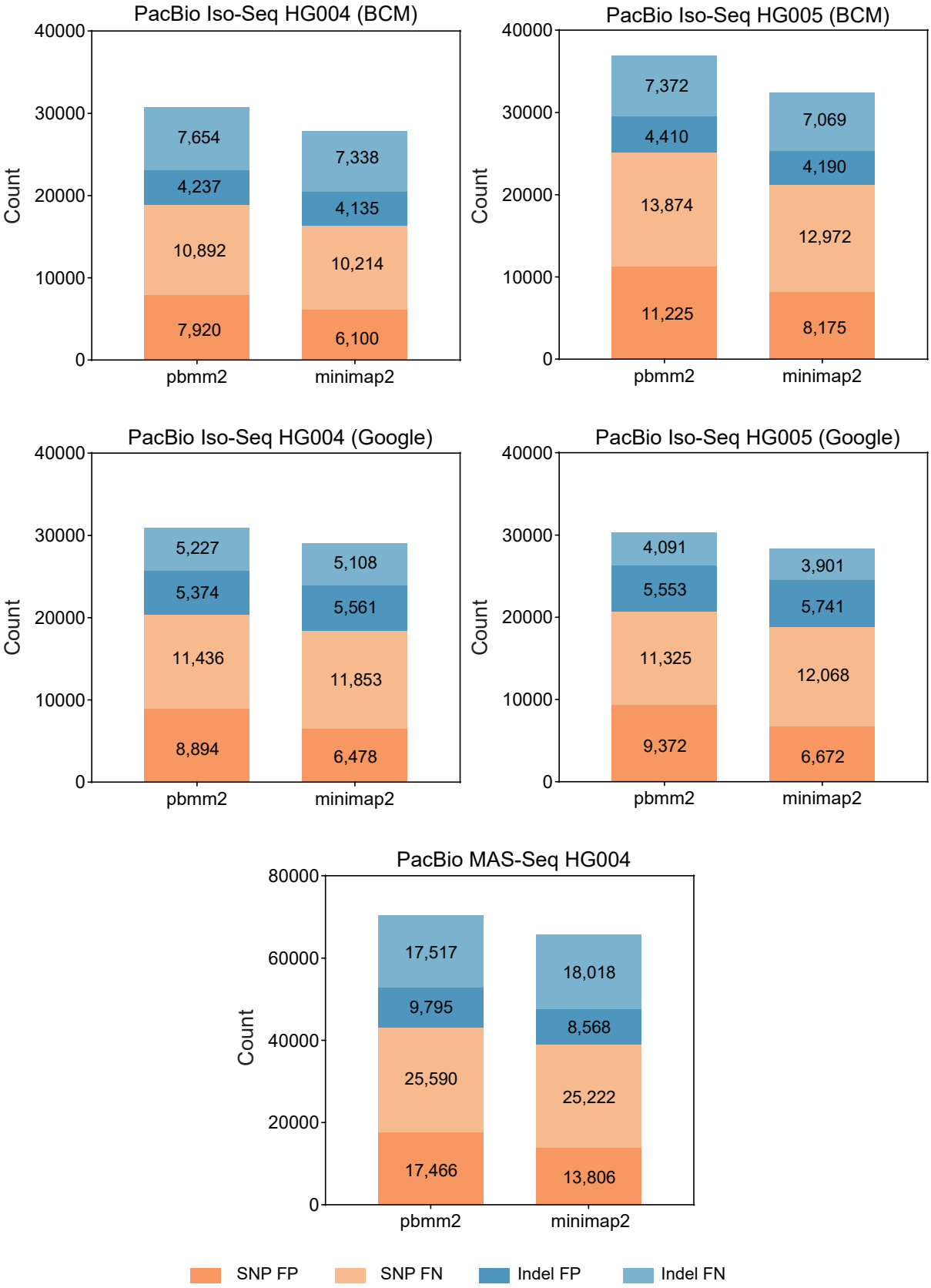


Figure 8. Clair3-RNA PacBio performance using different aligners.

This is a comparison of false positive (FP) and false negative (FN) using different alignment methods on PacBio datasets. minimap2 aligner exhibits lower FPs and FNs than pbmm2 in benchmarking

991 datasets. The alignment parameter settings for pbmm2 and minimap2 are described in the
992 **Supplementary Notes** – **Command line used** section.
993