Detecting Foldback Artifacts in Long Reads

Jakob M. Heinz[1,2,3], Matthew Meyerson[3,4,5], Heng Li[1,2,*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

[2]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, United States

[3]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[5]Department of Genetics, Harvard Medical School, Boston, MA, USA

*Corresponding author: hli@ds.dfci.harvard.edu

Abstract

Long-read sequencing data is useful for detecting large and complex structural variations; however, technical artifacts can lead to false structural variant calls. In our analyses, we became aware of a foldback artifact in long-read data. Therefore, we developed the open-source Breakinator tool to flag putative foldback artifact reads, as well as previously known chimeric artifacts. Through an alignment-based approach, Breakinator can detect artifacts missed by existing quality control tools. We profiled the occurrences of foldbacks and chimeric reads in both nanopore and single-molecule real-time sequences across a range of specimens, library types, sequencing chemistries, sequencing machines, and base-calling software.

Keywords

Long-read sequencing, Nanopore, RNA-Sequencing, Technical artifacts, Quality control

Background

Advances in long-read sequencing technologies, such as those by Pacific Biosciences (PacBio; single-molecule real-time sequencing) and Oxford Nanopore Technologies (ONT; nanopore sequencing), have made the reliable resolution of complex and repetitive regions of the genome possible, which cannot be achieved by short-read sequencing. The ability of long reads to span these complicated regions is useful for detecting large complex structural variations (SV) in the genome [1,2]. Typically, for long-read SV calling, reads are aligned to a reference genome to identify those with a supplementary alignment, a so-called split read, where a portion of the read maps to one genomic region, and the remainder maps to another distinct region causing the aligner to split the read into separate alignment blocks [3–5]. However, systemic technical artifacts in the reads can lead to false SV calls. Chimeric reads have been a pervasive issue throughout the history of long-read sequencing [6–9]. Technical artifacts can arise through issues in library preparation or computational errors in distinguishing independent molecules [10,11]

Here, we have become aware of another type of technical artifact, a foldback (also known as a duplicated inversion) [12]where half the read maps to one strand and the other half maps to the same location, but on the opposite strand, appearing as a near perfect concatemerization of the forward and reverse strands of a read. A previous study by Soneson *et al.* had noted an elevated rate of palindromic reads in ONT direct-cDNA data with supplementary alignments overlapping their primary alignment on the opposite strand. They hypothesized that these may represent un-split reads by the base caller [11]. However, to our knowledge, neither the cause nor the prevalence of these foldback artifacts was explored by any later studies with more recent data, such as the latest ONT nanopores (R10.4) or PacBio HiFi.

Results and Discussion

The Breakinator tool identifies putative foldback and chimeric artifacts by parsing a read alignment file. Our tool extracts the primary and all supplemental alignments of a read and classifies them as either a true breakpoint, chimeric, or foldback using the classification scheme in Fig. 1a. While all metrics can be set by the user, we classify a chimeric artifact as a read that has segments mapping to different chromosomes, or more than 1 Mb apart within the same chromosome. We classify a read as a foldback artifact if the supplementary alignment maps to the same location as the primary alignment (within 200 bps) on the opposite strand. For samples where true foldback events may be expected, as in cancers, we added a symmetry option to our tool to only report reads as foldback artifacts if the foldback occurs within a 5% margin on either side of the middle of the read, to retain true foldback event supporting reads.

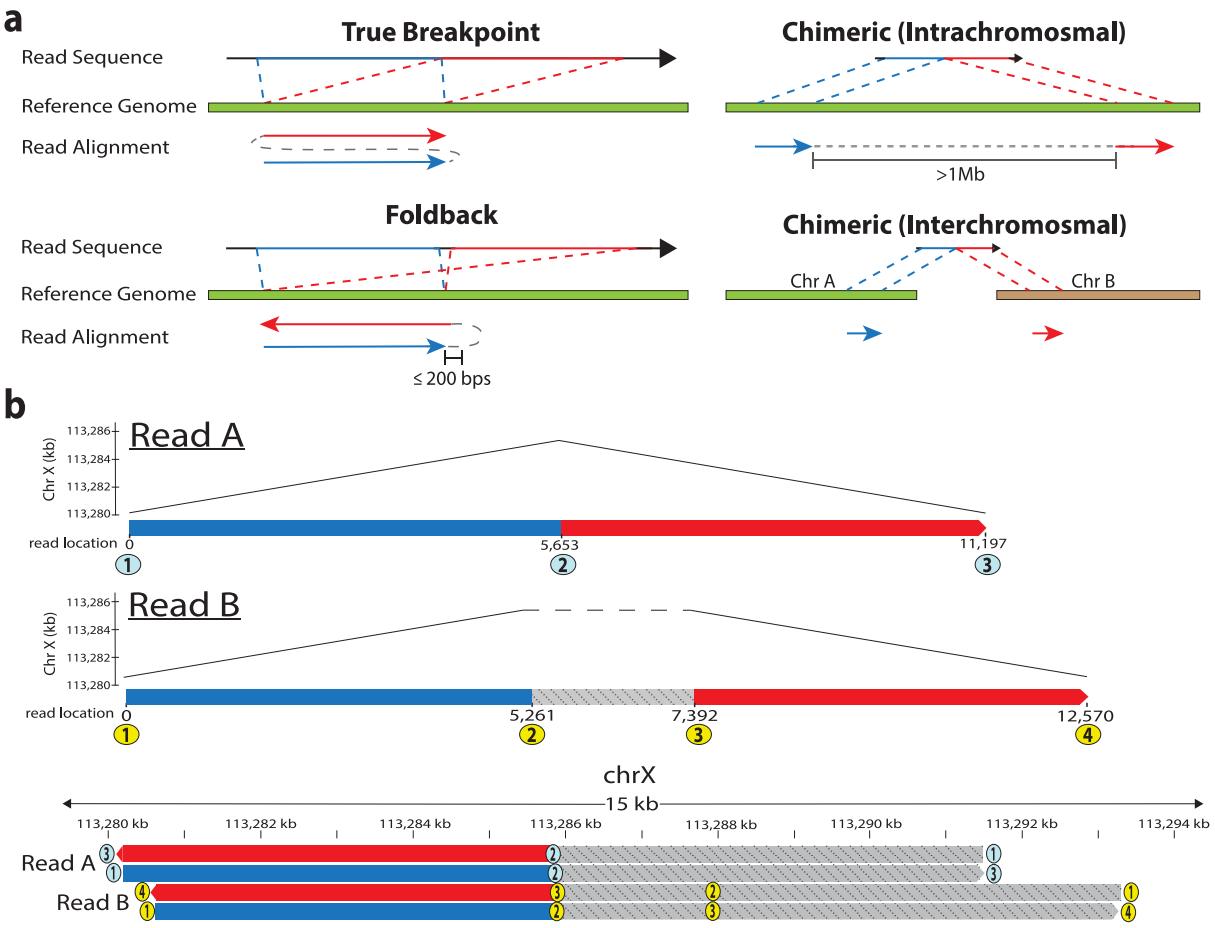


Fig. 1 a). Metrics used to define types of breakpoints observed in this study. A "True Break" is a breakpoint that is not found to be chimeric or a foldback, such as the tandem duplication example shown here. A "Foldback" is a read where the second segment maps to approximately the same location (within 200bps) as the first segment, but on the opposite strand. A "Chimeric" is where the second segment maps more than 1 Mb away on the same chromosome (intrachromosomal) or to a different chromosome (interchromosomal). b) IGV schematization of two foldback artifact reads from the HG002_PAW70337_sup sample supporting the same breakpoint at approximately 113,286 kb on chromosome X of the HG002 diploid assembly. Read A (5f80430c-2b83-408b-89b2-63b61b1173cc) is an example of a read with no sequence between the two segments of the

65 foldback. Read B (115c006d-16a6-47b1-bf5c-c5eb9bb85f69) is an example where there are ~2kb
66 of sequence between the segments.
67
68 To verify that our tool properly identifies foldback events, we used PacBio's samples of
69 self-complementary adeno-associated viruses (scAAV) CBA-eGFP as a positive control. Since
70 scAAV-CBA-eGFP constructs were designed to be perfectly palindromic, we expected our tool to
71 detect them as foldbacks. With the Breakinator, we find that 97.49% CBA-eGFP sample reads are
72 scAAV compared to the 97.69% reported by Talevich *et al*. [13]. This verification also presented
73 an interesting additional use case for the Breakinator for the rapid identification of scAAVs in a
74 sample.
75
76 We first noticed a potential technical artifact when analyzing transcriptomic data of the
77 HCC1395 cell line generated by Cotto *et al.* [14]. In the ONT direct-cDNA samples, we observed
78 many reads with supplementary alignments overlapping the primary alignment on the opposite
79 strand, indicative of a foldback SV, that were not present in the ONT direct-RNA sample run on
80 the same RNA replicate. Manual inspection of the unique breakpoint locations in IGV [15] found
81 that those detected from direct-RNA had supporting evidence from PacBio HiFi gDNA sequences,
82 while most of the direct-cDNA breakpoints did not, leading us to suspect that many of these events
83 were caused by a technical artifact. After developing the Breakinator, we used it to identify all
84 breakpoints in the sample and merge breakpoints in the same region (see Methods). We found
85 35,843 unique breakpoint locations with at least five supporting reads in the direct-cDNA samples,
86 compared to just eight in the direct-RNA. While foldback SVs occur in cancer genomes, they have

87 not been observed at such a high frequency [12], and are not expected to vary so greatly by library
88 type.
89
90 To explore whether this was a dataset-specific sequencing result or a systemic technical
91 artifact, we downloaded publicly available datasets representing four library types: direct-RNA,
92 direct-cDNA, cDNA, and gDNA, from four sample types (HG002 cell line, K562 cell line,
93 HCC1395 cell line, and mouse tissue). A range of sequencing chemistries, machines, and base-
94 calling software were evaluated through these samples (Additional file 1). The HG002 samples
95 offer a useful neutral background to investigate technical artifacts because they have a high-quality
96 personal diploid reference assembly to which the reads can be aligned [16]Therefore, we would
97 expect no germline variants, given that this is the reference, only variations between samples of
98 the cell line, which are likely limited [17]. Each sample was aligned to the corresponding reference
99 genome with the appropriate Minimap2 [3] parameters for each data type (see Methods).
100
101 While an alignment-based artifact detection method, as is implemented by the Breakinator,
102 may lead to issues in distinguishing true SV reads from artifact reads, the increased specificity
103 offered by alignment can be useful. Commonly used quality control tools for long reads, such as
104 fastplong, Porechop, pychopper, and Restrander [18–21], were not able to identify many of the
105 foldback or chimeric reads flagged by the Breakinator (Additional file 2). We suspect this is due
106 to their reliance on identifying adapter sequences within the middle of a read sequence to identify
107 a concatemerized read. We found that typically less than half of the putative artifactual reads
108 detected by the Breakinator had more than 11 bps of non-mapping sequence between alignments
109 (Additional file 3), making it unlikely they will be detected with existing methods. This is

110 highlighted in Fig. 1b, where two foldback reads identified by the Breakinator are shown. The
111 reads are both nearly perfectly palindromic and map to the same breakpoint location, yet Read A
112 has no sequence between mappings, while Read B has ~2kb of sequence that could be searched
113 for retained adaptors. 27 other reads (not shown) cover this foldback location, with no indication
114 of a breakpoint.
115
116 Summary statistics across library types and samples evaluated with the Breakinator are
117 shown in Fig. 2a. ONT direct-cDNA samples were found to have substantially elevated levels of
118 foldback artifacts (~10-25% of all reads). Meanwhile, at most one foldback read was observed in
119 each of the ONT direct-RNA samples. The ONT cDNA samples had the lowest rates of foldback
120 artifacts (~0.1 – 0.2). We suspect this low rate is due to the cDNA samples being the only datatype
121 in this study that undergoes PCR amplification, and palindromic sequences such as these foldbacks
122 are known to be difficult to PCR amplify as they will bind to themselves [22,23]. In ONT gDNA,
123 we observed foldback rates of ~0.5–0.7%. While only three PacBio samples were investigated,
124 they did not show any evidence of systemic foldback artifacts. There were virtually no foldback
125 reads observed in the PacBio HiFi (Revio) or Mas-Seq (Vega) data, while 0.04% of the Iso-Seq
126 (Sequel II) reads were considered foldbacks. For five of the samples, from K562 and HCC1395
127 cells, the same RNA replicate was sequenced with different technologies, allowing for direct
128 comparisons in the rates of foldbacks and chimeras across libraries without replicate bias. It was
129 observed that for all replicates, the rates of foldbacks and chimeras were much higher in ONT
130 direct-cDNA library samples, as shown in Fig. 2b.

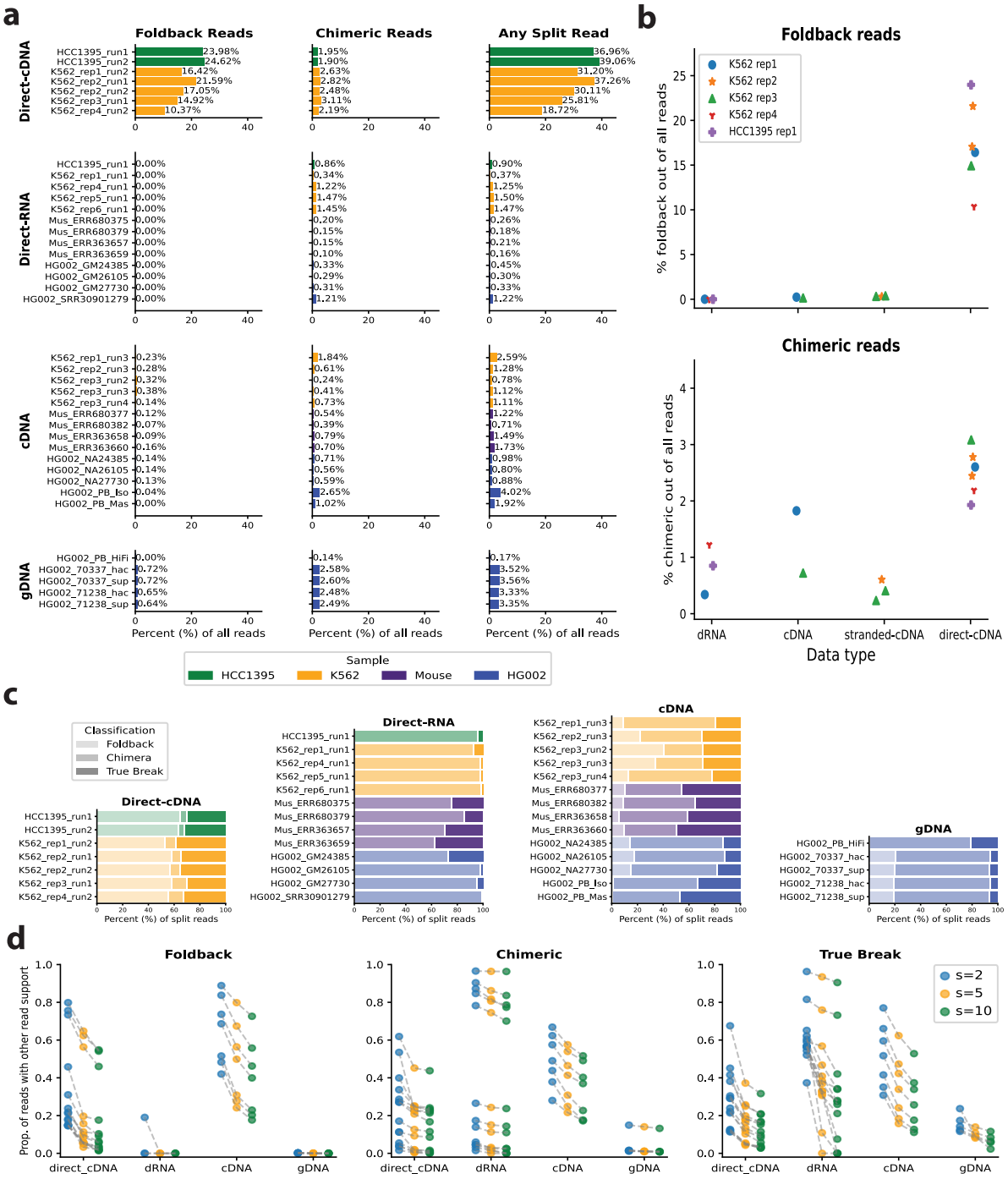


Fig. 2. a) The observed rates of Foldback, Chimeric, and any split reads out of all reads that passed the filtering criteria for direct-cDNA, direct-RNA, cDNA, and gDNA data from four different sample types: HCC1395, K562, mouse tissue, and HG002. B) Paired comparisons for the five RNA replicates that were sequenced with different technologies for both foldback and chimeric

136 reads. c) The proportion of Foldback, Chimeric, and True Break reads in relation to the number of
137 split reads. d) Recurrence of breakpoints at 2, 5, and 10 supporting reads for Foldback, Chimeric,
138 and True Break reads.
139
140 While the rates of putative foldback and chimeric artifacts across all reads are relatively
141 low for libraries other than ONT direct-cDNA, they represent a far larger proportion of the reads
142 used for SV calling, namely, reads with split alignments. Fig. 2c highlights that these artifacts can
143 cause issues even with ONT cDNA or gDNA data. In cDNA, foldback artifacts account for 10-
144 20% of all split reads, while chimeras are ~30-40%. In gDNA, foldbacks are also 15-20% of split
145 reads, and chimeric reads are 50-60% of all split reads (Fig. 2.c). These artifactual reads were also
146 observed to be recurrent at the same locations, which made it more likely they would be called as
147 a false positive SV. Fig. 2d highlights that many of the locations to which these reads align have
148 more than 2, 5, or 10 reads of support.
149
150 Given that the rates of foldback artifacts appear relatively consistent among runs of the
151 same data type, but vary between data types that are prepared with different chemistries, we believe
152 it is most likely that these artifacts arise during library preparation, rather than a failure of the base-
153 caller to recognize a pore-open signal or an adapter sequence, as was hypothesized by Soneson *et*
154 *al*. [11]. Additionally, we observed both forward and reverse strands of foldback and chimeric
155 reads, which can only arise during library preparation, compared to an informatic chimera, which
156 would only occur in one reading direction. Finally, we investigated the raw signal of eight putative
157 artifactual reads from the SGNEx K562 direct-cDNA replicate 2 run 1 to visualize the electrical
158 signal interpreted by the base-caller. We did not observe any notable difference in signal around

159 the breakpoint region for artifacts compared to true breakpoints (Additional file 4). Thus, we
160 conclude that Breakinator can identify foldback artifacts in sequencing data, which may arise due
161 to an experimental feature in library preparation.

162 Conclusions

163 We present evidence of a new technical artifact, a foldback read. If these technical artifacts
164 are not properly removed, they can lead to tangles in an assembly graph or false-positive SV calls.
165 We developed the open-source Breakinator tool to flag putative foldback and chimeric technical
166 artifacts in long-read sequences. Many current quality control tools rely on detecting incorrectly
167 incorporated adapter sequences, but through a reference alignment-based approach, we can detect
168 reads that may not have retained any non-sample sequence, but still likely represent a non-real
169 biological concatenation. In this study, we profiled both ONT and PacBio data across a range of
170 specimens, library types, sequencing chemistries, sequencing machines, and base-callers. We find
171 that foldback artifacts occur throughout ONT library types and are particularly high in direct-
172 cDNA libraries. We believe the Breakinator will offer a useful tool for quality control in the
173 analysis of long-read data.

174 Methods

175 All data used in this study were downloaded from publicly available sources, with the links
176 available in Additional file 1. The K562 samples were generated by the SGN-Ex consortium [24].
177 HCC1395 samples were generated by Cotto *et al.* [14]. The mouse tissue samples were generated
178 by Sessegolo *et al.* [25]. The HG002 gDNA samples were downloaded from ONT's 2025 Genome
179 in a Bottle Data (GIAB) Release [26]. The HG002 direct-RNA002, PacBio Iso-seq, and Mas-seq

180 were made available by GIAB. The HG002 dRNA sample SRR30901279 and the HG002 cDNA
181 data were generated by Zheng *et al.*[27] The HG002 PacBio HiFi gDNA is a public dataset made
182 available by PacBio [28]. The PacBio scAAV sample was generated by Talevich *et al*. [13]
183
184 The samples were aligned to their respective reference genomes using Minimap2 [3] K562
185 and HCC1395 samples were aligned to T2T-CHM13v2.0 [29,30]. The mouse samples were
186 aligned to GRCm39 [31]. HG002 samples were aligned to the HG002v1.0.1 diploid genome
187 [16]To align dRNA ONT, the following parameters were used: "-cx splice -uf -k14 --
188 secondary=no". For ONT direct-cDNA and cDNA alignments, "-cx splice --secondary=no"
189 parameters were used. ONT gDNA data were aligned with "-cx map-ont --secondary=no"
190 parameters. For PacBio Iso-Seq and Mas-Seq data, we used "-cx splice:hq -uf --secondary=no"
191 parameters. Finally, PacBio HiFi data were aligned with "-cx map-pb-- secondary=no." Aligning
192 to a diploid genome decreases the mapping quality of many reads, as there will likely be at least
193 two locations where they map well. To address this issue, HG002 sample reads were aligned to the
194 HG002v1.0.1 diploid genome by aligning the reads to each haplotype separately using Minimap2
195 with the flags "--secondary=no" and "--paf-no-hit" specified. We wrote a preprocessing script for
196 the Breakinator to compare the alignment of every read to the maternal and paternal haplotype
197 assemblies, choosing to keep the alignment with the highest alignment score. If the alignment
198 scores were equal, then the maternal haplotype alignment was chosen.
199
200 The PAF files were generated using either "paftools.js sam2paf -p" on a SAM file or
201 Minimap2 with the "--secondary=no" flag. We designed the Breakinator to parse the input PAF
202 file to detect reads with supplementary alignments. Secondary alignments were ignored.

203 Alignments with mapping quality less than 10 or lengths less than 200 bps were filtered out. The
204 primary and all supplementary alignments for a given read were identified and sorted by read start
205 location. If the first read segment mapped to the forward strand, then the end of the alignment of
206 the first segment and the start of the second segment were identified as the breakpoint location. If
207 the first segment mapped to the reverse strand, the reciprocal start and end locations were taken
208 for the breakpoint. As shown in Fig. 1a, if the breakpoint was to the opposite strand within 200
209 bps of the end of the alignment of the first segment, it was classified as a foldback. If the breakpoint
210 was to another chromosome (interchromosomal) or more than 1 Mb away (intrachromosomal), it
211 was classified as a chimeric read. The 1 Mb default threshold for intrachromosomal chimeras was
212 chosen because only two introns, out of 177,839, on MANE selected protein-coding genes were
213 longer than 1Mb [32]. If the breakpoint was not chimeric or a foldback, then it was considered a
214 true breakpoint. The identified breakpoints and their classification were generated for every
215 sample. While uncommon, it is possible to have multiple breakpoints on the same read. To classify
216 at the read level, we used the following approach: if there was any foldback or chimeric breakpoint
217 on the read, regardless of how many true breakpoints there were, it was classified as an artifact. If
218 there were multiple foldback or chimeric breakpoints on the read, the read was classified according
219 to which breakpoint occurred most frequently. If they were equal, then the read was classified as
220 chimeric. All summary statistics are generated using the read classification.
221
222 To compare how existing tools detect the reads we had identified as artifactual, we filtered
223 all reads with fastplong, porechop, and restrander, respectively [18,19,21]. Fastplong was run with
224 the following command: "fastplong -i sample.fastq -o filter_pass.fastq --failed_out
225 failed_reads.fastq." Restrander was run only on our cDNA samples as it didn't support any other

226 data types, with the command “restrander sample.fastq filter_pass.fastq config.json” Restrander
227 required a configuration file of adaptor sequences, so for samples that did not have their library
228 version supported by Restrander, we ran them with the closest version of the library supported by
229 Restrander. For example, DCS108 samples were run with the SQK-DCS109 config file. Porechop
230 was also only run on the transcriptomic data using the command: “porechop -i sample.fastq --
231 discard_middle -o filter_pass.fastq.” For each tool, the read IDs that passed their respective filter
232 were used to subset the reads we found to pass the Breakinator default filter of a minimum of 200
233 bps of alignment and a mapping quality of 10. The reads classified as foldback, chimeric, or any
234 breakpoint were similarly subset to include only the reads that passed the respective filter to
235 calculate the proportions reported in Additional file 2. PyChopper was only run on the three
236 HCC1395 samples, as we found it to be too slow to run on all samples. The length of the unaligned
237 sequence between aligning segments of the breakpoint was identified by running the Breakinator
238 with the “--rcoord” flag to output the read coordinates of the breakpoint along with the reference
239 coordinates of the breakpoint. The reads were subset to foldback, chimeric, and true breakpoints,
240 and the proportions of reads with at least 11 bps between the read coordinates of the breakpoint
241 were calculated for each subset in Additional file 3.
242
243 To generate summary statistics for Fig.2, the Breakinator was run on every sample using
244 the “-i in.paf --tabular -o out.txt” parameters. The PacBio scAAV sample was evaluated with the
245 Breakinator using the stricter parameters of a symmetry filter and half the default margin (--sym -
246 -margin 0.025). To determine which breakpoints were recurrent, we extracted the subset of
247 breakpoints that were flagged as either “Foldback”, “Chimeric”, or “True Break”, respectively,
248 then for each file we merged any breakpoints that were within 100 bps of each other to account for

249 inexact alignments around the breakpoint, to determine the number of reads supporting a
250 breakpoint using our merge_breaks.py script. The merging process was run with the -s flag set to
251 2, 5, and 10, which sets the minimum number of reads needed to support a breakpoint. The reads
252 that had support of other reads at the respective support thresholds were summed and divided by
253 the total number of reads in the subset to determine the proportion of reads with support, as shown
254 in Fig. 2d.
255
256 The SGNEx K562 direct-cDNA replicate 2 run 1 sample was chosen to investigate the raw
257 electrical signal that the ONT base caller interprets because its fast5 files were publicly available,
258 and we discovered numerous putative artifacts in the reads. The fast5 files were indexed with
259 nanopolish using "nanopolish index -d fast5_directory reads.fastq" [33]. Nanopolish eventalign
260 was used to determine which region of the raw signal corresponds to the breakpoint location in the
261 read. To explore the signal in read coordinates, we converted the read fastq file to a fasta file to
262 use as the genome input, then aligned the reads to themselves with "minimap2 -ax map-ont -o
263 reads2reads.sam reads.fastq reads.fa " and used this as the BAM input for the eventalign command
264 as follows: "nanopolish eventalign -r reads.fastq -b reads2reads.bam -g reads.fasta --signal-index".
265 We chose eight reads in total to investigate the raw signal. We investigated foldbacks, chimeras,
266 and true breakpoints with and without unaligned sequences in the middle, as well as a read with
267 no breakpoint. The start and end of the breakpoint locations in the read were determined from the
268 Breakinator output and converted to the signal coordinates using the nanopolish eventalign output.
269 The fast5 samples were graphed, and the breakpoint coordinates in the signal were highlighted to
270 manually investigate the signal for any signs of increased noise around the breakpoint.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Additional files

Additional_file_1.xlsx – All public data used in this study and their sources. Includes metadata on datatype, sequencer, pore, chemistry, and base-caller.

Additional_file_2.xlsx – Proportions of reads remaining, including artifact subsets, after filtering with other current tools (fastplong, Porechop, pychopper, and Restrander).

Additional_file_3.xlsx – Proportions of reads that had 11 bps or greater of unmapped read sequence between alignment blocks.

Additional_file_4.html – SGNEx K562 direct-cDNA replicate 2 run 1 raw signal analysis around identified breakpoints for eight reads.

Availability of data and materials

The Breakinator tool and additional software are available on GitHub at https://github.com/jheinz27/breakinator. All data used in this study were previously made publicly available, and their sources are listed in Additional file 1. The alignment files for every sample, as well as all breakpoints identified and their Breakinator classification, are available on Zenodo at: https://zenodo.org/records/15882143.

Competing interests

M.M. holds equity in Bayer, Delve Bio, Isabl, and Karyoverse; consults for Delve Bio; receives research funding from Bayer; and receives patent licensing payments from Bayer and Labcorp.

Funding

J.M.H. is supported by the U.S. National Institutes of Health training grant T32HG002295. M.M. is supported by NIH grants R35CA197568 and U24CA264029. H.L. is supported by NIH grants U24CA294203 and R01HG010040

Authors' contributions

J.M.H., M.M., and H.L conceived the project. J.M.H. implemented the method, analyzed the data, and drafted the manuscript. M.M. and H.L provided edits. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank Drs. Neng Huang, Owen Hirschi, Alaina Shumate, Maximillian Marin, and Victoria Popic, as well as all members of the Meyerson and Li labs, for their helpful discussions.

References

1. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20.
2. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing.

311 Nature Methods 2018 15:6 [Internet]. 2018 [cited 2025 Apr 17];15:461–8. Available
312 from: https://www.nature.com/articles/s41592-018-0001-7
313 3. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics [Internet]. 2018
314 [cited 2025 Apr 15];34:3094–100. Available from:
315 https://dx.doi.org/10.1093/bioinformatics/bty191
316 4. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection
317 of mosaic and population-level structural variants with Sniffles2. Nature Biotechnology
318 2024 42:10 [Internet]. 2024 [cited 2025 Apr 15];42:1571–80. Available from:
319 https://www.nature.com/articles/s41587-023-02024-y
320 5. Keskus A, Bryant A, Ahmad T, Yoo B, Aganezov S, Goretsky A, et al. Severus: accurate
321 detection and characterization of somatic structural variation in tumor genomes using
322 long reads. medRxiv [Internet]. 2024 [cited 2025 Apr 15];2024.03.22.24304756.
323 Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10996739/
324 6. Buck D, Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, et al. Comprehensive
325 comparison of Pacific Biosciences and Oxford Nanopore Technologies and their
326 applications to transcriptome analysis. F1000Res [Internet]. 2017 [cited 2025 Jun
327 13];6:100. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5553090/
328 7. Eccles D, White R, Pellefigues C, Ronchese F, Lamiable O. Investigation of chimeric reads
329 using the MinION. F1000Res [Internet]. 2017 [cited 2025 Jun 13];6:631. Available from:
330 https://pmc.ncbi.nlm.nih.gov/articles/PMC5600009/
331 8. Yang C, Lo T, Nip KM, Hafezqorani S, Warren RL, Birol I. Characterization and simulation of
332 metagenomic nanopore sequencing data with Meta-NanoSim. Gigascience [Internet].

333 2022 [cited 2025 Jun 13];12:1–11. Available from:
334 https://dx.doi.org/10.1093/gigascience/giad013
335 9. Stock W, Rousseau C, Dierickx G, D'hondt S, Amadei Martínez L, Dittami SM, et al.
336 Breaking free from references: a consensus-based approach for community profiling with
337 long amplicon nanopore data. Brief Bioinform [Internet]. 2024 [cited 2025 Jun 13];26.
338 Available from: https://dx.doi.org/10.1093/bib/bbae642
339 10. Li Y, Wang T-Y, Guo Q, Ren Y, Lu X, Cao Q, et al. A Genomic Language Model for Chimera
340 Artifact Detection in Nanopore Direct RNA Sequencing. bioRxiv [Internet]. 2024 [cited
341 2025 Jun 13];2024.10.23.619929. Available from:
342 https://pmc.ncbi.nlm.nih.gov/articles/PMC11526916/
343 11. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A
344 comprehensive examination of Nanopore native RNA sequencing for characterization of
345 complex transcriptomes. Nature Communications 2019 10:1 [Internet]. 2019 [cited 2025
346 Apr 17];10:1–14. Available from: https://www.nature.com/articles/s41467-019-11272-z
347 12. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic
348 structural variation in human cancer genomes. Nature 2020 578:7793 [Internet]. 2020
349 [cited 2025 Apr 17];578:112–21. Available from:
350 https://www.nature.com/articles/s41586-019-1913-9
351 13. Talevich E, Tseng E, Diallo A, Sellami N, Elliott A, Cantarel BL, et al. Standardized
352 Nomenclature and Reporting for PacBio HiFi Sequencing and Analysis of rAAV Gene
353 Therapy Vectors. bioRxiv [Internet]. 2024 [cited 2025 Jun 11];2024.05.07.592296.
354 Available from: https://www.biorxiv.org/content/10.1101/2024.05.07.592296v1

14. Cotto KC, Feng YY, Ramu A, Richters M, Freshour SL, Skidmore ZL, et al. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. Nature Communications 2023 14:1 [Internet]. 2023 [cited 2025 Apr 17];14:1–18. Available from: https://www.nature.com/articles/s41467-023-37266-6

15. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol [Internet]. 2011 [cited 2025 Apr 18];29:24. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3346182/

16. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol [Internet]. 2023 [cited 2025 Jun 8];41:1474–82. Available from: https://www.nature.com/articles/s41587-023-01662-6

17. Liu Y, Mi Y, Mueller T, Kreibich S, Williams EG, Van Drogen A, et al. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat Biotechnol [Internet]. 2019 [cited 2025 May 26];37:314–22. Available from: https://www.nature.com/articles/s41587-019-0037-y

18. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta [Internet]. 2023 [cited 2025 May 17];2:e107. Available from: /doi/pdf/10.1002/imt2.107

19. rrwick/Porechop: adapter trimmer for Oxford Nanopore reads [Internet]. [cited 2025 May 17]. Available from: https://github.com/rrwick/Porechop

20. epi2me-labs/pychopper: cDNA read preprocessing [Internet]. [cited 2025 May 17]. Available from: https://github.com/epi2me-labs/pychopper

21. Schuster J, Ritchie ME, Gouil Q. Restrander: rapid orientation and artefact removal for long-read cDNA data. NAR Genom Bioinform [Internet]. 2023 [cited 2025 May 17];5:lqad108. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10748469/

22. Devine SE, Chissoe SL, Eby Y, Wilson RK, Boeke JD. A Transposon-Based Strategy for Sequencing Repetitive DNA in Eukaryotic Genomes. Genome Res [Internet]. 1997 [cited 2025 Jun 9];7:551–63. Available from: http://genome.cshlp.org/content/7/5/551.full

23. Tanaka H, Tapscott SJ, Trask BJ, Yao MC. Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. Proc Natl Acad Sci U S A [Internet]. 2002 [cited 2025 Jun 9];99:8772–7. Available from: /doi/pdf/10.1073/pnas.132275999?download=true

24. Chen Y, Davidson NM, Wan YK, Yao F, Su Y, Gamaarachchi H, et al. A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. Nature Methods 2025 22:4 [Internet]. 2025 [cited 2025 Jun 10];22:801–12. Available from: https://www.nature.com/articles/s41592-025-02623-4

25. Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, et al. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. Sci Rep [Internet]. 2019 [cited 2025 Jun 2];9:1–12. Available from: https://www.nature.com/articles/s41598-019-51470-9

26. Genome in a Bottle Data Release 2025.01 | EPI2ME Blog [Internet]. [cited 2025 Jun 10]. Available from: https://epi2me.nanoporetech.com/giab-2025.01/

27. Zheng Z, Yu X, Chen L, Lee Y-L, Xin C, On A, et al. Clair3-RNA: A deep learning-based small variant caller for long-read RNA sequencing data. [cited 2025 Jun 10]; Available from: https://doi.org/10.1101/2024.11.17.624050

28. Datasets - PacBio - Highly accurate long-read sequencing [Internet]. [cited 2025 Jun 10]. Available from: https://www.pacb.com/connect/datasets/

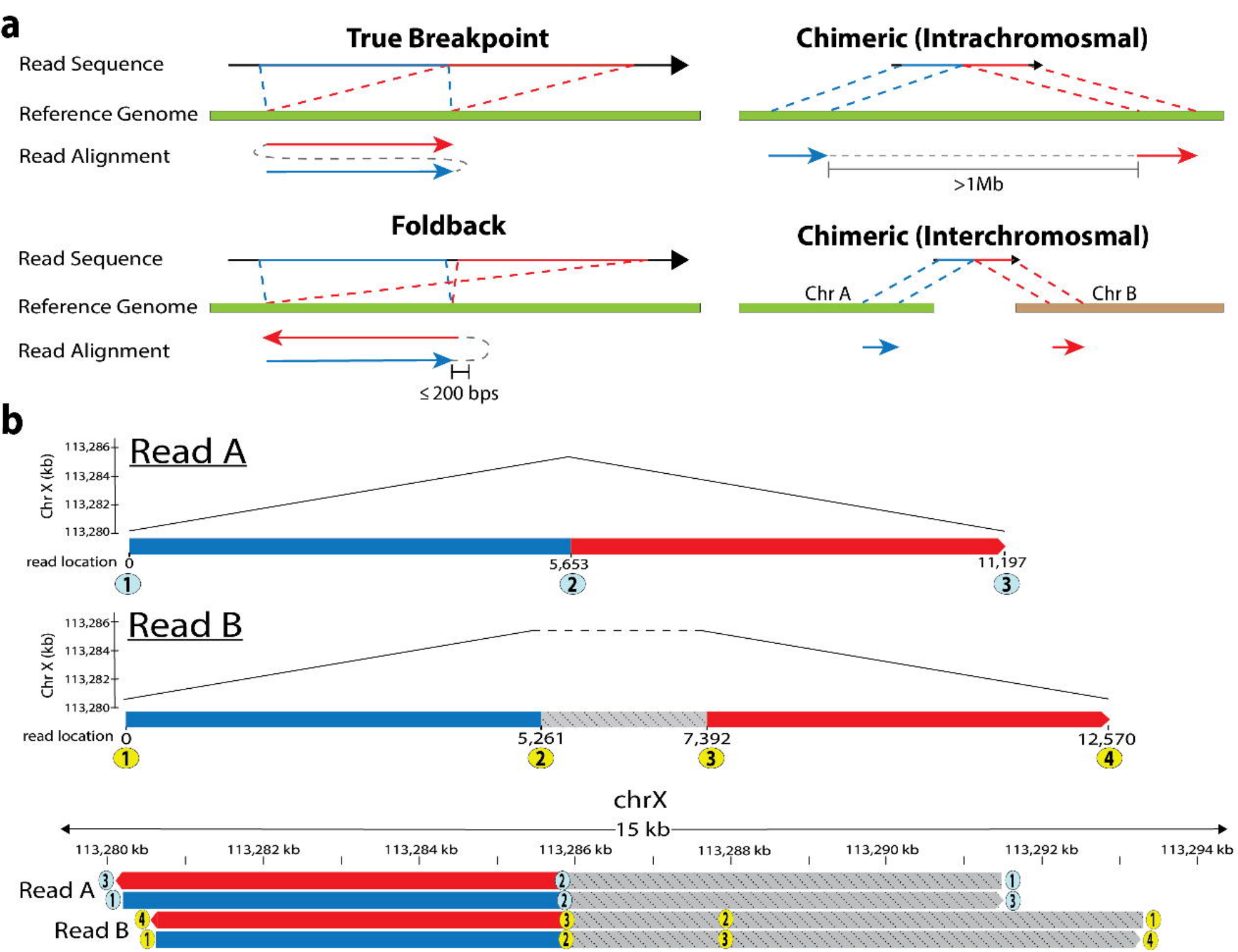
29. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al. The complete sequence of a human genome. Science (1979) [Internet]. 2022 [cited 2025 Apr 17];376:44–53. Available from: https://www.science.org/doi/10.1126/science.abj6987

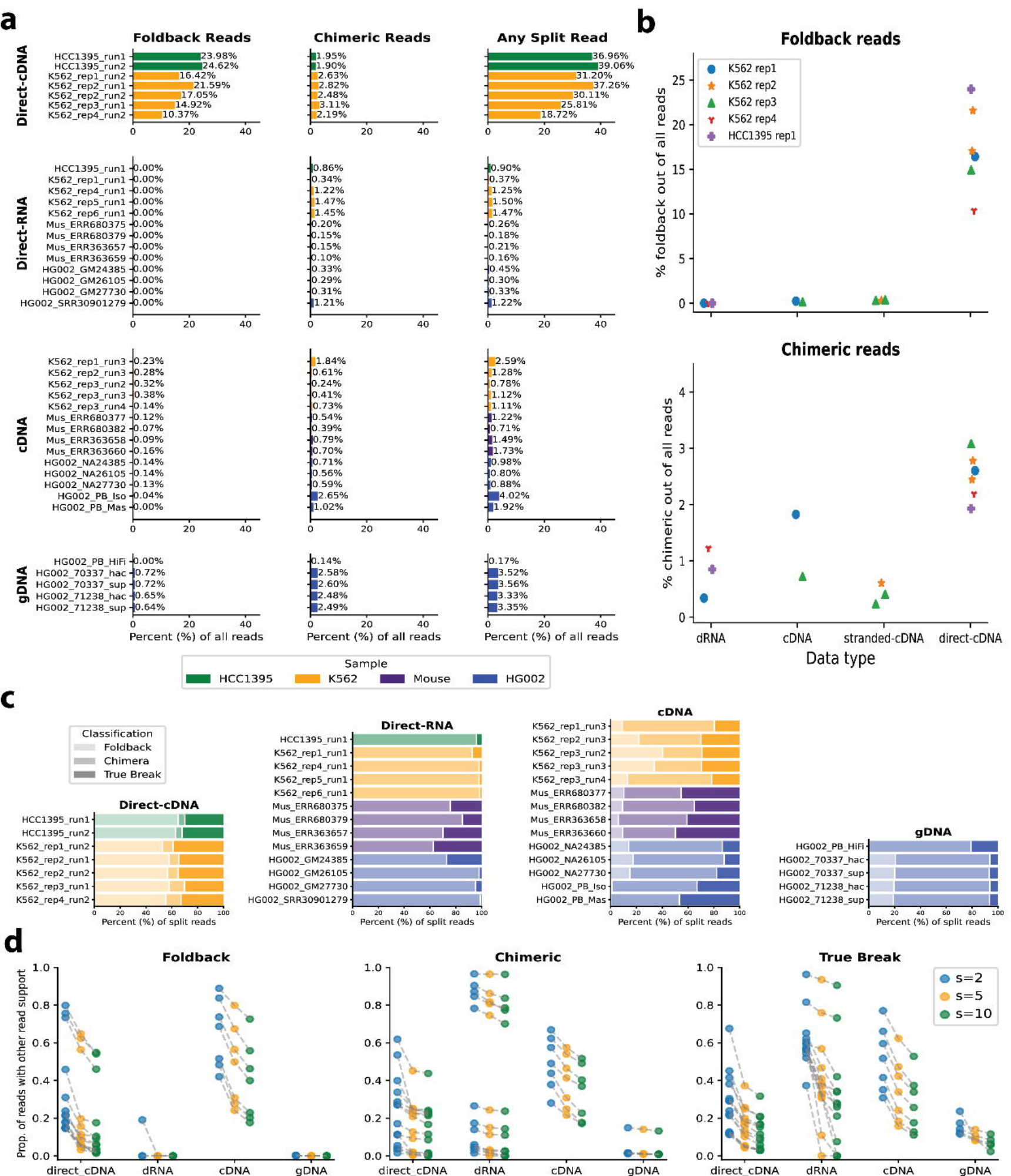
30. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. Nature 2023 621:7978 [Internet]. 2023 [cited 2025 Apr 17];621:344–54. Available from: https://www.nature.com/articles/s41586-023-06457-y

31. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature [Internet]. 2002 [cited 2025 Jul 8];420:520–62. Available from: https://pubmed.ncbi.nlm.nih.gov/12466850/

32. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature 2022 604:7905 [Internet]. 2022 [cited 2025 Jun 21];604:310–5. Available from: https://www.nature.com/articles/s41586-022-04558-8

33. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods 2015 12:8 [Internet]. 2015 [cited 2025 Jun 11];12:733–5. Available from: https://www.nature.com/articles/nmeth.3444


a
True Breakpoint
Read Sequence
Reference Genome
Read Alignment
Chimeric (Intrachromosmal)
>1Mb
Foldback
Read Sequence
Reference Genome
Read Alignment
≤ 200 bps
Chimeric (Interchromosmal)
Chr A
Chr B
b
Read A
Chr X (kb)
113,286
113,284
113,282
113,280
read location
0
5,653
11,197
1
2
3
Read B
Chr X (kb)
113,286
113,284
113,282
113,280
read location
0
5,261
7,392
12,570
1
2
3
4
chrX
15 kb
113,280 kb
113,282 kb
113,284 kb
113,286 kb
113,288 kb
113,290 kb
113,292 kb
113,294 kb
Read A
Read B


a
Foldback Reads
Chimeric Reads
Any Split Read
Direct-cDNA
HCC1395_run1 23.98% 1.95% 36.96%
HCC1395_run2 24.62% 1.90% 39.06%
K562_rep1_run2 16.42% 2.63% 31.20%
K562_rep2_run1 21.59% 2.82% 37.26%
K562_rep2_run2 17.05% 2.48% 30.11%
K562_rep3_run1 14.92% 3.11% 25.81%
K562_rep4_run2 10.37% 2.19% 18.72%
Direct-RNA
HCC1395_run1 0.00% 0.86% 0.90%
K562_rep1_run1 0.00% 0.34% 0.37%
K562_rep4_run1 0.00% 1.22% 1.25%
K562_rep5_run1 0.00% 1.47% 1.50%
K562_rep6_run1 0.00% 1.45% 1.47%
Mus_ERR680375 0.00% 0.20% 0.26%
Mus_ERR680379 0.00% 0.15% 0.18%
Mus_ERR363657 0.00% 0.15% 0.21%
Mus_ERR363659 0.00% 0.10% 0.16%
HG002_GM24385 0.00% 0.33% 0.45%
HG002_GM26105 0.00% 0.29% 0.30%
HG002_GM27730 0.00% 0.31% 0.33%
HG002_SRR30901279 0.00% 1.21% 1.22%
cDNA
K562_rep1_run3 0.23% 1.84% 2.59%
K562_rep2_run3 0.28% 0.61% 1.28%
K562_rep3_run2 0.32% 0.24% 0.78%
K562_rep3_run3 0.38% 0.41% 1.12%
K562_rep3_run4 0.14% 0.73% 1.11%
Mus_ERR680377 0.12% 0.54% 1.22%
Mus_ERR680382 0.07% 0.39% 0.71%
Mus_ERR363658 0.09% 0.79% 1.49%
Mus_ERR363660 0.16% 0.70% 1.73%
HG002_NA24385 0.14% 0.71% 0.98%
HG002_NA26105 0.14% 0.56% 0.80%
HG002_NA27730 0.13% 0.59% 0.88%
HG002_PB_Iso 0.04% 2.65% 4.02%
HG002_PB_Mas 0.00% 1.02% 1.92%
gDNA
HG002_PB_HiFi 0.00% 0.14% 0.17%
HG002_70337_hac 0.72% 2.58% 3.52%
HG002_70337_sup 0.72% 2.60% 3.56%
HG002_71238_hac 0.65% 2.48% 3.33%
HG002_71238_sup 0.64% 2.49% 3.35%
Percent (%) of all reads
Sample
HCC1395
K562
Mouse
HG002
b
Foldback reads
K562 rep1
K562 rep2
K562 rep3
K562 rep4
HCC1395 rep1
% foldback out of all reads
Chimeric reads
% chimeric out of all reads
dRNA
cDNA
stranded-cDNA
direct-cDNA
Data type
c
Classification
Foldback
Chimera
True Break
Direct-cDNA
Direct-RNA
cDNA
gDNA
Percent (%) of split reads
d
Foldback
Chimeric
True Break
Prop. of reads with other read support
s=2
s=5
s=10
direct_cDNA
dRNA
cDNA
gDNA