

1 **Long read sequencing reveals a novel class of structural aberrations in cancers:**
2 **identification and characterization of cancerous local amplifications**

3

4 Yoshitaka Sakamoto¹, Liu Xu¹, Masahide Seki¹, Toshiyuki T. Yokoyama¹, Masahiro
5 Kasahara¹, Yukie Kashima^{2,3}, Akihiro Ohashi³, Yoko Shimada⁴, Noriko Motoi⁵, Katsuya
6 Tsuchihara², Susumu Kobayashi³, Takashi Kohno⁴, Yuichi Shiraishi⁶, Ayako Suzuki¹,
7 Yutaka Suzuki^{1*}

8

9 Affiliations:

10 ¹Department of Computational Biology and Medical Sciences, Graduate School of
11 Frontier Sciences, The University of Tokyo, Chiba, Japan.

12 ²Division of Translational Informatics, Exploratory Oncology Research and Clinical
13 Trial Center, National Cancer Center, Chiba, Japan.

14 ³Division of Translational Genomics, Exploratory Oncology Research and Clinical Trial
15 Center, National Cancer Center, Chiba, Japan.

16 ⁴Division of Genome Biology, National Cancer Center Research Institute, Tokyo, Japan.

17 ⁵Department of Pathology, National Cancer Center Hospital, Tokyo, Japan.

18 ⁶Division of Cellular Signaling, National Cancer Center Research Institute, Tokyo,
19 Japan.

20

21 *To whom correspondence should be addressed. Yutaka Suzuki: Tel: +81 4 7136 3607;
22 Fax: +81 4 7136 3607; Email: ysuzuki@hgc.jp.

23

24 Email address:

25 Yoshitaka Sakamoto (sakamoto_yoshitaka_18@stu.cbms.k.u-tokyo.ac.jp)

26 Liu Xu (xu_liu_18@stu.cbms.k.u-tokyo.ac.jp)

27 Masahide Seki (mseki@edu.k.u-tokyo.ac.jp)

28 Toshiyuki T. Yokoyama (yokoyama_toshiyuki_17@stu.cbms.k.u-tokyo.ac.jp)

29 Masahiro Kasahara (mkasa@edu.k.u-tokyo.ac.jp)

30 Yukie Kashima (y Kashima@east.ncc.go.jp)

31 Akihiro Ohashi (aohashi@east.ncc.go.jp)

32 Yoko Shimada (yoshimad@ncc.go.jp)

33 Noriko Motoi (nmotoi@ncc.go.jp)

34 Katsuya Tsuchihara (ktsuchih@east.ncc.go.jp)

35 Susumu Kobayashi (sukobaya@east.ncc.go.jp)

36 Takashi Kohno (tkkohno@ncc.go.jp)

37 Yuichi Shiraishi (yuishira@ncc.go.jp)

38 Ayako Suzuki (asuzuki@edu.k.u-tokyo.ac.jp)

39 Yutaka Suzuki (ysuzuki@hgc.jp)

40

41 Keywords:

42 Long read sequencing, Lung cancer, Local structural aberrations

43

44 Running title: Cancerous local copy-number lesions in lung cancers

45 **ABSTRACT**

46 Here we report identification of a new class of local structural aberrations in lung
47 cancers. The whole-genome sequencing of cell lines using a long read sequencer,
48 PromethION, demonstrated that typical cancerous mutations, such as point mutations,
49 large deletions and gene fusions can be detected also on this platform. Unexpectedly, we
50 revealed unique structural aberrations consisting of complex combinations of local
51 duplications, inversions and micro deletions. We further analyzed and found that these
52 mutations also occur *in vivo*, even in key cancer-related genes. These mutations may
53 elucidate the molecular etiology of patients for whom causative cancerous events and
54 therapeutic strategies remain elusive.

55 **INTRODUCTION**

56 Recent cancer sequencing projects, such as the International Cancer Genome
57 Consortium (ICGC) and The Cancer Genome Atlas (TCGA), have revealed causative
58 mutations in various types of cancers^{1,2}. Among them, lung adenocarcinomas are one of
59 the most well-studied cancers regarding such “cancer driver” mutations³⁻⁶. In lung
60 cancer patients, more than half of the cases have characteristic point mutations in the
61 *EGFR* and *KRAS* genes or gene fusions in the *ALK*, *RET* and *ROS1* genes. These
62 mutations are utilized as “biomarkers”, providing fundamental information about the
63 most appropriate therapeutic strategy. Patients are separated based on their genomic
64 mutation statuses and are matched to the most appropriate treatment⁷⁻¹¹. Despite this
65 general success, approximately 20-30 % of the lung adenocarcinoma patients remain
66 undiagnosed with respect to their cancerous mutations⁷.

67 Current information on the cancer mutations has been mostly obtained by
68 short read sequencing. Short read sequencing data, generally consisting of tens of
69 millions reads of up to 200-300 bases in length¹², are the most powerful in detecting
70 point mutations such as single nucleotide variants (SNVs) and short indels^{5,13}.
71 Significant efforts have been made to enable the identification of fusion genes by short
72 read sequences¹⁴. However, it is still difficult to detect more complex or larger-scale
73 structural aberrations, such as chromosome aneuploidy, copy number aberrations and
74 rearrangements solely based on short read sequencing data. There are inherent
75 drawbacks even in the latest bioinformatics pipelines for this purpose, which hampers
76 their practical use without careful validation¹⁵.

77 Recently developed long read sequencing technologies are changing this
78 situation. Several pioneering papers have reported the precise analysis of complicated
79 genomic regions, and large-range aberration detection is enabled by the long read
80 sequencing. For example, a single molecule real-time (SMRT) sequencer, PacBio RS, has
81 been utilized to analyze *BCR-ABL1* rearranged transcripts and their TKI-resistant
82 mutations in chronic myeloid leukemia (CML)^{16,17}. In Chromophobe Renal Cell
83 Carcinoma (ChRCC), structural alterations in the *TERT* promoter region have been
84 identified and characterized by PacBio sequencing¹⁸. Very recently, a nanopore-type
85 sequencer, MinION, was first utilized to characterize the pathogenic sequence
86 expansion of intronic repeats in Benign Adult Familial Myoclonic Epilepsy (BAFME)¹⁹.
87 Particularly for cancer applications, we and others have shown that cancer-associated
88 structural variants (SVs) could be detected by nanopore sequencing approaches^{20,21}.
89 Additionally, transcriptome sequencing by MinION has been shown to provide a
90 powerful analytical platform, where the complete splicing pattern of a given mRNA can

91 be thoroughly represented by a single read²². Further improvements in MinION
92 sequencing have been achieved by the parallelization of the nanopores in a given flow
93 cell, a platform named PromethION. PromethION can now produce over 100 Gb reads
94 per flow cell.

95 In this study, we attempted long read sequencing of whole human cancer
96 genomes using PromethION. We first demonstrate that PromethION sequencing can
97 identify point mutations as well as large structural aberrations and fusion genes
98 relatively easily. Moreover, we unexpectedly identified that mutations containing
99 complex combinations of small and middle-sized structural aberrations are quite
100 common, constituting a previously undefined unique class of mutations. Hereafter, we
101 will call those mutations Cancerous Local Copy-number Lesions (CLCLs). These CLCLs
102 resided even within the key cancer genes or drug target genes, such as the *STK11*, *NF1*
103 *SMARCA4* and *PTEN* genes. Additionally, taking advantage of long read sequencing,
104 we characterized the full-length transcript structures by full-length cDNA sequencing of
105 the transcriptome.

106 We initially used lung cancer cell lines for which we had previously collected
107 detailed information on multi-omics features, such as whole-genome sequencing,
108 RNA-seq, and ChIP-seq of Illumina reads²³. Then, we used clinical samples to
109 demonstrate that those CLCLs are not restricted to cell lines.

110 RESULTS

111 *Long read sequencing of cancer cell lines*

112 We conducted long read and whole-genome sequencing analysis using the nanopore type
113 sequencers, MinION and its latest high-throughput derivative, PromethION. We first
114 validated the performance of the new PromethION instrument by sequencing the
115 genome of LC2/ad, which is a lung cancer cell line derived from a Japanese lung
116 adenocarcinoma patient^{24,25} (**Fig. 1** and **Supplementary Table S1**). As a reference, we
117 collected the whole-genome sequencing data from a total of 33 MinION runs (R9.5 flow
118 cells) to cover the whole human genome at an overall sequencing depth of 31× by a total
119 of 7,282,846 reads (93,813,338,154 base pairs (bp)). The maximum length and N50
120 length of the reads were 2,495,160 bp and 30,606 bp, respectively. In total, 67.5 % of the
121 reads were mapped to the human reference genome UCSC hg38 using Minimap2²⁶. The
122 calculated overall sequence identity was 82 % on average. The average length of the
123 mapped reads was 16,452 bp, which was significantly longer than previous long read
124 whole human cancer genome sequencing analyses^{27–29} (**Table 1**). The PromethION
125 sequencing required approximately three flow cells to generate a total of 10,064,668
126 reads (100,440,433,160 bp) for an overall coverage of 33× (**Fig. 1**) (this number
127 decreased for subsequent cell types; see **Methods** for more details). The maximum
128 length and N50 length of reads were 987,834 bp and 32,710 bp, respectively. Using
129 Minimap2, 69.4 % of the reads were mapped to the reference genome. The average
130 length of the mapped reads was 13,620 bp, and the average identity was 85 % (**Table 1**).
131 Notably, because sample preparation need not be performed for each run, the required
132 total amount of starting DNA used for PromethION could be reduced by more than
133 tenfold compared with MinION.

134 To examine whether PromethION sequencing was compatible with MinION
135 sequencing, we compared the features of the obtained two datasets. The overall
136 distribution of read lengths was similar (**Fig. 1A**). Both datasets included a substantial
137 fraction of long reads over 50 kb (MinION: 360,786 reads, PromethION: 451,698 reads).
138 Detailed analysis of the mapping results showed that more than 50 % of the human
139 genome region was covered at more than 20× sequencing depth in both of the datasets
140 (**Fig. 1B**). For the sequence accuracy, both datasets showed an overall fidelity of more
141 than 80 % (**Fig. 1C**), which is similar to that of a previous study³⁰. We concluded that
142 PromethION should be an effective analytical method for whole cancer genome
143 sequencing.

144 Having finished the initial evaluation of the data obtained from MinION and
145 PromethION, we scaled the MinION and PromethION sequencing for an additional four

146 lung cancer cell lines (A549, RERF-LC-KJ, RERF-LC-MS and PC-14; detailed cellular
147 profiles are described **Supplementary Table S1**). The data production proceeded
148 similarly to the case of LC2/ad cells, for example, in the sequencing of RERF-LC-KJ,
149 5,986,875 reads were generated (57,062,227,853 bp, at 18.5×), with the max and N50
150 lengths of the reads being 922,768 bp and 23,442 bp, respectively. Other detailed
151 statistics are shown in **Supplementary Figure S1** and **Supplementary Table S2**.

152 To evaluate the quality of the sequence data at the individual base level, we
153 examined the known driver mutations of the corresponding cells by manually reviewing
154 the mapping results with Integrative Genomics Viewer (IGV)^{31,32}. In A549, eleven reads
155 illustrated the cancerous mutation *KRAS* G12S as the point mutation (left, **Fig. 1D**). In
156 PC-14, eight reads represented the driver *NRAS* Q61K point mutation (right, **Fig. 1D**).
157 Conversely, we also confirmed the absence of any driver mutations in the RERF-LC-KJ
158 and RERF-LC-MS cell lines at well-known driver genes. All of these results are
159 consistent with those of previous reports²³. These results collectively indicated that
160 mutation calling at the single-base level is also possible using only the long read
161 sequencer, at least when the cancer cell contents are as high as in the cultured cells.

162

163 *Identification of large-scale genomic aberrations*

164 Using the long read sequencing data, we then attempted to detect structural
165 aberrations larger than point mutations (**Fig. 2A**). From the MinION/PromethION
166 dataset of LC2/ad, we successfully identified 12 reads directly overlapping the junction
167 point of the *CCDC6-RET* fusion gene, which is the known “cancer driver mutation” for
168 this cell line^{24,25} (**Fig. 2B**; for details of the bioinformatics pipeline, see **Methods**). We
169 further attempted to identify large deletions. A large deletion around the *CDKN2A* gene,
170 which is a well-known tumor suppressor gene³, was previously reported to occur in
171 LC2/ad, A549 and PC-14 cells²³ (**Fig. 2C**). Using the MinION/PromethION datasets in
172 this study, we re-confirmed the deletion of this gene in the respective cells. In addition,
173 we found that the precise junction point of each of the *CDKN2A* deletions was different
174 between the cell types. Large deletions in other cancer-related genes are described in
175 **Supplementary Figure S2**.

176 We could also detect novel gene fusions by employing the split alignment
177 method (see **Methods**). We identified three novel rearrangements, which were further
178 validated by the Illumina short reads (**Fig. 2A** and **Supplementary Fig. S3**). These genes
179 were fused to *NELL1-CCSER1* and *EFNA5-IKBKB* in LC2/ad and *UTS2B-GRM4* in
180 RERF-LC-KJ. In each of these cases, the long read sequencing precisely identified the
181 junction at single-base resolution (**Supplementary Table S3**).

182 We further attempted to decipher perhaps the most difficult case, the
183 rearrangement of the *MYC* gene. We identified copy number aberrations of the *MYC*
184 gene in LC2/ad²³. The amplification was estimated to extend over approximately an 8
185 Mb locus having the *MYC* gene at the center. Even using long read sequencing, it was
186 still difficult to completely reconstruct its structure, which included complex rearranged
187 patterns, expanding to 8 Mb in chromosome 8 at an estimated aneuploidy of eight (**Fig.**
188 **2D**). Particularly for the *MYC* region, we attempted to identify the correct structure by
189 the optical mapping method, Bionano Saphyr. Even using the Saphyr, the precise
190 structure of the *MYC* region remained elusive, though the results from this analysis
191 support the *MYC* amplification spanning the 8 Mb region with approximately 8 copies
192 (**Fig. 2E**).

193

194 *Identification of a new class of cancerous local genomic lesions, CLCL*

195 During the attempts to identify the above structural aberrations of the established
196 classes, we unexpectedly found a new type of local structural aberration (**Fig. 3**). These
197 aberrations consisted of complex combinations of copy number changes, inversions and
198 deletions. As it appears that these aberrations do not precisely belong to the above
199 categories, we named them Cancerous Local Copy-number Lesions (CLCLs). As we will
200 describe below, we found it difficult to identify and characterize these CLCLs regarding
201 their precise junctions solely based on short read sequencing, even though some
202 suggestive data could be occasionally obtained.

203 The first example was found in the *STK11* gene locus. In our previous study of
204 lung cancer whole-genome sequencing using Illumina, we noticed a possible local
205 copy-number lesion in the *STK11* gene region in RERF-LC-KJ cells. The sequencing
206 depth increased from the middle of intron 1 to the end of the gene²³. There were short
207 read split tags (see Methods for details), suggesting that the inversions may occur in
208 this region. Despite the substantial number of sequencing reads mapped in this region,
209 we could not reconstruct its precise structure.

210 We examined the long reads to decipher the aberration in the *STK11* gene locus
211 (**Fig. 3A**). It revealed the aberration as follows: The first rearrangement occurred as an
212 inversion starting from intron 1 (chr19: 1,216,572; breakpoint II) and jumping
213 downstream of the gene (chr19: 1,228,569; breakpoint IV). The inverted sequence
214 continued back to the middle of the intron 1 (chr19: 1,216,360; breakpoint I), which was
215 212 bases upstream of the initial breakpoint II. Then, the sequence reverted back and
216 jumped to intron 3 (chr19: 1,219,538; breakpoint III). The following sequence continued
217 to the end of the gene locus. The detected junctions, breakpoints II/IV and I/III, were

218 represented by seven and nine PromethION reads, respectively. When we re-examined
219 the Illumina reads, the sequencing depth increased at the two regions, between
220 breakpoints I and II and between breakpoints III and IV (boxed region in **Fig. 3A**). We
221 also looked for the short reads using the soft-clipped method. We found that it was
222 difficult to detect two of the breakpoints, I and III, using the short read split tags, partly
223 because the junctions were resided in the repetitive regions.

224

225 *Identification of CLCLs in other genes and cell lines*

226 To more generally identify CLCLs in other loci in all lung cancer cell lines, we
227 constructed a new analytical bioinformatics pipeline (see **Supplementary Fig. S4** and
228 **Methods**). Briefly, we utilized the information of the split alignments from the mapping
229 results. We sorted the mapping information by the position of the reads and extracted
230 the CLCL candidates. The associated reads were reassembled to reconstruct their
231 structures.

232 As a result, we successfully identified the following numbers of CLCLs in the
233 other cell lines as well: sixteen in LC2/ad, one in A549, seven in RERF-LC-KJ, seven in
234 RERF-LC-MS, and eleven in PC-14 (**Table 2**). Importantly, CLCLs were found to occur
235 even in key cancer genes, such as the *STK11*, *NF1*, *SMARCA4* and *PTEN* genes. The
236 aberrant structures varied, and most of them would not be easily detected by the
237 conventional short-read-based approaches because of their complex structures and the
238 size of the affected regions. A relatively simple one was that which was detected in the
239 *NF1* gene in RERF-LC-MS cells (**Fig. 3B**). This was a tandem duplication of the region
240 between intron 9 (chr17:31,200,948) and the downstream region of the last exon 36
241 (chr17: 31,278,880; it was supported by six reads at the junction). In another case, the
242 structure of the *SMARCA4* CLCL showed a duplication from intron 1 (chr19:10,973,314)
243 to intron 20 (chr19: 11,022,573; supported by eight reads at the junction; **Fig. 3C**). A
244 more complex case was found in the structure of *PTEN* in PC-14. This CLCL was found
245 to be a combination of inversion and deletion (**Fig. 3D**). In these relatively simple cases,
246 remapping of the Illumina short reads to the discovered junctions validated the precise
247 identification of the reconstructed structure.

248 Indeed, although the presence of these mutations was partly suspected in a
249 previous study²³, their precise structures remained elusive before this study. We and
250 others had previously suspected the presence of large deletions, frameshift indels and
251 splice site mutations based on short read sequencing for those cases. However, by
252 conventional aberration detection based on the short reads, we could not detect some
253 cases, which were first identified as CLCLs in this study (indicated by black dots in **Fig.**

254 **3E).**

255 We also examined the genomic context of the CLCLs. In total, 64 % (28/44) of
256 the CLCLs had at least one junction overlapping with a long interspersed nuclear
257 element (LINE), short interspersed nuclear element (SINE) or long terminal repeat
258 (LTR), and 13 %, 24 % and 4 % (12/92, 22/92, and 4/92, respectively) of the junctions of
259 the CLCLs were in a LINE, SINE or LTR, respectively (**Fig. 3F**). It is possible that their
260 unique locations may hamper the precise identification of CLCLs by short read
261 sequencing.

262

263 *Aberrant transcriptional events associated with CLCLs*

264 After the new CLCL-type aberrations were identified in a number of key genes in a
265 number of cell types, the immediately raised question was in what manner they have
266 transcriptional or epigenomic consequences.

267 To characterize how the CLCL aberrations are reflected in the transcriptomes,
268 we newly generated and analyzed full-length cDNA sequencing data using MinION. We
269 also utilized the previous Illumina short read RNA-seq and ChIP-seq data. In
270 RERF-LC-KJ cells, short read sequences indicated that the *STK11* transcript is
271 abnormally spliced at intron 1 and that transcription jumped just before the CLCL
272 structure²³. MinION reads representing the full-length transcripts further specified the
273 precise splice pattern and the transcription termination sites (**Fig. 4A**). For almost all of
274 the transcripts, the first splicing occurred at the abnormal position (from
275 chr19:1,216,268) and transcription occurred according to the CLCL structure (RNA-seq
276 reads covered breakpoints II-IV from chr19:1,216,572 to chr19:1,228,569). Some
277 aberrant transcription was also observed within the downstream CLCL region (middle
278 panel, **Fig. 4A**). Such an aberrant transcription pattern was not observed in PC-14 cells,
279 where the *STK11* gene is wild-type (lower panel, **Fig. 4A**).

280 We examined the epigenome marks in the regions surrounding the CLCL as
281 represented by the ChIP-seq of H3K4me3, H3K9/14ac and RNA polymerase II. We
282 found that chromatin normally formed the active structure at the promoter regions and
283 that transcription was initiated normally at the correct position regardless of whether
284 the cell line harbored the CLCL or wild-type *STK11* locus (**Fig. 4B**). However, in only the
285 RERF-LC-KJ cells harboring the CLCL, the H3K36me3 mark disappeared in the
286 middle of intron 1, indicating that transcriptional elongation should be disrupted
287 exactly where the CLCL started. Illumina RNA-seq data also supported that the RNAs
288 were abnormally spliced in the middle of intron 1 and transcribed according to the
289 CLCL structure. The expression levels of these aberrant transcripts were measured as

290 2.8 rpkm. No normal transcripts were detected. However, the aberrant transcripts
291 retained a substantial expression level, although somewhat lower than that of the wild
292 type.

293 We conducted a similar analysis for the other CLCLs. For the *PTEN* gene in
294 PC-14 (**Fig. 4C**), the CLCL resided at exon 6. As a result, this exon was completely
295 skipped from the transcripts of *PTEN*. Accordingly, the resulting transcript should be
296 frame-shifted and thus should be likely to lead to functional loss of the *PTEN* gene. We
297 also examined the RNA expression levels in the *STK11*, *NF1*, *SMARCA4* and *PTEN*
298 genes harboring CLCLs based on the Illumina RNA-seq data. The results indicated that
299 CLCLs are generally likely to result in reduced gene expression levels (**Fig. 4D**).
300 Nevertheless, in some cases, gene expression levels remained significant, such as the
301 *NF1* transcripts in RERF-LC-MS cells and the *PTEN* transcripts in PC-14 cells.

302 To address the biological significance of the CLCLs, we examined how the
303 CLCL-affected locus invokes changes in protein expression levels and their related
304 signaling pathways. We conducted Western blotting analysis. As expected, we found
305 that the proteins of STK11, NF1, SMARCA4, and PTEN were completely lost in cells
306 harboring CLCLs in these genes (**Fig. 4E**). We further examined the activation status of
307 the downstream proteins. The expected disruptions of the pathways were observed for
308 all of the examined cases. PTEN suppresses the phosphorylation of AKT, and
309 phosphorylated AKT (phospho-AKT) consequentially activates the mTOR signaling
310 pathway³³. Aberrant upregulation of phospho-AKT was observed, reflecting the
311 functional loss of PTEN in PC-14 cells (PTEN-CLCL). AMPK is a gene that plays an
312 important role in maintaining cellular homeostasis, and the phosphorylation of the
313 AMPK protein at its alpha subunit is activated by STK11³⁴. Its activation is impaired in
314 RERF-LC-KJ cells (STK11-CLCL). The NF1 gene, which is a negative regulator of
315 RAS³⁵. Phospho-ERK, which is downstream of the RAS signaling pathway³⁶, was
316 aberrantly upregulated in RERF-LC-MS cells (NF1-CLCL). Interestingly, despite the
317 clear protein losses of the corresponding genes in all of the cases, either by conventional
318 aberrations or CLCLs, their consequences somewhat varied depending on the cases. For
319 example, even though the STK11 protein similarly disappeared in both RERF-LC-MS
320 cells (STK11-loss) and RERF-LC-KJ cells (STK11-CLCL), the enhanced ratio of
321 phospho-AMPK α was higher in the RERF-LC-KJ cells. The effects of NF1 in
322 RERF-LC-OK (NF1-loss) were almost undetectable, while the effects were significant in
323 RERF-LC-MS cells (NF1-CLCL). It is possible that other pathways can sometimes
324 complement the loss of the key protein.

325

326 *Identification of CLCLs in clinical lung cancer specimens*

327 To examine whether CLCLs are also present in clinical cancer lung adenocarcinoma
328 cases, we conducted similar PromethION whole-genome sequencing for the surgical
329 specimens of nine Japanese lung adenocarcinoma patients (**Table 3** and **Supplementary**
330 **Table S4**). The detected driver mutations for each patient are shown in **Table 3**. For
331 these cases, we generated 43,953,136,203 bp sequences on average for each case (more
332 than 10× depth; **Supplementary Table S5**). For case S10, we also sequenced normal
333 counterparts to eliminate possible normal variations and dubious CLCLs derived from
334 the mapping errors.

335 Here, again, we successfully detected CLCLs. To our surprise, six of the nine
336 specimens harbored at least one CLCL in their tumor genomes. Again, several key
337 cancer genes were included. For example, we identified an *RNF20* CLCL in case S8.
338 This patient is a female patient and had been shown to have an *EGFR* exon 19 deletion
339 as a driver mutation. However, the other cancerous mutations remained elusive. In this
340 case, the CLCL of the *RNF20* gene occurred as a tandem duplication between intron 2
341 (chr9: 101,536,324) and intron 6 (chr9: 101,544,752, **Supplementary Fig. S5**), which is
342 very likely to lead to the functional loss of this gene. The *RNF20* gene encodes an E3
343 ubiquitin ligase with a tumor suppressor function, and it is frequently mutated,
344 particularly in lung cancer³⁷. In the other cases, the indications obtained for the
345 molecular etiology underlying the carcinogenesis of the patients are summarized in
346 **Table 3**. Further scaling the long read sequencing would be needed to more precisely
347 identify the frequencies of the CLCLs and the preference of the genes harboring CLCLs.
348

349 *Re-evaluation of possible CLCLs based on public short read sequencing data*

350 As the first step for scaling CLCL analysis, we attempted to utilize pre-existing
351 Illumina short read data. We hoped that we might be able to identify CLCL candidates
352 even from the short read sequence data. If this is possible, there are tens of thousands of
353 whole-genome/exome sequencing datasets publicly available for the various cancer
354 types. We were also interested in how these CLCLs had been represented by the
355 previous short read sequences.

356 To identify putative CLCLs starting from the short read sequences, we
357 employed a soft-clipping program, GenomonSV
358 (<https://github.com/Genomon-Project/GenomonSV>)³⁸. We selected the “split” reads as
359 the “soft-clipped” reads and the paired-end reads, which may span the junctions of the
360 SVs (see **Methods** for details). As the model dataset, we first analyzed the whole-genome
361 short read data obtained for the five lung cancer cell lines that were used for the above

362 PromethION sequencing. We could extract an average of 182 “soft-clipped” junction
363 points in genic regions for each cell line (**Supplementary Fig. S6**). We defined tandem
364 duplication structures as putative CLCLs in the short read data. An average of 26 genes
365 were affected with putative CLCLs in the cell lines. We compared CLCLs detected from
366 short reads with those from long reads (**Supplementary Table S6**). Among the CLCLs
367 detected by PromethION, 72 % of the genes were also detected from the short read
368 sequence data (**Fig. 5A**). However, the precision rates were limited to 25 % because of
369 the general high rates of false-positive detection.

370 We then collected and analyzed the whole-genome short read data at a
371 sequencing depth of approximately 63× for nine clinical cases, as shown in **Table 3**. The
372 obtained results were inspected regarding the possible occurrence of CLCLs. We
373 identified an average of nine genes that may be affected by putative CLCLs. As shown
374 in **Figure 5B**, CLCLs were detected starting from the short read data at an estimated
375 sensitivity of 73 %. However, the precision rate was supposed to be limited to 14 % for a
376 variety of reasons inherent to the shorter read sequencing.

377 Despite the limited estimated precision and recall rates of CLCL detection
378 using short read data at 21 % and 72 %, respectively, for all the cell lines and clinical
379 samples taken together, we applied the constructed analytical pipeline to the
380 whole-exome sequencing data of 514 TCGA lung adenocarcinoma (TCGA-LUAD)³ and
381 97 Japanese lung adenocarcinoma (Japanese LUAD)⁵ samples. We expected that the
382 detection rate would be inherently lower, reflecting the fact that they are exome
383 sequencing datasets. We detected a total of 269 and 50 junction points with tandem
384 duplication structures, which are likely to correspond to the CLCLs, by soft-clipped
385 reads from TCGA-LUAD and Japanese LUAD cases, respectively (ranging from 1 to 29
386 genes per case). In total, we extracted CLCL candidates from 155 (30 %) TCGA-LUAD
387 and 39 (40 %) Japanese LUAD cases (**Fig. 5C**).

388 In particular, we considered whether there were any suspected cases harboring
389 the CLCL candidates for the 299 genes that are considered to be the most relevant
390 “cancer-associated genes”³⁹. We detected 16 cases (2.6 % in 514 + 97 cases), harboring
391 potential CLCLs in 17 genes (**Fig. 5D**). Interestingly, nine of these cases harbored no
392 known driver mutations. For example, in the case of TCGA-49-4512 (female,
393 nonsmoker), we identified a potential CLCL in the kinase domain of the *EGFR* gene.
394 This duplication was previously reported⁴⁰ and might cause aberrant activation of
395 EGFR, thus serving as a driver mutation of this case. Importantly, this patient’s
396 therapeutic target should be addressed by EGFR inhibitors such as afatinib⁴⁰. Putative
397 CLCLs associated with *ERBB2* were also detected in two other cases (both male and

398 smoked). Aberrant duplications seemed to occur between the *ERBB2* genic region and
399 the downstream intergenic or genic regions. Other patients were found to harbor
400 putative CLCLs in other important tumor suppressor genes, such as *STK11* and
401 *PBRM1*, for which mutation statuses could be utilized as putative markers for immune
402 checkpoint inhibitors^{41–43}. For these cases, the precise structure as well as the
403 functional relevance of the putative CLCLs are still unknown, and thus, they should be
404 subjected to detailed long read sequencing analyses.

405

406

407 DISCUSSION

408 In this paper, we have described the identification and characterization of structural
409 aberrations in lung cancer genomes using PromethION.

410 We were able to identify the precise junctions of chromosomal rearrangements
411 and large-scale deletions relatively easily. For example, the junction points of the
412 *CDKN2A* gene were precisely detected (**Fig. 2C**). In most cases, the proximal genes were
413 simultaneously deleted. In LC2/ad, the deletion spanned from the *MIR31HG* and *MTAP*
414 gene loci to the *DMRTA1* gene locus (supported by 9 reads). In A549 cells, the deletion
415 started from the *MTAP* locus and reached the *CDKN2B* gene locus (supported by 15
416 reads). In PC-14, 22 genes were deleted in addition to the *CDKN2A* gene (supported by
417 8 reads). Several studies have reported that *CDKN2A* codeleted genes are involved in
418 the hidden molecular features of cancers. The *MTAP* gene encodes
419 5-methylthioadenosine phosphorylase, which is associated with the purine and
420 methionine salvage pathways, located to the adjacent region of the *CDKN2A* gene and
421 frequently codeleted in cancers. MTAP-deficient cancers are known to acquire
422 vulnerability to arginine methyltransferase PRMT5 depletion, which may be a novel
423 target of an anticancer drug^{44–46}. It is important to determine the precise junction by the
424 long read approach to completely understand what genes or regions are affected by
425 these genomic aberrations.

426 We also identified the 8 Mb amplification for the *MYC* gene locus in LC2/ad
427 cells. We attempted to further characterize it and found that even the latest long read
428 sequencing technologies could not comprehend the precise mutation pattern in this
429 locus. Eight Mb may have been too large, and the internal rearrangement may have
430 been too complicated, although this is the only region where we could not reassemble
431 the structure. Interestingly, we found no sequences suggesting aberrations occurring
432 within the internal region of the *MYC* gene locus itself. Outside of the *MYC* gene, at
433 least four break points were detected by nanopore reads, which were further confirmed

434 by the Illumina short reads (**Fig. 2D**). There may be a unique selective pressure exerted
435 on this gene specifically, retaining the gene function itself intact, at the same time,
436 enhancing its gene expression.

437 Most importantly, in this study, we unexpectedly identified a unique aberration
438 pattern, the CLCL. We found that CLCLs exist even in pivotal cancer-related genes,
439 such as the *STK11*, *NF1*, *SMARCA4*, and *PTEN* genes. Recent papers have reported
440 that immune checkpoint inhibitors are less effective for lung cancers with *STK11*
441 mutations^{41,42}. Therefore, the therapeutic strategy for each patient would be different
442 depending on whether there is a mutation in the *STK11* gene or not. Additionally,
443 CLCLs were identified in the *PTPN13*, *RPTOR*, and *RHEB* genes in LC2/ad cells (**Table**
444 **2**). *PTPN13* encodes a protein tyrosine phosphatase. The *RPTOR* and *RHEB* genes are
445 members of the mTOR signaling pathway. The functional loss of these genes should be
446 related to tumorigenesis and malignancy of the cancer, although further studies will be
447 needed to clarify the relationship between those aberrations and the molecular etiology
448 of the cancers in more detail.

449 We could also analyze the causes and consequences of genomic SVs. In total,
450 67 % of the CLCLs had at least one junction in LINE, SINE and LTR regions,
451 suggesting that transposable elements were likely to contribute to the formation of
452 CLCLs. Using epigenome and transcriptome data, we also showed that CLCLs led to
453 the formation of abnormal transcripts and functional loss of their encoded proteins in
454 most cases.

455 We further conducted long read sequencing of clinical samples. We successfully
456 demonstrated that CLCLs occur in the *in vivo* genomes of lung adenocarcinoma patients.
457 In six cases, at least one CLCL was detected in the genes of important functions, giving
458 the complementary therapeutic indication for the patients. Finally, we reanalyzed short
459 read sequencing data from clinical samples that were previously published^{3,5},
460 particularly focusing on detecting CLCLs in cancer-related genes, such as the driver
461 genes of lung adenocarcinoma. Although the current precision rate is limited, the recall
462 rate was reasonably high. We believe it is important to subject those cases to further
463 detailed long read sequencing. We suggest that CLCLs occurring in cancer genomes
464 might have important roles in the phenotypic features of cancers, including responses to
465 anticancer drugs.

466 Lastly but not less importantly, we found that the visualization of the detected
467 mutation is also important. So-called “genome graph” databases should play an
468 indispensable role in representing the diverse nature of cancer genomes and thus
469 further enhance the accuracy of future genome analyses^{47,48}. This is the first study that

470 has utilized PromethION sequencing for cancer genomics. Obviously, further
471 improvements in the sequencing method itself, coupled with refinements of the
472 computational tools, are needed to reach further goals. Indeed, this study may have
473 presented more questions than answers. In that sense, this is only the first study
474 paving the way towards a more comprehensive understanding of the complicated
475 genomic aberrations of cancers and further in-depth study of their biology.

476 **MATERIAL AND METHODS**

477 *Cell lines and clinical samples*

478 The Lung adenocarcinoma cell lines LC2/ad, A549, RERF-LC-KJ, RERF-LC-MS, and
479 PC-14 were cultured as previously described²³. Cell pellets were washed with cold PBS
480 and cryopreserved.

481 Clinical samples were obtained with the appropriated informed consent at the
482 National Cancer Center Japan. Surgical specimens from 10 patients were
483 pathologically checked, and one was removed because of low tumor content
484 (**Supplementary Fig. S7** and **Supplementary Table S4**). All nine patients were diagnosed
485 with primary lung cancer, including eight adenocarcinomas and one large cell
486 carcinoma (**Table 3**). Fresh frozen surgical specimens were used to extract genomic DNA
487 (gDNA) and total RNA as described below.

488

489 *Whole-genome sequencing using MinION*

490 High-molecular-weight (HMW) gDNA was extracted from the lung cancer cell lines
491 LC2/ad and A549 with Smart DNA prep(a) kit (Analytikjena). In the case of LC2/ad,
492 WGS data were produced from 1D sequencing (SQK-LSK108), 1D² sequencing
493 (SQK-LSK308), rapid sequencing (RAD003), and in the case of A549, WGS data were
494 produced from only 1D² sequencing. In summary, 4 µg HMW gDNA was used for 1D
495 sequencing and DNA repair, end-prep, and adapter ligation were conducted. DNA repair
496 was performed using NEBNext FFPE DNA Repair Mix (M6630, NEB). End-prep was
497 performed using NEBNext Ultra II End Repair/dA-Tailing Module (E7546L, NEB).
498 Adapter ligation was performed using NEBNext Blunt/TA Ligase Master Mix (M0367L,
499 NEB) and Ligation Sequencing Kit 1D (SQK-LSK108, Oxford Nanopore Technologies).
500 In summary, 1D² sequencing, 4 or 5 µg HMW gDNA was used as input, and DNA repair,
501 end-prep, first adapter ligation, and second adapter ligation were conducted. DNA
502 repair and end-prep were the same protocol as the 1D sequencing. First and second
503 adapter ligations were performed using NEBNext Blunt/TA Ligase Master Mix and
504 Ligation Sequencing Kit 1D² (SQK-LSK308, Oxford Nanopore Technologies). DNA
505 purifications in each step of 1D and 1D² sequencing were performed using Agencourt
506 AMPure XP (A63882, Beckman Coulter). In summary, 15 µl gDNA was used for rapid
507 sequencing, and a Rapid Sequencing Kit (RAD003) was used.

508

509 *Whole-genome sequencing using PromethION*

510 The HMW gDNA extraction method was the same as MinION sequencing for LC2/ad,
511 A549, and RERF-LC-MS. Forty-eight microliters of 1.5 or 2 µg gDNA plus nuclease free

512 water (NFW), 3.5 μ l of NEBNext FFPE DNA Repair Buffer, 2 μ l of NEBNext FFPE DNA
513 Repair Mix (M6630, NEB), 3.5 μ l of NEBNext Ultra II End Prep Reaction Buffer, and 3
514 μ l of NEBNext Ultra II End Prep Enzyme Mix (E7545L, NEB) were mixed gently in 1.5
515 ml Eppendorf tube. After spinning down, the sample was incubated at 20 °C for 5
516 minutes and 65 °C for 5 minutes. Then, 60 μ l of AMPure XP beads (A63882, Beckman
517 Coulter) was added to the tube and the tube was mixed by flicking. The sample was
518 incubated for 5 minutes at room temperature (R.T.) using a rotator mixer. After
519 spinning down, the tube was set on a magnetic stand, and the supernatant was pipetted
520 off. The beads were washed twice with 200 μ l of 70 % ethanol. The tube was spun down
521 and back on the magnet. Any residual ethanol was pipetted off, and the tube was dried
522 for 30 seconds. The tube was removed from the magnetic stand, the pellet was
523 resuspended in 61 μ l NFW, and the tube was incubated for 2 minutes. The tube was set
524 on a magnet, 60 μ l of the sample was used in the next step, and 1 μ l was used for quality
525 check by Qubit. 60 μ l of end-prepped DNA, 25 μ l of Ligation Buffer (LNB), 10 μ l of
526 NEBNext Quick T4 DNA Ligase (NEB, E6056S), and 5 μ l of Adapter Mix (AMX) were
527 mixed in a 1.5 ml Eppendorf tube. After spinning down, the sample was incubated at
528 R.T. for 10 minutes. Then, 40 μ l of AMPure XP beads was added to the tube and mixed
529 by flicking. The sample was incubated for 5 minutes at R.T. using a rotator mixer. After
530 spinning down, the tube was set on the magnetic stand. The supernatant was pipetted
531 off, and the beads were washed twice with 250 μ l of Fragment Buffer (LFB). The sample
532 was spun down, and the tube was replaced on the magnet. Any residual supernatant
533 was pipetted off, and the tube was dried for 30 seconds. The tube was removed from the
534 magnet, and the pellet was resuspended in 25 μ l of Elution Buffer (EB). The sample was
535 incubated at R.T. for 10 minutes. Next, 24 μ l of the sample was used in the next step,
536 and 1 μ l was used for quality check by Qubit. A total of 46 μ l of Flush Tether (FLT) was
537 added directly to the tube of PromethION Flush Buffer (PFB), and the solution was
538 mixed by pipetting (Priming Mix). Then, 800 μ l of Priming Mix was loaded onto the
539 PromethION flow cell. Next, 75 μ l of SQB, 51 μ l of LB, and 24 μ l of the DNA library
540 were mixed in a 1.5 ml Eppendorf tube in 5 minutes (Loading Library). Then, 200 μ l of
541 Priming Mix was loaded onto the flow cell, and 150 μ l of Loading Library was loaded
542 onto the flow cell. We started the PromethION run. The LNB, AMX, LFB, EB, FLT, PFB,
543 SQB, and LB are in Ligation Sequencing Kit 1D (SQK-LSK109, Oxford Nanopore
544 Technologies). From RERF-LC-KJ, PC-14, and lung adenocarcinoma clinical samples,
545 HMW gDNA was extracted with the MagAttract HMW DNA Kit (Qiagen). For the
546 LC2/ad cells, we performed five runs, and each throughput was 10.4 Gb, 9.2 Gb, 33.0 Gb,
547 28.5 Gb, and 19.3 Gb, respectively.

548

549 *Full-length transcriptome sequencing using MinION*

550 Full-length transcriptome analysis using MinION was performed as previously
551 described⁴⁹. RNA was extracted from lung cancer cell lines using the RNeasy Mini kit
552 (Qiagen). The extracted RNA was converted to cDNA using SMART-seq v4 Ultra Low
553 input RNA kit (Takara). Then, we used cDNA as input for 1D² MinION sequencing.

554

555 *Computational analysis of long read sequencing data*

556 MinION fast5 data were basecalled using albacore 2.0.2 and converted fastq files.
557 PromethION fast5 data were basecalled using guppy and converted fastq files. Our
558 MinION and PromethION data set were mapped to the human reference genome, hg38,
559 using Minimap2 (with the “-ax map-ont” option, 2.9-r720 version). MinION 1D²
560 sequencing outputs two types of fastq files, 1D and 1D². 1D means that reads were
561 generated using single strand information. 1D² reads integrate the double strand
562 information. There were some overlapping reads between 1D files and 1D² files.
563 Therefore, reads used as 1D² were removed from 1D files. In addition, there were some
564 overlapping reads in the 1D² files. These reads were removed from the 1D² files and
565 used as 1D reads.

566

567 *Detection of driver mutations for cell lines in long read data*

568 We detected known driver mutations by IGV. For the point mutations, we directly
569 explored the known positions of the mutations. For the driver mutation of LC2/ad cells,
570 the *CCDC6-RET* fusion gene, we explored the reads split-aligned to both *RET* and
571 *CCDC6* genes in the reference genome and extracted the alignments with SAM format
572 from IGV. Then, we extracted the information of split alignment (chromosome, position
573 of reference, read strand, and position of read) from the file and sorted the information
574 by the position of reads. We filtered out reads with small MAPQ (< 30) and counted the
575 number of supporting reads.

576

577 *Detection of structural variants from long read data*

578 To detect gene rearrangements and CLCLs, we used the information of split alignments.
579 First, we mapped sequencing data (fastq) to the human reference genome, hg38. Then,
580 we extracted reads with split alignment from mapping data (bam) using the command
581 “samtools view -f 2048”. We filtered out reads with multiple hits (flag: 256). Then, we
582 extracted the information of split alignments (chromosome, position of reference, read
583 strand, and position of read) from the file and sorted the information by the position of

584 reads. We filtered out reads with low MAPQ (< 30) from the dataset. We extracted
585 junction candidates of gene rearrangements and CLCLs (tandem duplications and
586 inversions) considering the position of reads (removing junctions with large differences
587 in read position, >300 bp), annotated the junctions by genes from DBKERO
588 (<http://kero.hgc.jp/>) and merged the junctions less than 50 bp from the junctions of
589 CLCL candidates. The threshold for the number of reads supporting the junctions was
590 four in the clinical samples and five in the cell lines. We removed junctions with less
591 than 2,000 bp between the junctions. Finally, we checked the structure of the candidates
592 for gene rearrangements and CLCLs by IGV and manually removed questionable
593 rearrangements and CLCLs. To detect deletions, we used the information of split
594 alignments and CIGAR strings in SAM format files. We extracted the information of
595 split alignments in the same way as the detection of gene rearrangements and CLCLs.
596 We extracted the CIGAR strings from reads with primary alignments using SAMtools
597 (flags: 0 or 16). Then, we detected deletions over 2,000 bp.

598

599 *Optical mapping using the Saphyr system*

600 Optical mapping analysis using the Saphyr system (Bionano Genomics) was performed
601 for LC2/ad cells. Briefly, HMW DNAs were isolated from frozen cells using a Bionano
602 Prep kit (Bionano Genomics) and measured by Qubit BR assay (Invitrogen). The
603 extracted DNAs were fluorescently labeled with DLE-1 using a Bionano DLS kit
604 (Bionano Genomics). Data were collected on the Saphyr instrument (Bionano Genomics).
605 The figures were created by Bionano Access (version 1.3.0, Bionano Genomics).

606

607 *Western blotting*

608 We performed Western blotting as described previously to quantify proteins from genes
609 with CLCL structures⁵⁰. For Western Blotting, we performed protein extraction and
610 quantification. We used a Pierce BCA Protein Assay kit (23225, Thermo Fisher
611 SCIENTIFIC) and prepared proteins of 1 mg/ml. In summary, we performed
612 electrophoresis of proteins, membrane transferring, blocking, reaction of the first
613 antibody, reaction of the second antibody, and visualization of bands. We conducted
614 electrophoresis using a tank and used 10 or 15 μ l of proteins as input. For blocking, we
615 used 4 % BSA blocking buffer. In the reaction of the first antibody, we used LKB1
616 (27D10) Rabbit mAb (3050S, CST) for STK11, phospho-AMPK- α (Thr172) Antibody
617 (2535S, CST) for phospho-AMPK α , AMPK- α Antibody Rabbit mAb (2603S, CST) for
618 AMPK α , Anti-Neurofibromin (NF1) (rabbit polyclonal IgG) (07-730, upstate) for NF1,
619 Phospho-p44/42 MAPK (Erk1/2) (Thr202/Thr204) (E10) Mouse mAb (9106S, CST) for

620 phospho-ERK, p44/42 MAPK (Erk1/2) Antibody (9102S, CST) for ERK, Brg1 (D1Q7F)
621 Rabbit mAb (49360, CST) for SMARCA4, PTEN (138G6) Rabbit mAb (CST, 9559S) for
622 *PTEN* protein, Phospho-Akt (Ser473) (D9E) XP[®] Rabbit mAb (4060S, CST) for
623 phospho-AKT, and Akt (pan) (11E7) Rabbit mAb (4685S, CST) for AKT. As a control, we
624 used GAPDH and GAPDH (14C10) Rabbit mAb (2118S, CST) as an anti-body. In the
625 reaction of the second antibody, we used an antibody corresponding to the animal that
626 generated the first antibody. We used Anti-rabbit IgG, HRP-linked antibody (7074S,
627 CST) for rabbit and Anti-mouse IgG, HRP-linked antibody (7076S, CST) for mouse. For
628 the visualization of bands, we used ImageQuant LAS 4000 mini (GE Healthcare). The
629 normalized ratio indicated the fraction of relative density of phosphorylated protein to
630 control protein and relative density of total protein to control protein, and we set the
631 normalized ratio of the wild type as one (**Fig. 4E**). The density of the bands was
632 calculated by ImageJ software⁵¹.

633

634 *Whole-genome short read sequencing of clinical samples*

635 Genomic DNA was extracted from surgical specimens using the MagAttract HMW DNA
636 Kit (QIAGEN). Whole-genome sequencing libraries were constructed using the TruSeq
637 Nano DNA Library Prep kit (Illumina) and sequenced by NovaSeq according to the
638 manufacturers' instructions. In summary, 100 ng gDNA was used for library
639 preparation as input, and DNA fragmentation, end repair, adenylation of the 3' end,
640 adapter ligation, and condensation of DNA fragments were conducted. We performed
641 DNA fragmentation using Covaris (Covaris) and used a protocol for 350 bp of insert size.
642 After preparing the library, we denatured the library with NaOH and then started the
643 NovaSeq run.

644

645 *Analysis of short read sequencing data*

646 Whole-genome short read sequences were mapped to the human reference genome
647 (hg38) using BWA-MEM (version 0.7.15). After mapping, sorted BAM files were created,
648 and PCR duplicates were marked by SAMtools. For detection of driver mutations in the
649 ten clinical samples, point mutations and variations were called using GATK Mutect2
650 (version 4.0.12.0).

651 For transcriptome and epigenome analysis of the cell lines, we used RNA-seq
652 and ChIP-seq data that were previously obtained (DRA001846 and DRA001860) and
653 mapped to the reference genome hg19. IGV was used for visualization.

654 To detect SV junctions, GenomonSV (version 2.6.1) was used with paired-end
655 read sequencing data as listed; 5 whole-genome datasets from cancer cell lines

656 (DRA001859; 101PE, Illumina HiSeq2500), nine whole-genome datasets from Japanese
657 lung cancer patients (150PE, Illumina NovaSeq), 514 whole-exome datasets from
658 TCGA-LUAD and 97 whole-exome datasets from Japanese lung adenocarcinoma
659 patients (JGAS0000000001; 76PE, Illumina GAIIX). After conducting Genomon
660 (version 2.6.1) with the recommended parameters
661 (<https://genomon.readthedocs.io/ja/latest/>; reference: hg19), GenomonSV filt was
662 performed with the options “--min_junc_num 1” and “--non_matched_control_junction”
663 with control panels that were constructed from ICGC/TCGA data. For clinical samples
664 harboring matched normal data, we also set the option “--matched_control_bam”. For
665 the cell lines, matched normal data of case S1 were used for normal control data. After
666 GenomonSV filt, we eliminated SV candidates with tumor VAF < 0.05 and a distance of
667 less than 2 kb between the junctions. We analyzed at least one SV junction within genic
668 regions. For validation of SVs, junction points were compared between long read and
669 short read data, allowing 100-bp margins after performing conversion of the reference
670 genome version (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>).

671 Genes with putative CLCLs are defined as those affected by SV junctions with
672 a tandem duplication structure. To evaluate the detection power of short read data,
673 genes with CLCL (and putative CLCLs) were classified into three categories: true
674 positives detected in both long read and short read data; false positives detected in short
675 read data only; and false negatives detected in long read data only. The precision and
676 recall rates of CLCL detection of short read data were calculated using the number of
677 true positives, false positives and false negatives. For CLCLs in cancer-related genes,
678 we checked 299 cancer-related genes that were previously reported as cancer driver
679 genes. The cancer driver genes were classified in gene types as follows: genes significant
680 in lung adenocarcinoma and pan-cancer types as “LUAD and multiple cancer types”,
681 genes significant in multiple and pan-cancer types as “Multiple cancer types”, and genes
682 significant in other cancer types than lung adenocarcinoma as “Unique to other cancer
683 types”³⁹. For known driver mutation status, 13 genes (*EGFR*, *KRAS*, *BRAF*, *HRAS*,
684 *NRAS*, *RET*, *MAP2K1*, *ALK*, *ROS1*, *ERBB2*, *MET*, *NF1* and *RIT1*) were examined in
685 the 16 cases by cBioPortal (Lung adenocarcinoma; TCGA; PanCancer Atlas)⁵²⁻⁵⁴.

686

687

688 DATA ACCESS

689 All sequencing data of cell lines were published in the DNA Data Bank of Japan (DDBJ)
690 under the accession numbers, DRA007423 (DRX143541, DRX143542, DRX143543,
691 DRX143544), DRA007941, DRA008154 and DRA008295. The data were also deposited

692 in DBKERO (<https://kero.hgc.jp>)⁵⁵. Sequencing data of clinical samples were deposited
693 at the Japanese Genotype-phenotype Archive (JGA, <http://trace.ddbj.nig.ac.jp/jga>),
694 which is hosted by the National Bioscience Database Center (NBDC) and DDBJ, under
695 the accession numbers, JGAS00000000065 (JGAD00000000252 and
696 JGAD00000000253).

697

698

699 **ACKNOWLEDGEMENTS**

700 We thank K. Imamura, K. Abe, M. Kimura, H. Wakaguri, Y. Kuze and T. Horiuchi for
701 their technical assistance. The Bionano data were generated at Bionano Genomics and
702 K. Hong and A. Pang collected and analyzed the data. This work was supported by
703 AMED P-CREATE Grant Number JP19cm0106539. This work was also supported in
704 part by The National Cancer Center Research and Development Fund (29-A-6). This
705 work was also supported by JSPS KAKENHI Grant Number 16H06279. The results
706 shown in this study are in part based on data generated by the TCGA Research
707 Network (<https://www.cancer.gov/tcga>). The super-computing resource was provided by
708 Human Genome Center, the University of Tokyo (<http://sc.hgc.jp/shirokane.html>).

709

710

711 **AUTHOR CONTRIBUTIONS**

712 Y. Sakamoto, A.S. and Y. Suzuki designed the study. Y. Sakamoto, L.X., and M.S.
713 performed sequencing experiments. Y. Sakamoto, T.K., M.K., Y. Shiraishi and A.S.
714 contributed computational analysis of sequencing data. Y. Sakamoto, Y.K., A.O. and S.K.
715 conducted Western blotting. Y. Shimada, N.M. and T.K. contributed and analyzed
716 clinical specimens. K.T., S.K., T.K. and Y. Suzuki interpreted the findings and
717 supervised the study. Y. Sakamoto, A.S. and Y. Suzuki wrote the manuscript. All
718 authors approved the final version of the manuscript.

719

720

721 **COMPETING INTERESTS**

722 The authors declare no competing interests.

723

724

725 **REFERENCES**

- 726 1. Hudson, T. J. *et al*. International network of cancer genome projects. *Nature* **464**,
727 993–998 (2010).

- 728 2. The Cancer Genome Atlas Research Network *et al* The Cancer Genome Atlas
729 Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
- 730 3. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of
731 lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- 732 4. Seo, J. S. *et al* The transcriptional landscape and mutational profile of lung
733 adenocarcinoma. *Genome Res.* **22**, 2109–2119 (2012).
- 734 5. Suzuki, A. *et al* Identification and Characterization of Cancer Mutations in Japanese
735 Lung Adenocarcinoma without Sequencing of Normal Tissue Counterparts. *PLoS*
736 *One* **8**, e73484 (2013).
- 737 6. Imielinski, M. *et al* Mapping the hallmarks of lung adenocarcinoma with massively
738 parallel sequencing. *Cell* **150**, 1107–1120 (2012).
- 739 7. Kohno, T. *et al* RET fusion gene: Translation to personalized lung cancer therapy.
740 *Cancer Sci.* **104**, 1396–1400 (2013).
- 741 8. Yoh, K. *et al* Vandetanib in patients with previously treated RET-rearranged
742 advanced non-small-cell lung cancer (LURET): an open-label, multicentre phase 2
743 trial. *Lancet Respir. Med.* **6**, 2–10 (2016).
- 744 9. Mainardi, S. *et al* SHP2 is required for growth of KRAS-mutant non-small-cell lung
745 cancer in vivo. *Nature Medicine* 1–7 (2018). doi:10.1038/s41591-018-0023-9
- 746 10. Ruess, D. A. *et al* Mutant KRAS-driven cancers depend on PTPN11/SHP2
747 phosphatase. *Nature Medicine* 1–7 (2018). doi:10.1038/s41591-018-0024-8
- 748 11. Wong, G. S. *et al* Targeting wild-type KRAS-amplified gastroesophageal cancer
749 through combined MEK and SHP2 inhibition. *Nature Medicine* 1–10 (2018).
750 doi:10.1038/s41591-018-0022-x
- 751 12. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing:
752 Computational challenges and solutions. *Nature Reviews Genetics* **13**, 36–46 (2012).
- 753 13. Pleasance, E. D. *et al* A comprehensive catalogue of somatic mutations from a
754 human cancer genome. *Nature* **463**, 191–196 (2009).
- 755 14. Liu, S. *et al* Comprehensive evaluation of fusion transcript detection algorithms and
756 a meta-caller to combine top performing methods in paired-end RNA-seq data.
757 *Nucleic Acids Res.* **44**, (2015).
- 758 15. Guan, P. & Sung, W. K. Structural variation detection using next-generation
759 sequencing data: A comparative technical review. *Methods* **102**, 36–49 (2016).
- 760 16. Ardui, S., Ameer, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time
761 (SMRT) sequencing comes of age: applications and utilities for medical diagnostics.
762 *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky066
- 763 17. Cavelier, L. *et al* Clonal distribution of BCR-ABL1 mutations and splice isoforms by

- 764 single-molecule long-read RNA sequencing. *BMC Cancer* **15**, 45 (2015).
- 765 18. Davis, C. F. *et al.* The somatic genomic landscape of chromophobe renal cell
766 carcinoma. *Cancer Cell* **26**, 319–330 (2014).
- 767 19. Ishiura, H. *et al.* Expansions of intronic TTTC A and TTTTA repeats in benign adult
768 familial myoclonic epilepsy. *Nat. Genet.* **50**, 581–590 (2018).
- 769 20. Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R. & Timp, W. Nanopore
770 sequencing detects structural variants in cancer. *Cancer Biol. Ther.* **17**, 246–253
771 (2016).
- 772 21. Suzuki, A. *et al.* Sequencing and phasing cancer mutations in lung cancers using a
773 long-read portable sequencer. *DNA Res.* **24**, 585–596 (2017).
- 774 22. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J.
775 Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and
776 qualitative assessment of cDNA populations. *Sci. Rep.* **6**, (2016).
- 777 23. Suzuki, A. *et al.* Aberrant transcriptional regulations in cancers: Genome,
778 transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic
779 Acids Res.* **42**, 13557–13572 (2014).
- 780 24. Suzuki, M. *et al.* Identification of a lung adenocarcinoma cell line with CCDC6-RET
781 fusion gene and the effect of RET inhibitors in vitro and in vivo. *Cancer Sci.* **104**,
782 896–903 (2013).
- 783 25. Matsubara, D. *et al.* Identification of CCDC6-RET fusion in the human lung
784 adenocarcinoma cell line, LC-2/ad. *J. Thorac. Oncol.* **7**, 1872–1876 (2012).
- 785 26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 1–7
786 (2018). doi:10.1101/169557.
- 787 27. Seo, J. S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature*
788 **538**, 243–247 (2016).
- 789 28. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat.
790 Commun.* **7**, (2016).
- 791 29. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, Nakamura M,
792 Nagasaki M, Kinoshita K, Okamura Y, Miyatake S, Miyake N, M. N. A 12-kb
793 structural variation in progressive myoclonic epilepsy was newly identified by
794 long-read whole-genome sequencing. *J. Hum. Genet.* (2019).
- 795 30. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with
796 ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
- 797 31. Robinson, J. T. *et al.* Integrative Genome Viewer. *Nat. Biotechnol.* **29**, 24–6 (2011).
- 798 32. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer
799 (IGV): High-performance genomics data visualization and exploration. *Brief*

- 800 *Bioinform.* **14**, 178–192 (2013).
- 801 33. Carracedo, A., Salmena, L. & Pandolfi, P. P. SnapShot: PTEN Signaling Pathways.
802 *Cell* **133**, 550–550.e1 (2008).
- 803 34. Mihaylova, M. M. & Shaw, R. J. The AMP-activated protein kinase (AMPK) signaling
804 pathway coordinates cell growth, autophagy, & metabolism. *Nat. Cell Biol.* **13**,
805 1016–1023 (2012).
- 806 35. Wolman, M. A. *et al* Modulation of cAMP and Ras Signaling Pathways Improves
807 Distinct Behavioral Deficits in a Zebrafish Model of Neurofibromatosis Type 1. *Cell*
808 *Rep.* **8**, 1265–1270 (2014).
- 809 36. De Luca, A., Maiello, M. R., D'Alessio, A., Pergameno, M. & Normanno, N. The
810 RAS/RAF/MEK/ERK and the PI3K/AKT signalling pathways: role in cancer
811 pathogenesis and implications for therapeutic approaches. *Expert Opin. Ther.*
812 *Targets* **16**, S17–S27 (2012).
- 813 37. Sethi, G., Shanmugam, M. K., Arfuso, F. & Kumar, A. P. Role of RNF20 in cancer
814 development and progression – a comprehensive review. *Biosci. Rep.* **38**,
815 BSR20171287 (2018).
- 816 38. Kataoka, K. *et al* Aberrant PD-L1 expression through 3'-UTR disruption in multiple
817 cancers. *Nature* **534**, 402–406 (2016).
- 818 39. Bailey, M. H. *et al* Comprehensive Characterization of Cancer Driver Genes and
819 Mutations. *Cell* **173**, 371–385.e18 (2018).
- 820 40. Gallant, J. N. *et al* EGFR kinase domain duplication (EGFR-KDD) is a novel
821 oncogenic driver in lung cancer that is clinically responsive to afatinib. *Cancer Discov.*
822 **5**, 1155–1163 (2015).
- 823 41. Rizvi, H. *et al* Molecular determinants of response to anti-programmed cell death
824 (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with
825 non-small-cell lung cancer profiled with targeted next-generation sequencing. *J. Clin.*
826 *Oncol.* **36**, 633–641 (2018).
- 827 42. Skoulidis, F. *et al* STK11/LKB1 mutations and PD-1 inhibitor resistance in
828 KRAS-mutant lung adenocarcinoma. *Cancer Discov.* **8**, 822–835 (2018).
- 829 43. Miao, D. *et al* Genomic correlates of response to immune checkpoint therapies in
830 clear cell renal cell carcinoma. *Science (80-)*. **359**, 801–806 (2018).
- 831 44. Kryukov, G. V. *et al* MTAP deletion confers enhanced dependency on the PRMT5
832 arginine methyltransferase in cancer cells. *Science (80-)*. **351**, 1214–1218 (2016).
- 833 45. Mavrakis, K. J. *et al* Disordered methionine metabolism in MTAP/CDKN2A-deleted
834 cancers leads to dependence on PRMT5. *Science (80-)*. **351**, 1208–1213 (2016).
- 835 46. Marjon, K. *et al* MTAP Deletions in Cancer Create Vulnerability to Targeting of the

- 836 MAT2A/PRMT5/RIOK1 Axis. *Cell Rep.* **15**, 574–587 (2016).
- 837 47. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing
838 genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–881 (2018).
- 839 48. Rakocevic, G. *et al.* Fast and Accurate Genomic Analyses using Genome Graphs. *Nat.*
840 *Genet.* 194530 (2019). doi:10.1101/194530
- 841 49. Seki, M. *et al.* Evaluation and application of RNA-Seq by MinION. *DNA Res.* (2019).
842 doi:10.1093/dnares/dsy038
- 843 50. Ohashi, A. *et al.* Aneuploidy generates proteotoxic stress and DNA damage
844 concurrently with p53-mediated post-mitotic apoptosis in SAC-impaired cells. *Nat.*
845 *Commun.* **6**, 1–16 (2015).
- 846 51. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of
847 Image Analysis HHS Public Access. *Nat. Methods* **9**, 671–675 (2012).
- 848 52. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles
849 using the cBioPortal. *Sci. Signal* **6**, (2013).
- 850 53. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring
851 multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–4 (2012).
- 852 54. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of
853 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).
- 854 55. Suzuki, A. *et al.* DBTSS/DBKERO for integrated analysis of transcriptional
855 regulation. *Nucleic Acids Res.* **46**, D229–D238 (2018).

856

857

858 **FIGURE LEGENDS**

859 **Figure 1 Long read sequencing of cancer genomes**

860 (A) Left: raw read length distribution of LC2/ad MinION sequencing. Right: raw read
861 length distribution of LC/2ad PromethION sequencing. Both MinION and PromethION
862 datasets have many long reads (for example, over 50 kb, 361 K reads and 452 K reads,
863 respectively). (B) Cumulative depth curve of LC2/ad. Blue: PromethION, Red: MinION.
864 More than 50 % of the human genome region was covered by more than 20× in
865 sequencing depth in both of the datasets. (C) Violin plots of identity of LC/2ad MinION
866 sequencing (left) and PromethION sequencing (right). The points in the violin plots
867 indicate the average identities of the MinION and PromethION data (82.1 %, 84.8 %,
868 respectively, **Table 1**). The identities were concentrated at more than 80 % in both
869 datasets. (D) IGV image of point mutations in the driver genes *KRAS* and *NRAS* in
870 MinION/PromethION sequencing (upper) and in Illumina sequencing (lower). *KRAS*
871 point mutation (G12S) is a driver mutation in A549 cells and *NRAS* point mutation

872 (Q61K) is a driver mutation in PC-14 cells.

873

874 **Figure 2 Long read sequencing reveals structural variants**

875 (A) Circos plot of novel rearrangement candidates in LC2/ad. The second layer of the
876 Circos plot indicates the sequencing depth by MinION. Detailed information on this
877 rearrangement is shown in **Supplementary Table S3**. (B) IGV image of the *CCDC6-RET*
878 fusion gene of LC2/ad MinION sequencing. The *CCDC6-RET* fusion gene is a driver
879 mutation of LC2/ad cells. The number of reads supporting the junction is twelve. (C)
880 IGV image of a large deletion including *CDKN2A* of LC2/ad, A549, and PC-14. The
881 deletion of LC2/ad spanned approximately 941 kb. The deletion of A549 spanned
882 approximately 296 kb. The deletion of PC-14 spanned approximately 3,438 kb. The
883 parentheses indicate the range of length of reads supporting the deletions. (D) Depth
884 plotting around the *MYC* gene of LC2/ad. Arrows indicate junction candidates of the
885 amplification supported by MinION and Illumina reads. (○): the number of supporting
886 MinION reads (left) and Illumina paired-end reads (right). (E) The 8-copy *MYC* region
887 of LC2/ad represented by the optical mapping method. Optical maps with
888 rearrangements and chromosome 8 (reference) are represented in light blue and
889 yellow-green, respectively. The orange arrow indicates the *MYC* gene.

890

891 **Figure 3 Identification and characterization of CLCL**

892 (A) Structure of the *STK11* CLCL in RERF-LC-KJ. The *STK11* CLCL was constructed
893 by a combination of local inversions. We can trace the CLCL structure following the
894 ordered arrows, and the junctions are indicated by colored arrows. The CLCL spanned
895 12 kb in the human reference genome. (B) Structure of the *NFI* CLCL in RERF-LC-MS.
896 The structure of the CLCL was a tandem duplication between the junctions (indicated
897 by a yellow arrow and a blue arrow). The CLCL spanned 78 kb in the reference genome.
898 (C) Structure of the *SMARCA4* CLCL in PC-14. The structure of the CLCL was a
899 tandem duplication of the junctions (indicated by a yellow arrow and a blue arrow). The
900 CLCL spanned 50 kb in the reference genome. (D) Structure of the *PTEN* CLCL in
901 PC-14. The structure of the CLCL was a combination of a local inversion and deletion.
902 We can trace the CLCL structure following the ordered arrows, and the junctions are
903 indicated by colored arrows. The CLCL spanned 7 kb in the reference genome. (E)
904 Summary of mutation types of four-cancer-related genes in five cell lines. (F) The
905 number of CLCL junctions in each category of genomic contexts.

906

907 **Figure 4 Aberrant transcriptional events caused by CLCL**

908 (A) Structures of *STK11* transcripts in RERF-LC-KJ. Sequencing tags of whole-genome
909 sequencing (PromethION) and full-length RNA-seq (MinION) were visualized by IGV.
910 PC-14 RNA-seq was also shown as a wild-type control. (B) Multi-layered statuses in the
911 *STK11* region. Patterns of whole-genome sequencing, ChIP-seq and RNA-seq tags of
912 short read data were visualized by IGV. The status of RERF-LC-KJ and PC-9 (control) is
913 shown. (C) Structures of *PTEN* transcripts in PC-14. Sequencing tags of whole-genome
914 sequencing (PromethION) and full-length RNA-seq (MinION) were visualized by IGV.
915 The transcripts indicated that exon 6 was skipped (black arrow). (D) Expression levels
916 of *STK11*, *NFI*, *SMARCA4* and *PTEN* in 26 lung cancer cell lines. Cell lines with
917 deleterious mutations, such as large deletions, frameshift indels and nonsense SNVs,
918 are shown in black. Cell lines with CLCLs are also shown in black with a diagonal line.
919 (E) Western blotting of genes affected by CLCLs and their downstream targets. MW:
920 molecular weight. WT: wild type. DEL: large deletion. Bar charts indicate the
921 normalized ratio of density of phosphorylated proteins and total proteins. Each protein
922 is downstream of proteins encoded by CLCL genes.

923

924 **Figure 5 CLCL as an overlooked cancerous aberration in cancers**

925 (A B) Detection of CLCLs in cell lines (A) and clinical samples (B) in short read data.
926 The number of genes with CLCLs detected by short read data are shown in each
927 category (TP: true positives, FN: false negatives, FP: false positives; y-axis on the left
928 side; also see **Supplementary Table S6**). The precision and recall rates are shown (y-axis
929 on the right side). The legend and color key are shown in the margin in A. (C) The
930 number of putative and expected CLCLs in whole-exome sequencing data. TCGA-LUAD
931 and 97 Japanese lung adenocarcinoma datasets are shown in the left and right panels.
932 (D) Putative CLCLs of cancer-related genes in 16 LUAD cases (shown as diamond). The
933 driver mutation status of each case is shown at the bottom (blue and red). Gene types
934 indicate driver genes that are significantly functionally altered in the indicated cancer
935 types.

936

937 **TABLES**

938 **Table 1 General statistics of nanopore sequencing in LC2/ad**

Categories	MinION	PromethION
Number of reads	7,282,846	10,064,668
N50 length of reads	30,606 bp	32,710 bp
Coverage	31.0 ×	33.1 ×
Percentage of mapped reads	67.5 %	69.4 %
Average length of mapped reads	16,452 bp	13,620 bp
Average identity of mapped reads	82.1 %	84.8 %

939

940

941

942 **Table 2 Summary of CLCLs detected in lung cancer cell lines**

Cell line	Number of CLCLs	Genes
LC2/ad	16	<i>AGO3, GABBR2, SART3, DHRS13, TAOK1, RPTOR, PTPN13, RHEB, SPAG1, POLB*</i> , <i>FERIL6, GRM8, PTPRD, EFNA2, CLEC18A</i>
A549	1	<i>RABGAP1L</i>
RERF-LC-KJ	7	<i>COL11A1, GTF2H1, LGALS3, STK11, CPT1C, IFNAR1, TWSG1</i>
RERF-LC-MS	7	<i>LCK, HDAC1, CLEC2D, NFI, SLC8A2, MYT1L, ACTB</i>
PC-14	11	<i>SLC44A5, PTEN, NAP1L1, INTS2, SMARCA4, PTPRD, PRR5, SERPINB9, F13A1, AUTS2, TMEM38B</i>

943 *LC2/ad cells had two CLCL structures in the *POLB* gene.

944 **Table 3 Clinical information and CLCL candidates of clinical samples**

945

Case	Age	Gender	Pathological stage	Histological characters	Driver mutations	Number of CLCLs	Genes with CLCLs
S1	76	Female	IA3	Large cell carcinoma	Not detected	0	-
S2	57	Female	IIB	Ad, micropapillary	<i>EGFR</i> L858R	2	<i>CSMD3, SLC30A8</i>
S3	54	Male	IIB	Ad, solid	<i>NRAS</i> Q61L	1	<i>SNX25</i>
S5	38	Male	IIIA	Ad, micropapillary	Not detected	0	-
S6	53	Female	IA3	Ad, papillary	<i>EGFR</i> exon 19 deletion	0	-
S7	77	Female	IA3	Ad, micropapillary	<i>EGFR</i> exon 20 insertion	2	<i>NLGN1, MEF2A</i>
S8	85	Female	IB	Ad, papillary	<i>EGFR</i> exon 19 deletion	2	<i>ASTN2, RNF20</i>
S9	71	Female	IA1	Ad, lepidic	<i>EGFR</i> L858R	1	<i>TSPAN8</i>
S10	65	Female	IA3	Ad, papillary	<i>EGFR</i> L858R	3	<i>AGBL4, ZFP82, ZNF681</i>

946 Ad: Adenocarcinoma

947 **SUPPLEMENTARY FIGURES AND TABLES**

948 **Supplementary Figure S1 General statistics of MinION and PromethION in four cell**
949 **lines**

950 **Supplementary Figure S2 IGV images of large deletions detected by MinION and**
951 **PromethION**

952 **Supplementary Figure S3 A novel genomic rearrangement in RERF-LC-KJ**

953 **Supplementary Figure S4 Pipeline for the detection of CLCLs**

954 **Supplementary Figure S5 An example CLCL from the clinical samples**

955 **Supplementary Figure S6 SV junctions detected by GenomonSV**

956 **Supplementary Figure S7 Representative histological images of clinical lung cancer**
957 **specimens**

958 **Supplementary Table S1 Summary of lung cancer cell lines**

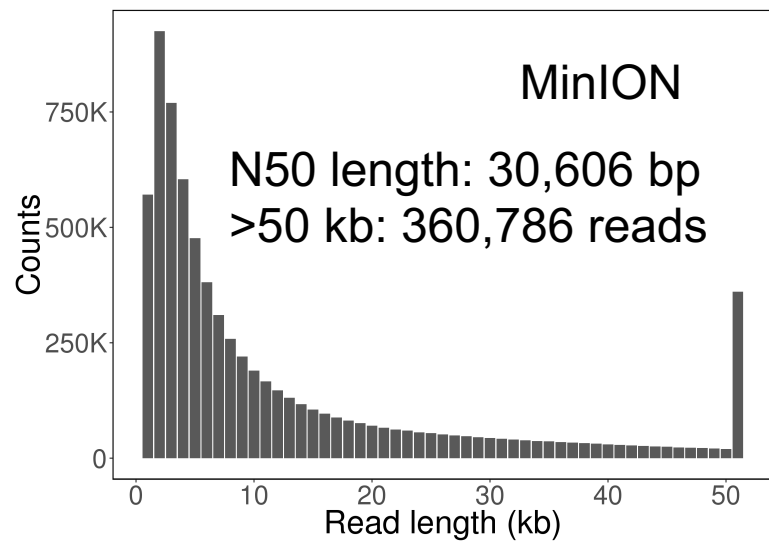
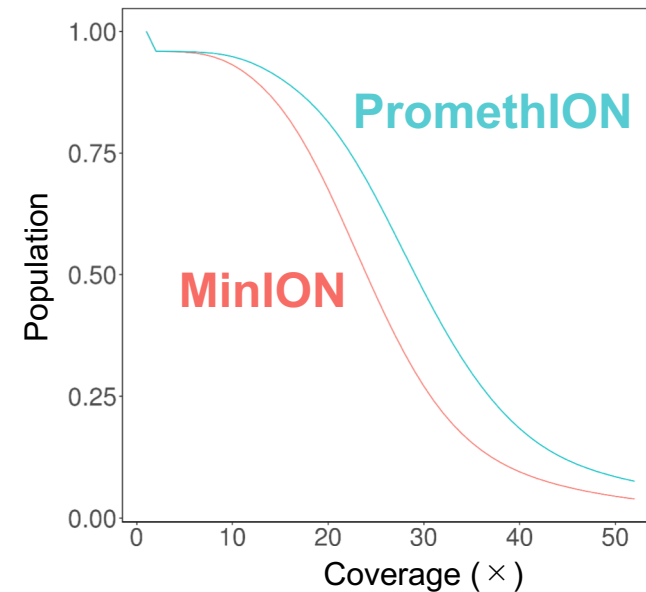
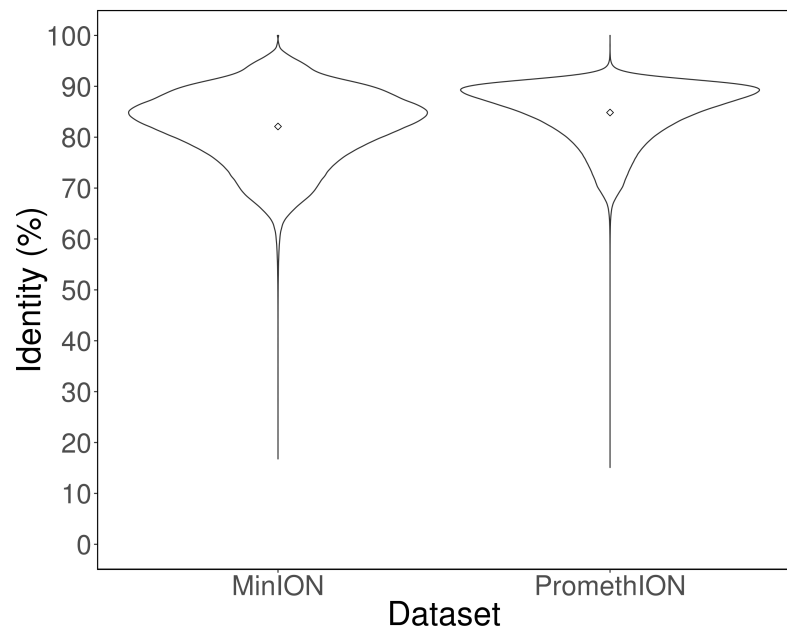
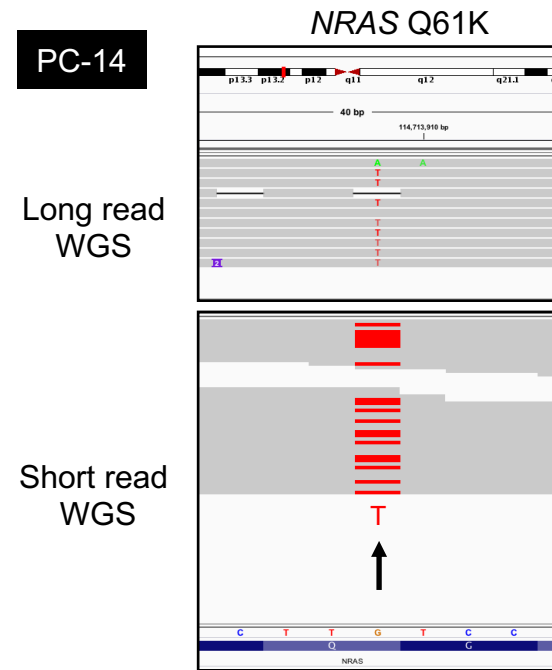
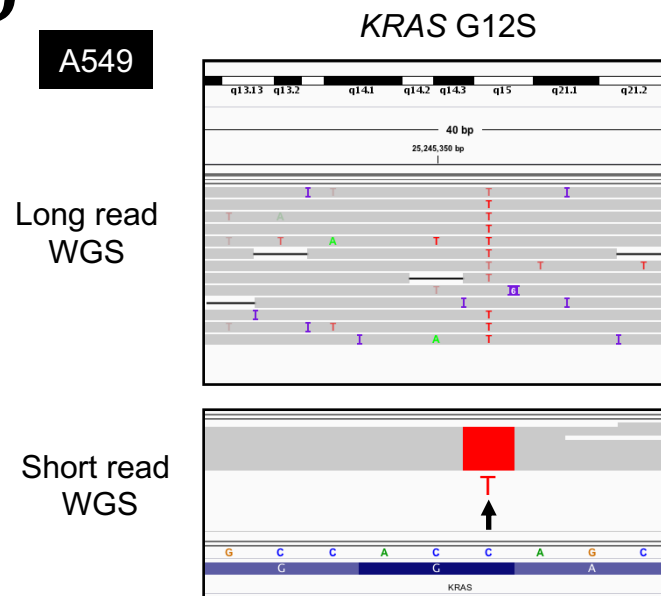
959 **Supplementary Table S2 Sequencing statistics of MinION and PromethION**

960 **Supplementary Table S3 Candidate novel fusion genes**

961 **Supplementary Table S4 Histopathological information on lung cancer clinical samples**

962 **Supplementary Table S5 General statistics of PromethION in lung cancer clinical**
963 **samples**

964 **Supplementary Table S6 Numbers of genes affected by CLCLs in cell lines and clinical**
965 **samples**

A**B****C****D****Figure 1**

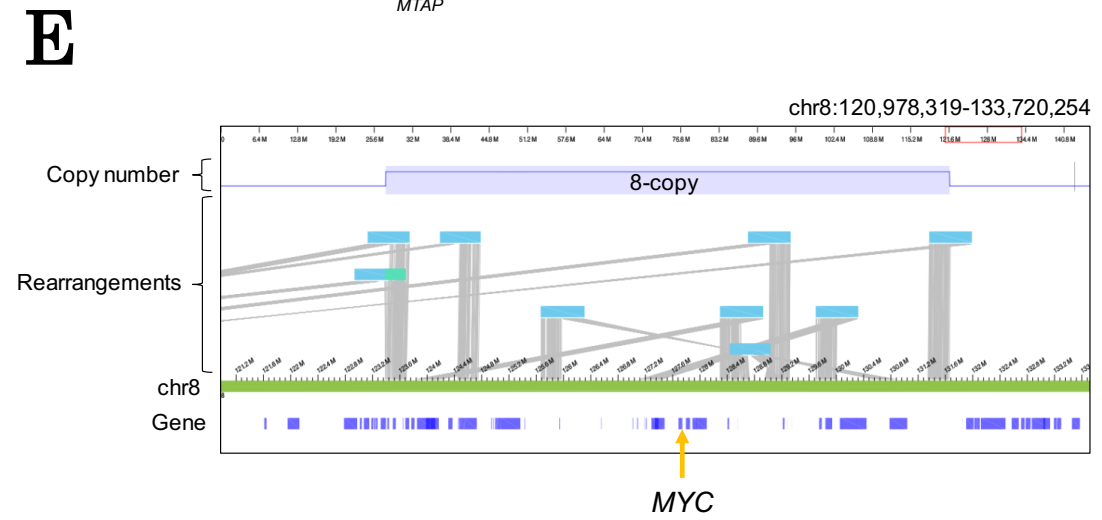
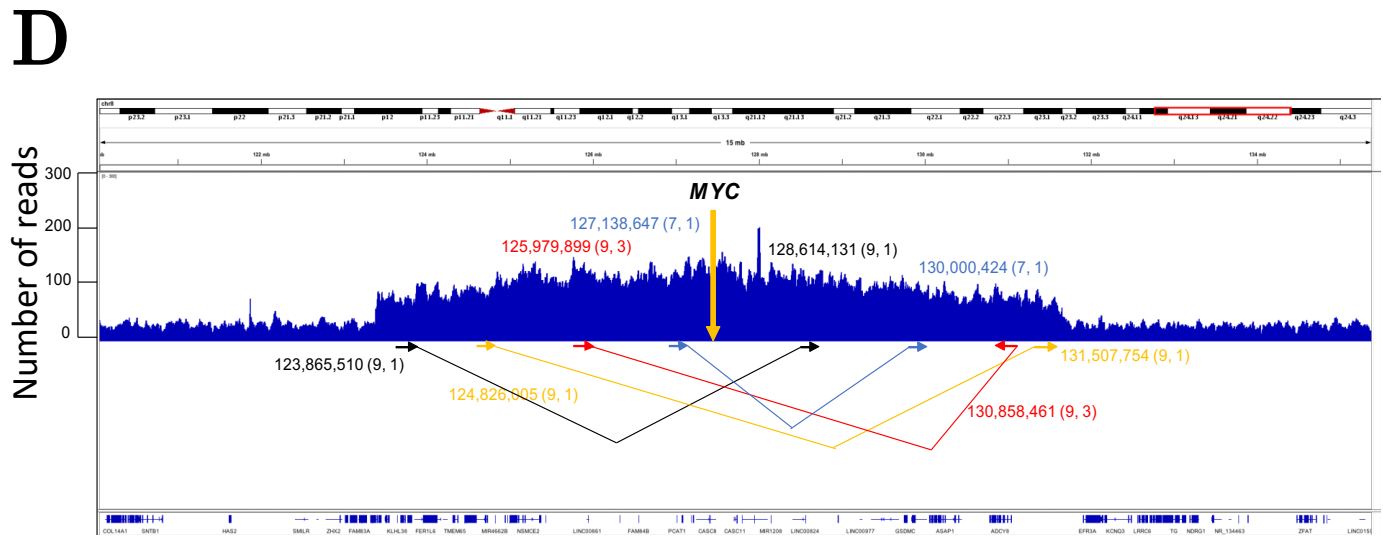
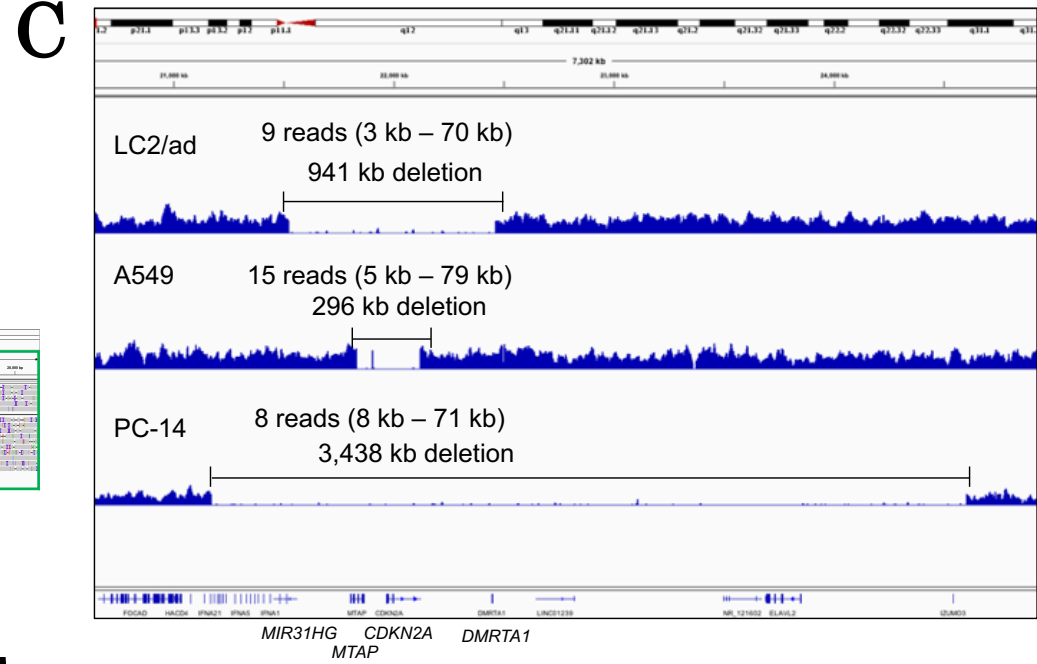
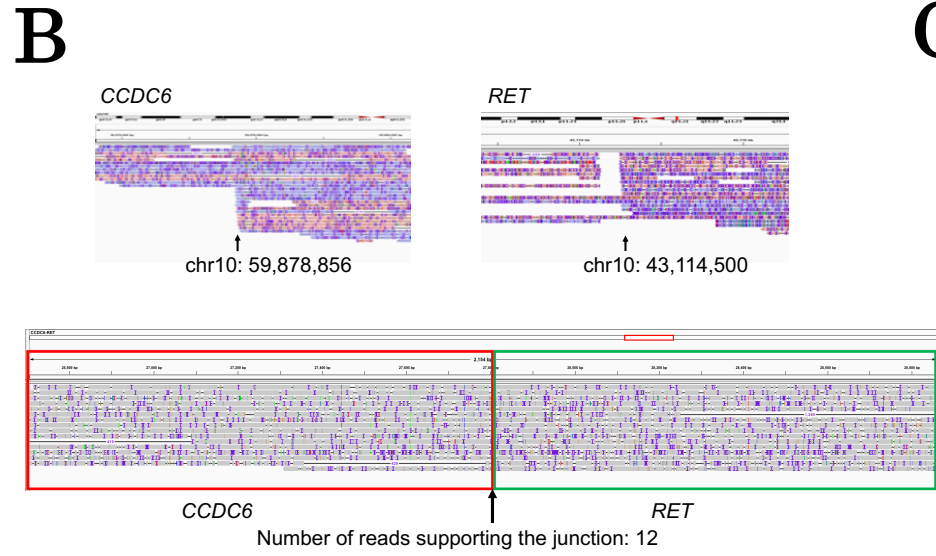
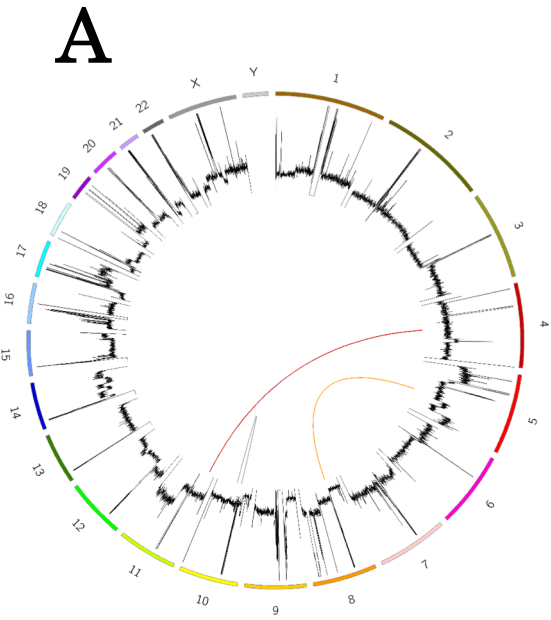
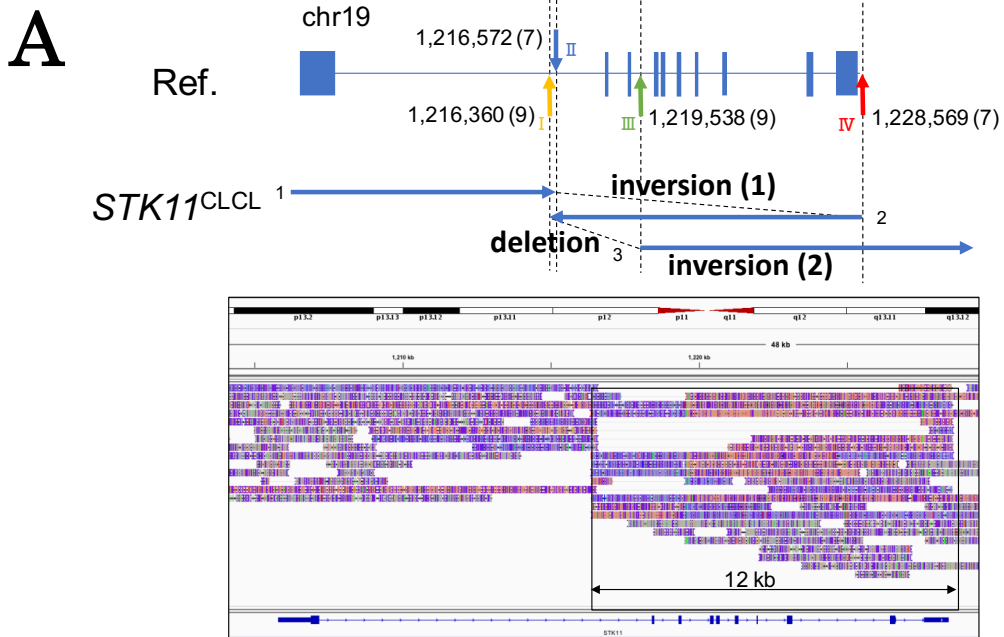
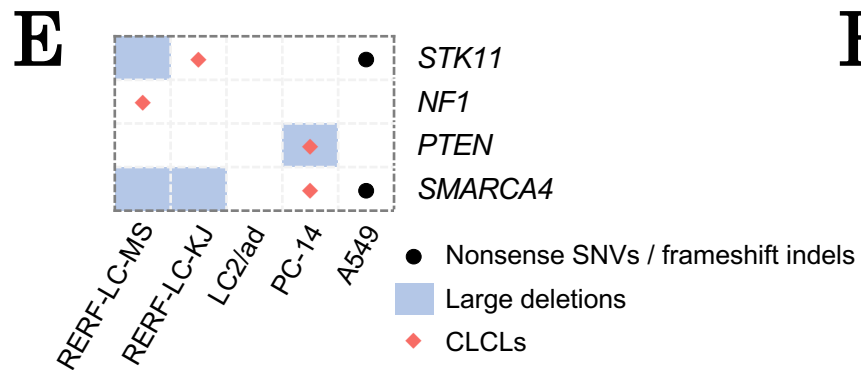
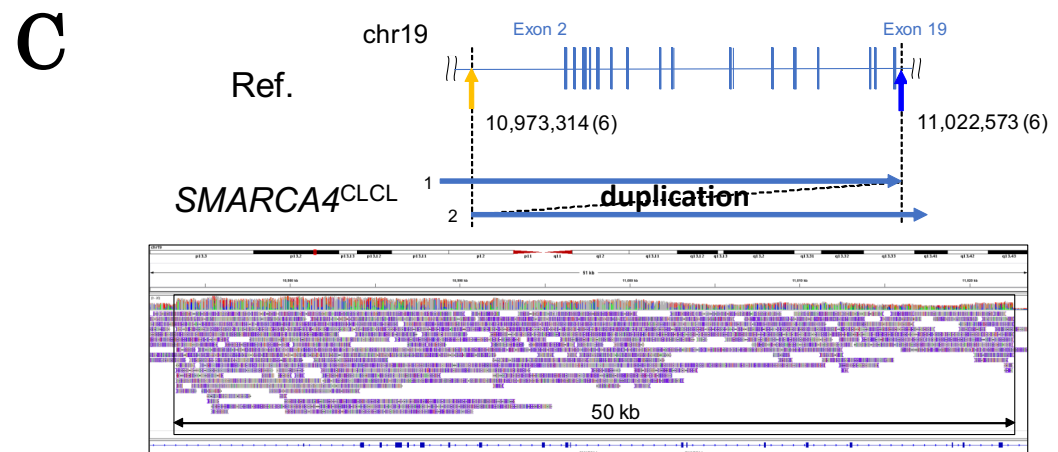
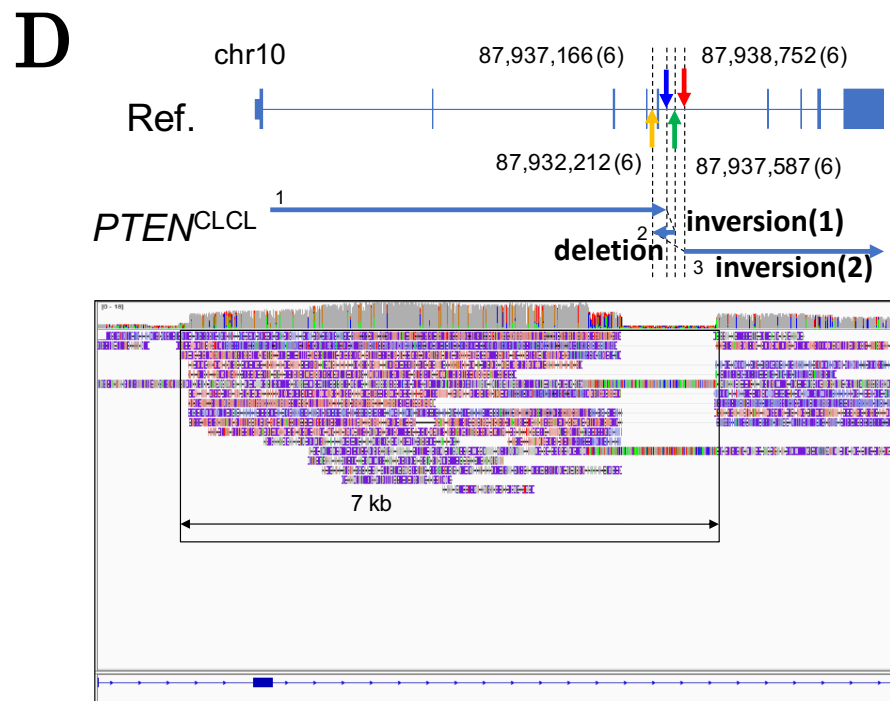


Figure 2



B

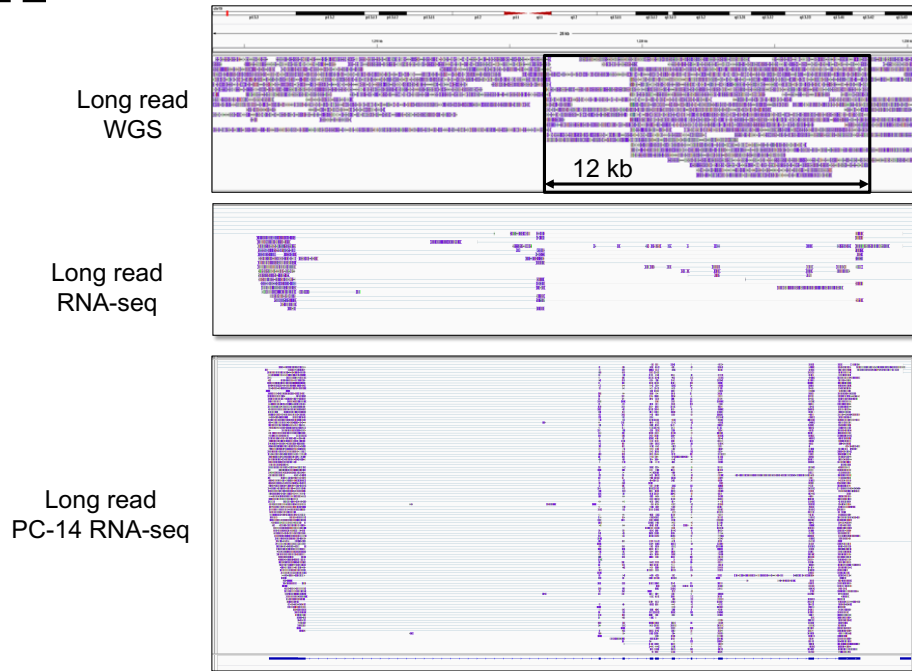
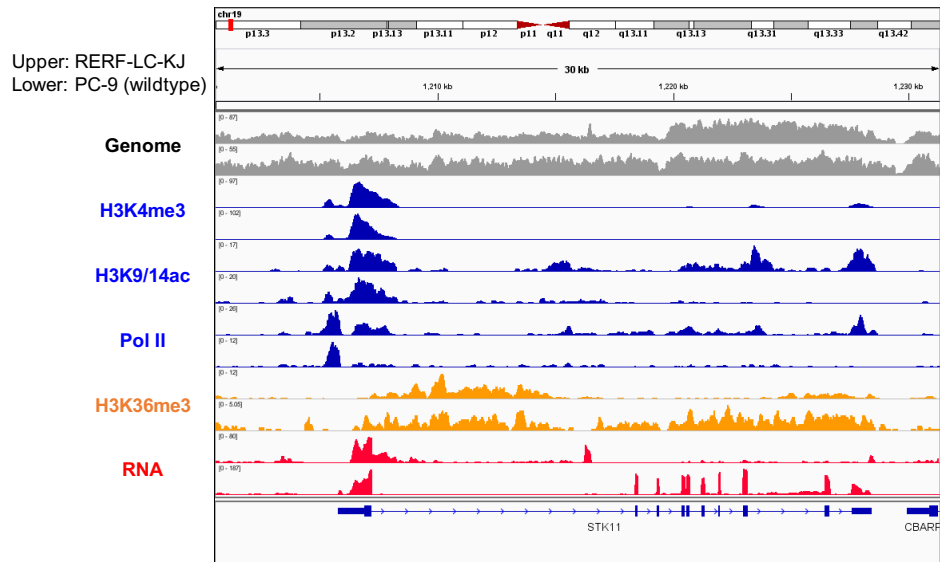
chr17
Ref. Exon 9 Exon 36
31,200,948 (6) 31,278,880 (6)
NF1^{CLCL} 1 duplication 2
78 kb



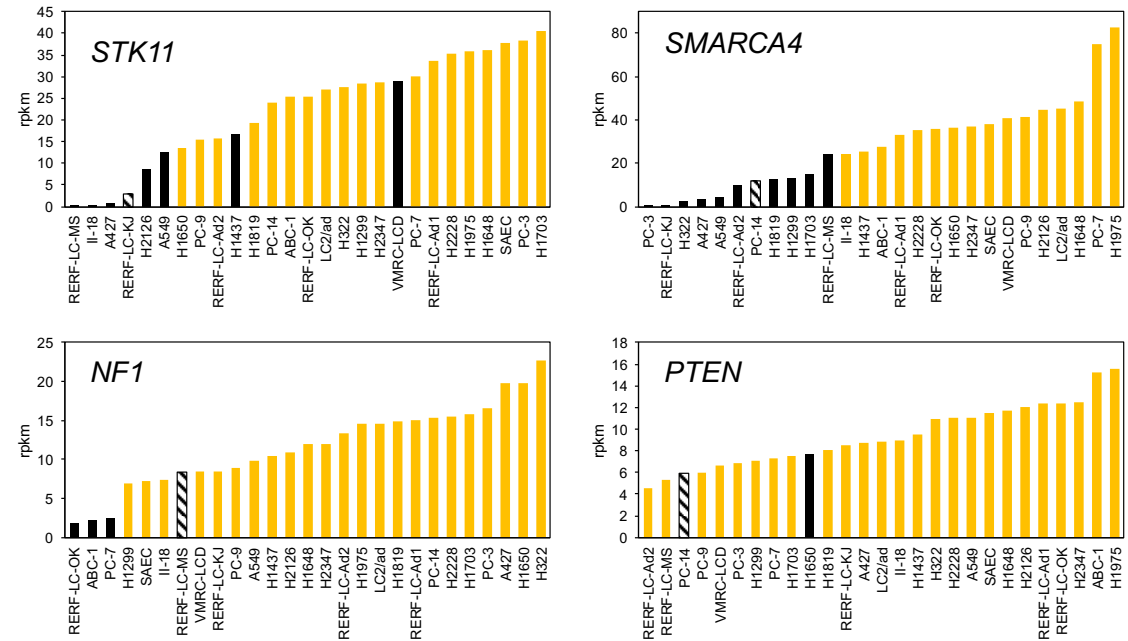
F

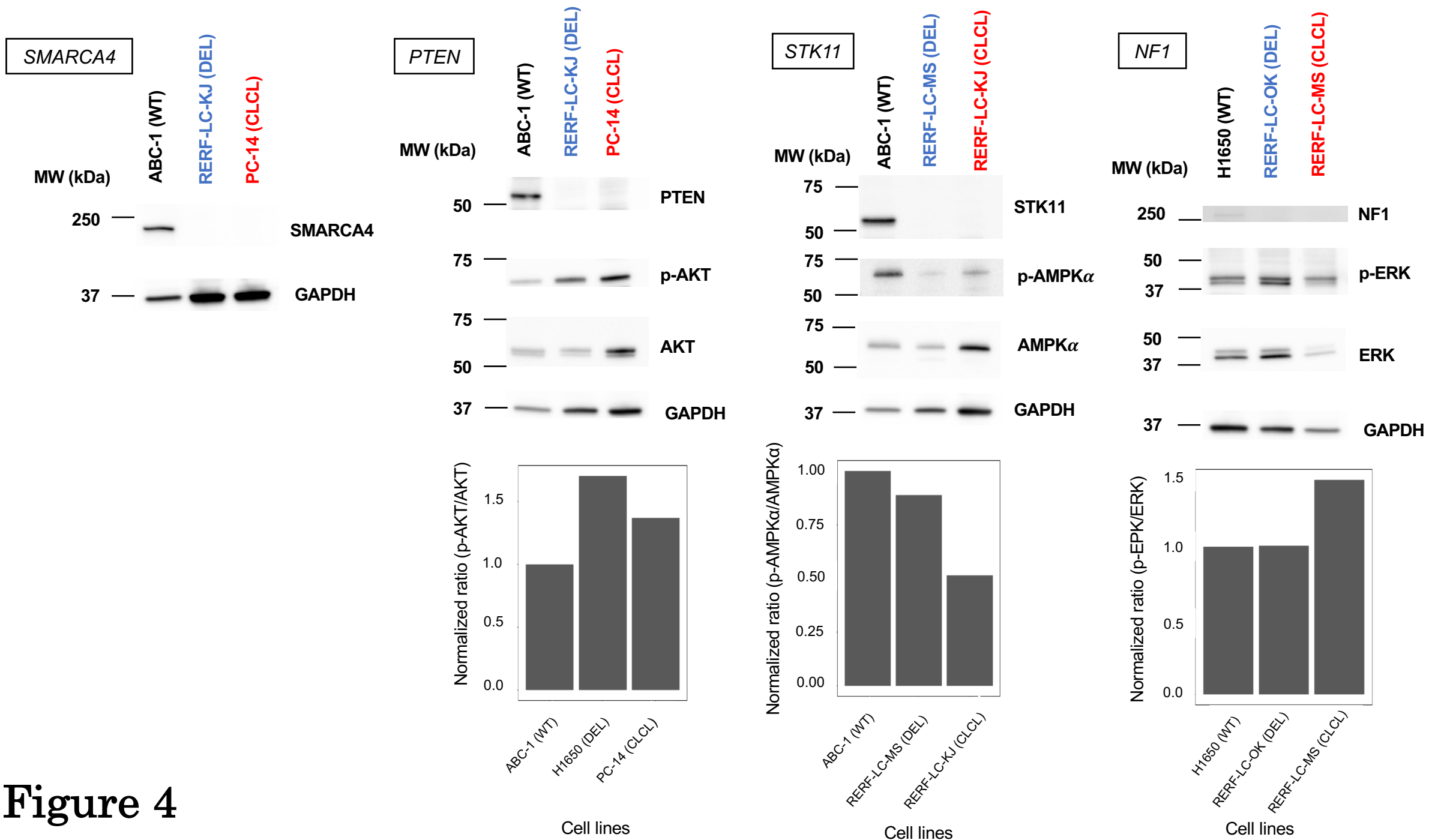
	LINE	SINE	LTR	Others
Number of junctions	12	22	4	54
Percentage (%)	13	24	4	59

Figure 3

A**STK11 in RERF-LC-KJ****B****C****PTEN in PC-14****D**

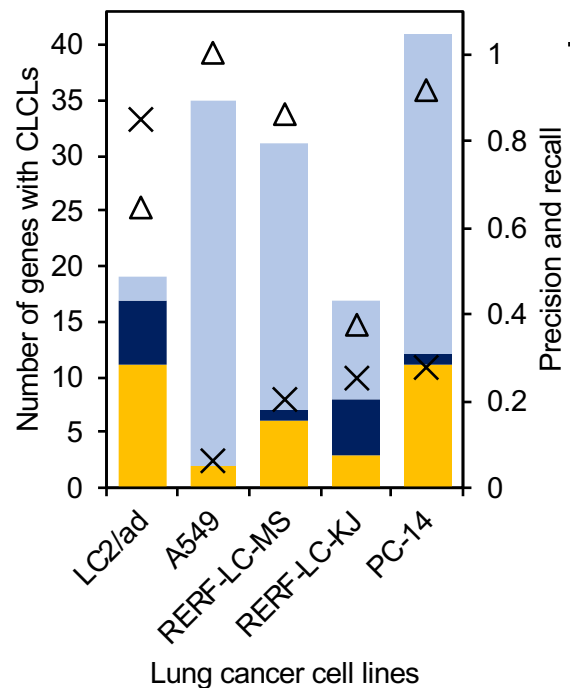
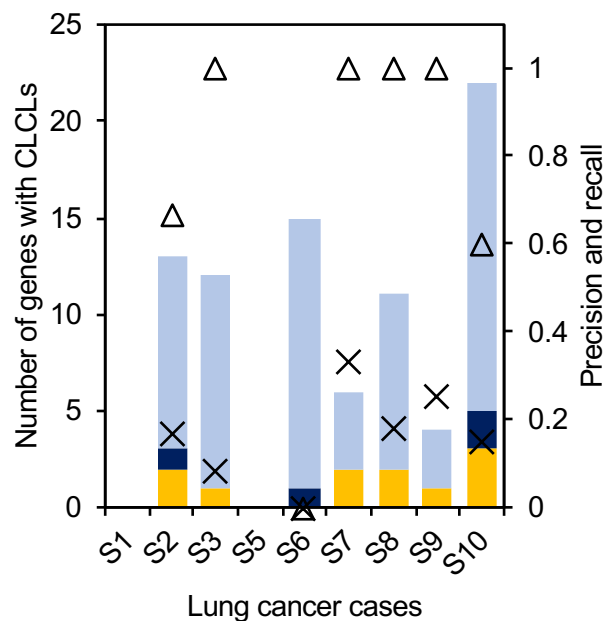
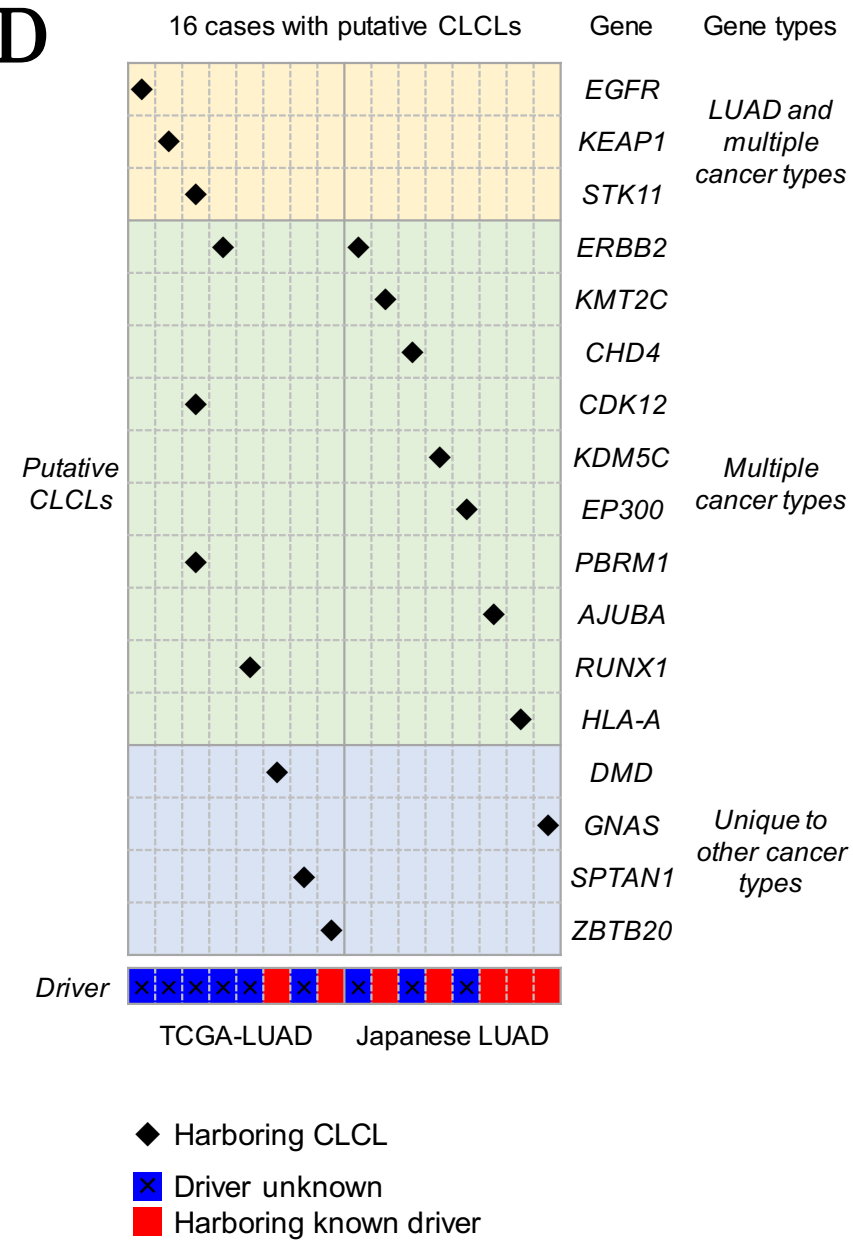
■ Large deletions, frameshift indels, nonsense SNVs detected by short reads
 ▨ CLCLs detected by long reads

**Figure 4**

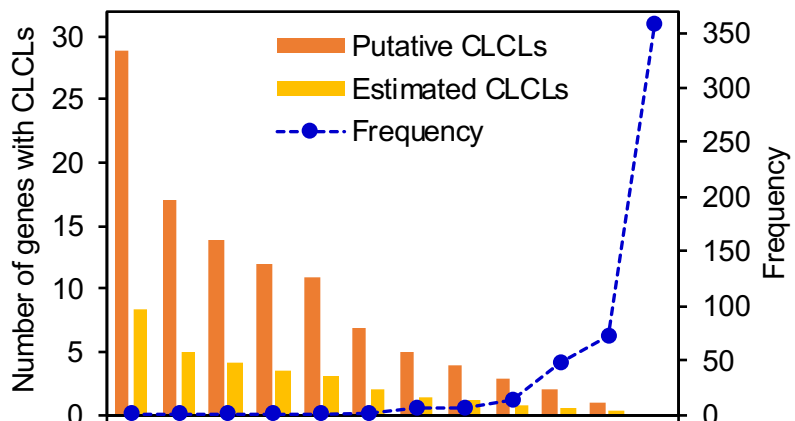
E**Figure 4**

A

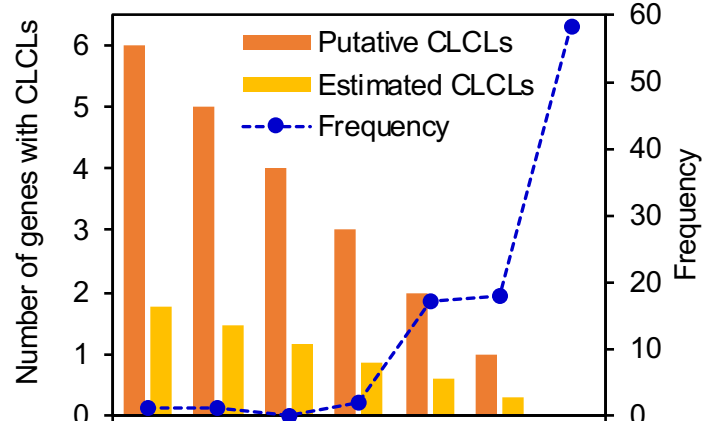
■ FP (short read only)
 ■ FN (long read only)
 ■ TP (both detected)
 × Precision
 △ Recall

**B****D****C**

TCGA-LUAD 514 cases



Japanese LUAD 97 cases



Putative CLCLs	29	17	14	12	11	7	5	4	3	2	1	0
Estimated CLCLs	8.5	5.0	4.1	3.5	3.2	2.0	1.5	1.2	0.9	0.6	0.3	0
Frequency	1	1	1	1	1	2	7	7	14	48	72	359

Putative CLCLs	6	5	4	3	2	1	0
Estimated CLCLs	1.8	1.5	1.2	0.9	0.6	0.3	0
Frequency	1	1	0	2	17	18	58

Figure 5