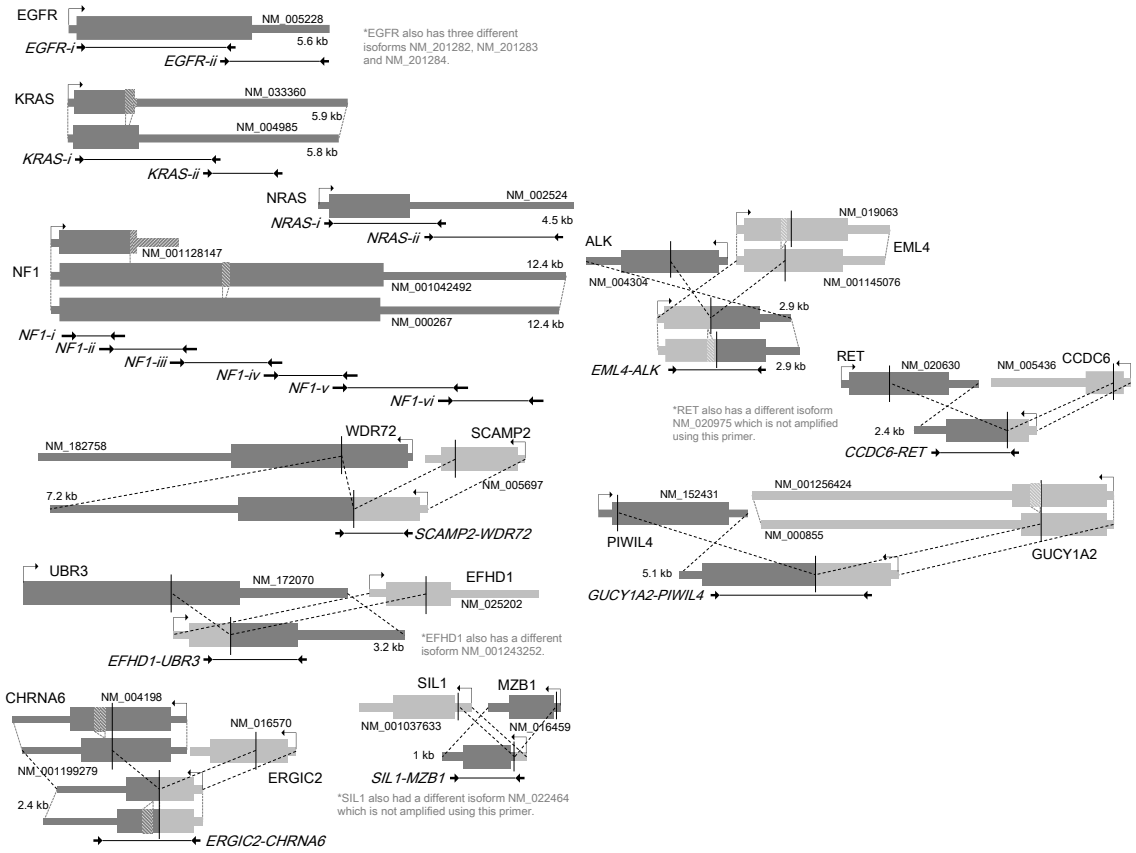
Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer

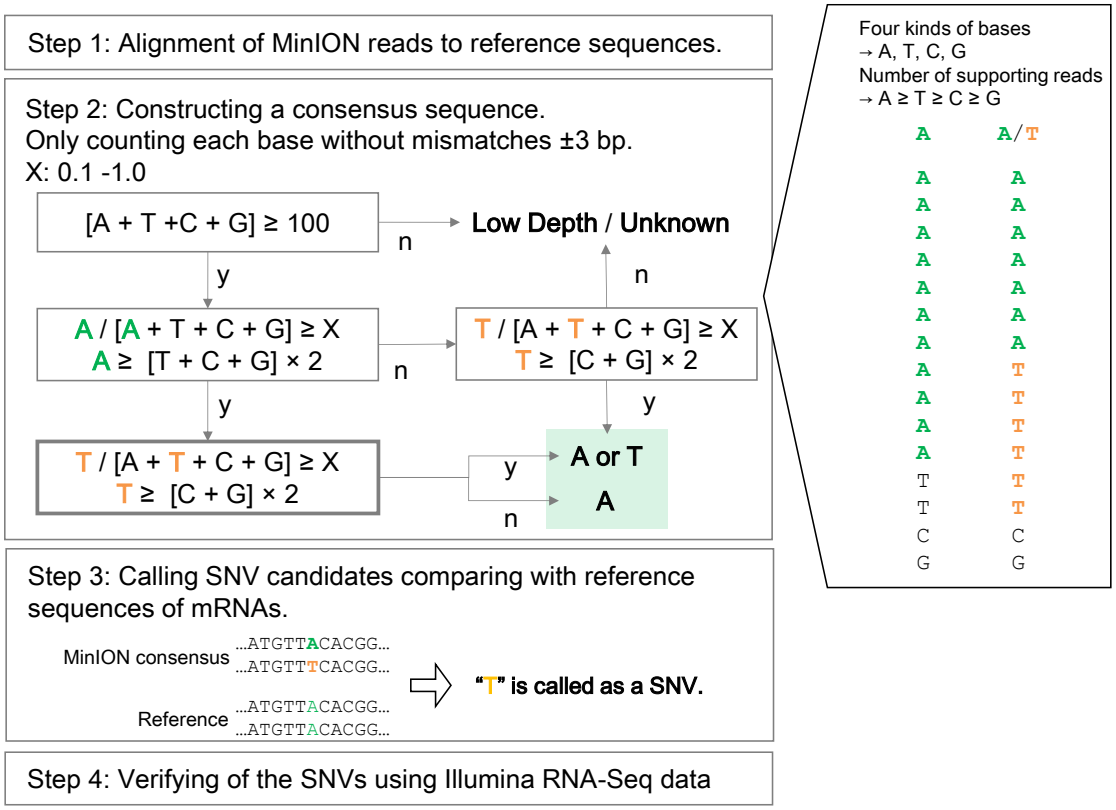
Ayako Suzuki, Mizuto Suzuki, Junko Mizushima-Sugano, Martin C. Frith, Wojciech Makałowski, Takashi Kohno, Sumio Sugano, Katsuya Tsuchihara and Yutaka Suzuki

Supplementary Figures



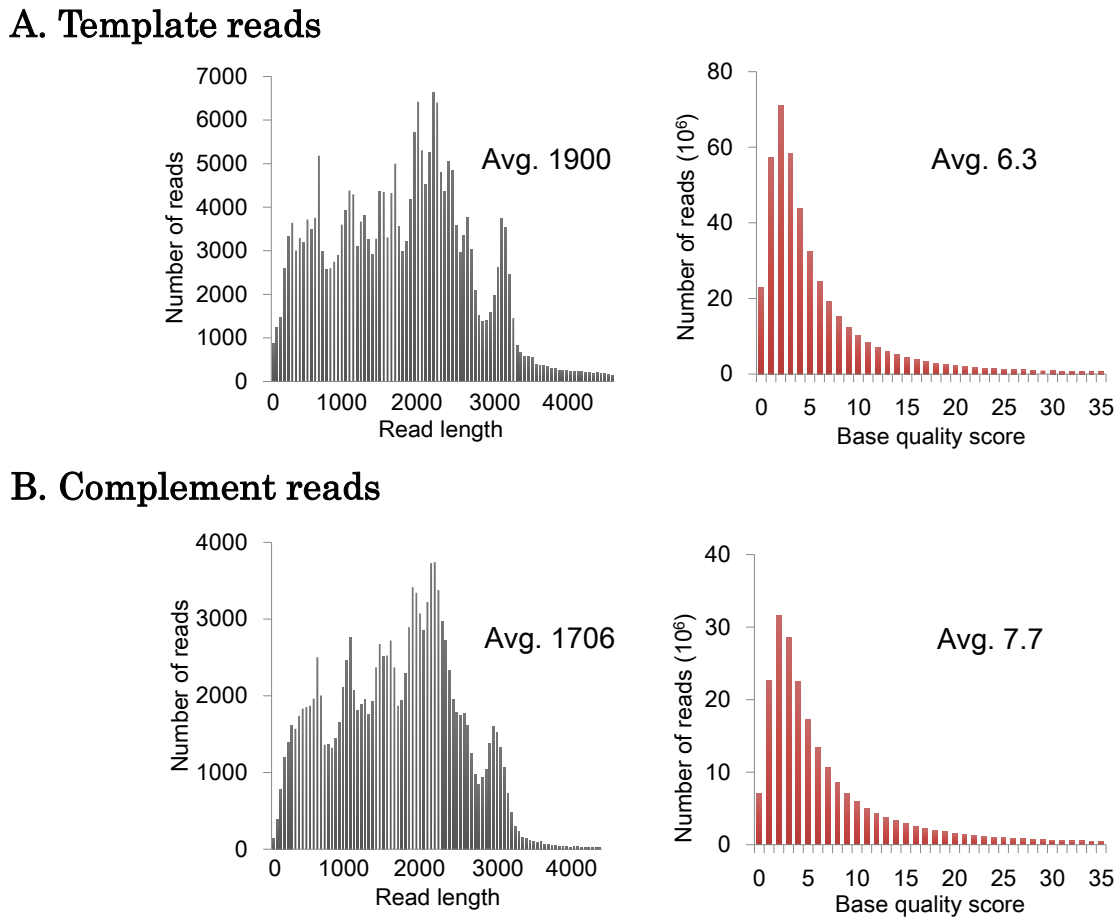
Supplementary Figure S1. Target regions for RT-PCR and MinION sequencing.

Target regions for RT-PCR and MinION sequencing of six known cancer-related genes and five potential fusion transcripts. Images of CDSs and UTRs in each of the transcripts are shown as thick and thin bands, respectively. The small arrows indicate the PCR primers. Some gene were amplified with separate primer sets because it is difficult to amplify large fragments by PCR with easy settings. Detailed information regarding the primers is provided in **Supplementary Table S2**.



Supplementary Figure S2. Method for SNV detection.

Using MinION reads, SNVs were detected as follows: 1) Aligning the reads to the PCR target regions, 2) calculating the depths of each base context in each position by only considering reads without errors ±3 bp of that position and constructing consensus sequences of the MinION reads (see the Methods section for detailed information), 3) comparing the consensus sequence with the reference sequences and detecting the SNV candidates, and 4) verifying the SNV candidates using the Illumina RNA-Seq data[1].

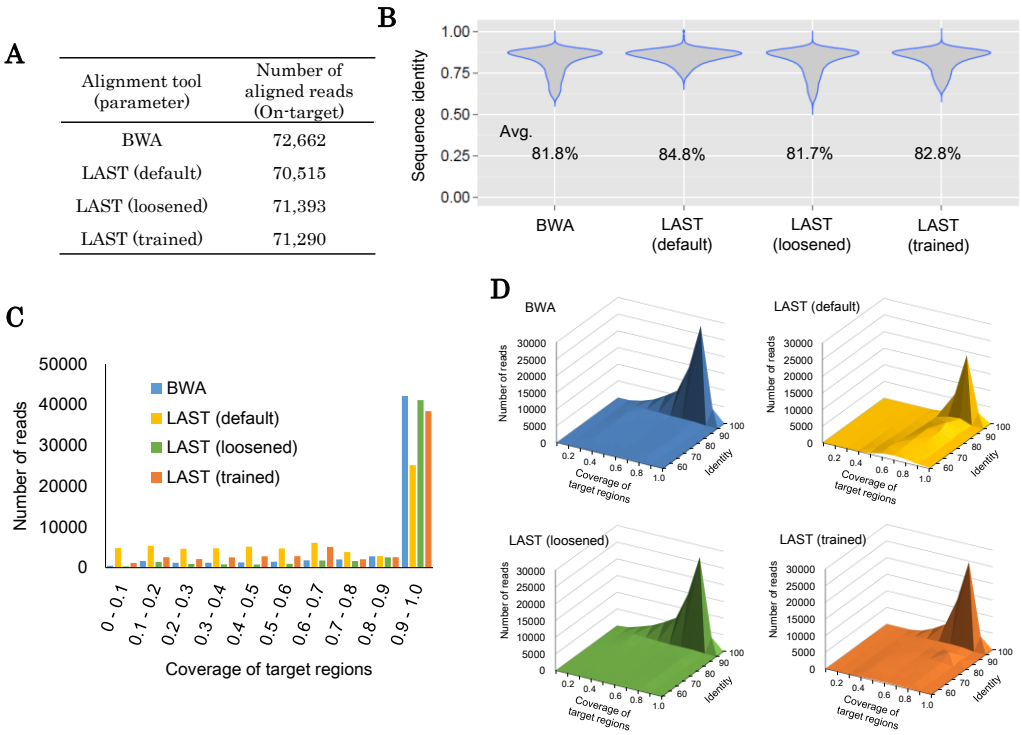


Supplementary Figure S3. Read lengths and QVs in template and complement reads.

The distributions of the read lengths and QVs in the template (**A**) and complement reads (**B**) are shown in the left and right panel, respectively. The average number is shown in the inset.

A

Alignment tool (parameter)	Number of aligned reads (On-target)
BWA	72,662
LAST (default)	70,515
LAST (loosened)	71,393
LAST (trained)	71,290



Supplementary Figure S4. Comparisons of the sequence identities and coverage of the target regions using different alignment tools and parameters.

Yields and sequence identities of the MinION reads depending on the alignment algorithms and parameters. The reads aligned to the target regions of four cancer-related genes were considered in this analysis. Four different tools/parameters were used in this analysis. (**A**) The number of the MinION reads aligned to the target regions. (**B**) Distributions of the sequence identities. The average numbers are shown in the inset. (**C**) Histograms of the target cover rates. (**D**) Two-dimensional histograms of the sequence identities and target cover rates.

Note that a higher coverage of the target regions is not necessarily "better" because the coverage can trivially be increased by weakening the mismatch and gap costs[2]. Such comparisons risk a race-to-the-bottom in alignment stringency.

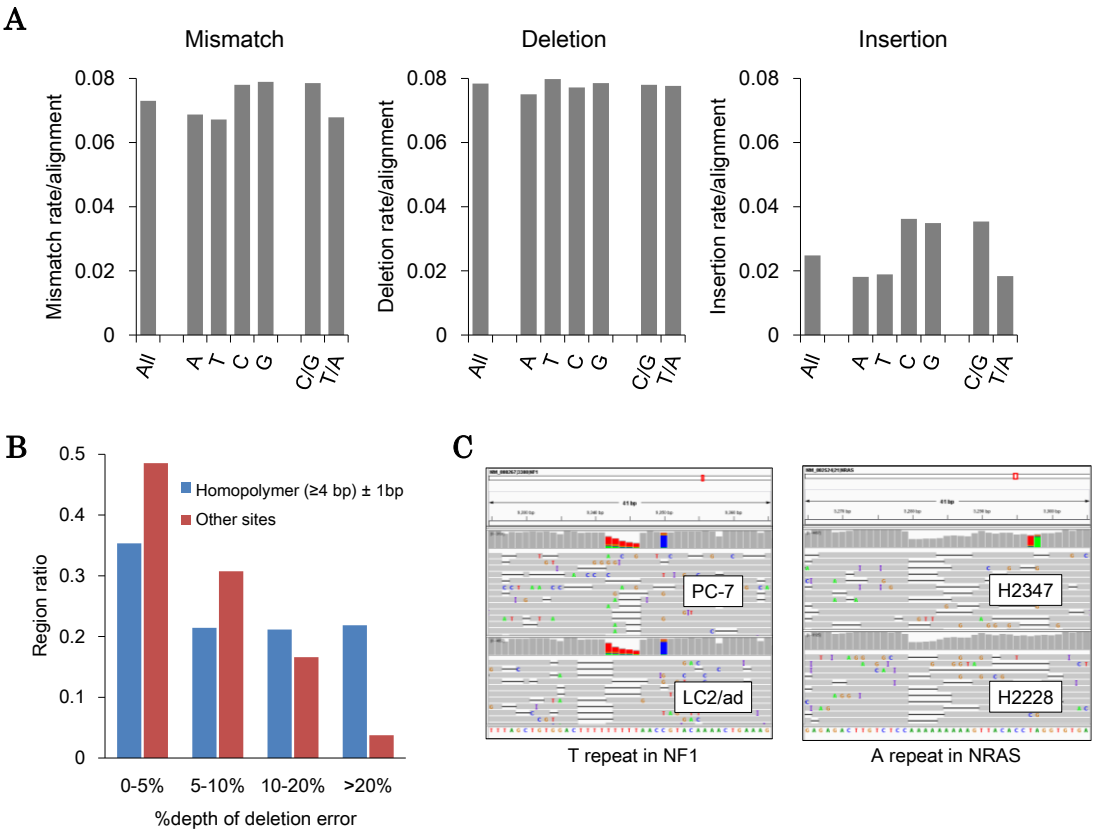
Command:

BWA[3]: bwa mem –x ont2d [ref] [fastq] > [out.sam]

LAST[4] (default): lastal -Q1 [ref] [fastq] > [out.maf]

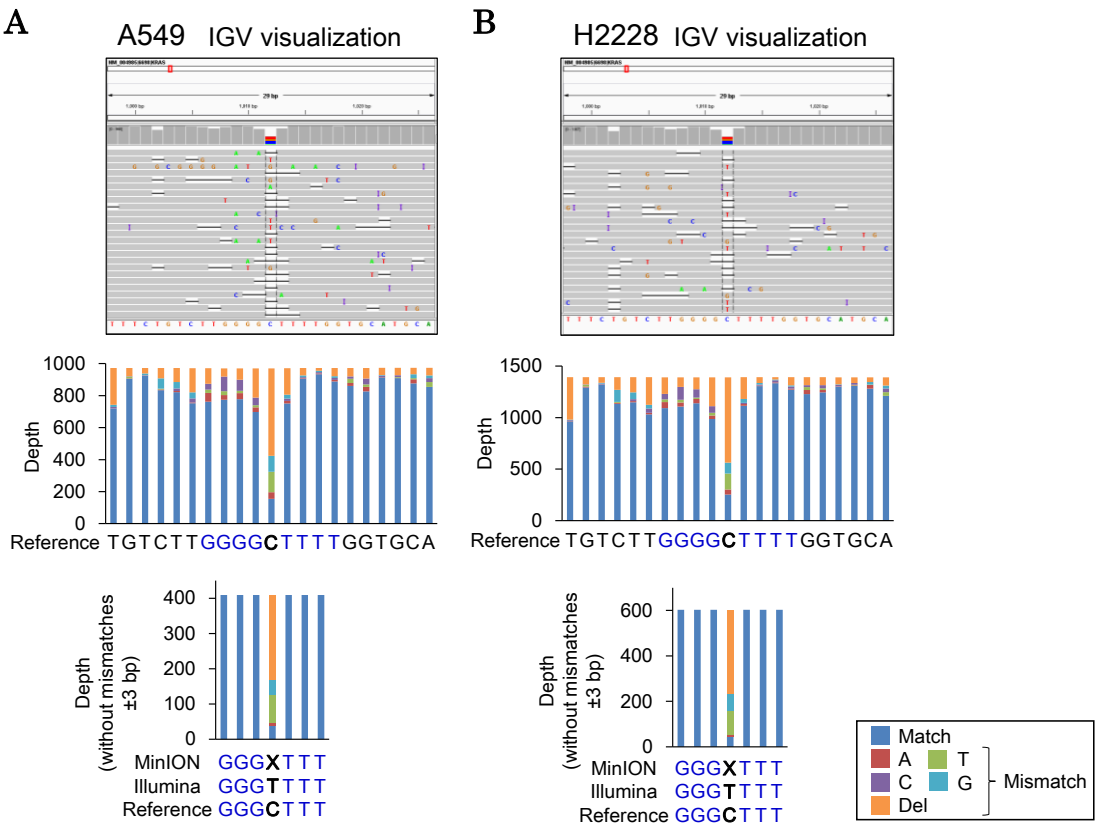
LAST (loosened): lastal -Q1 -r1 -a1 -b1 -q1 [ref] [fastq] > [out.maf]

LAST (trained[5]): lastal -Q1 -p [output of last-train] [ref] [fastq] > [out.maf]



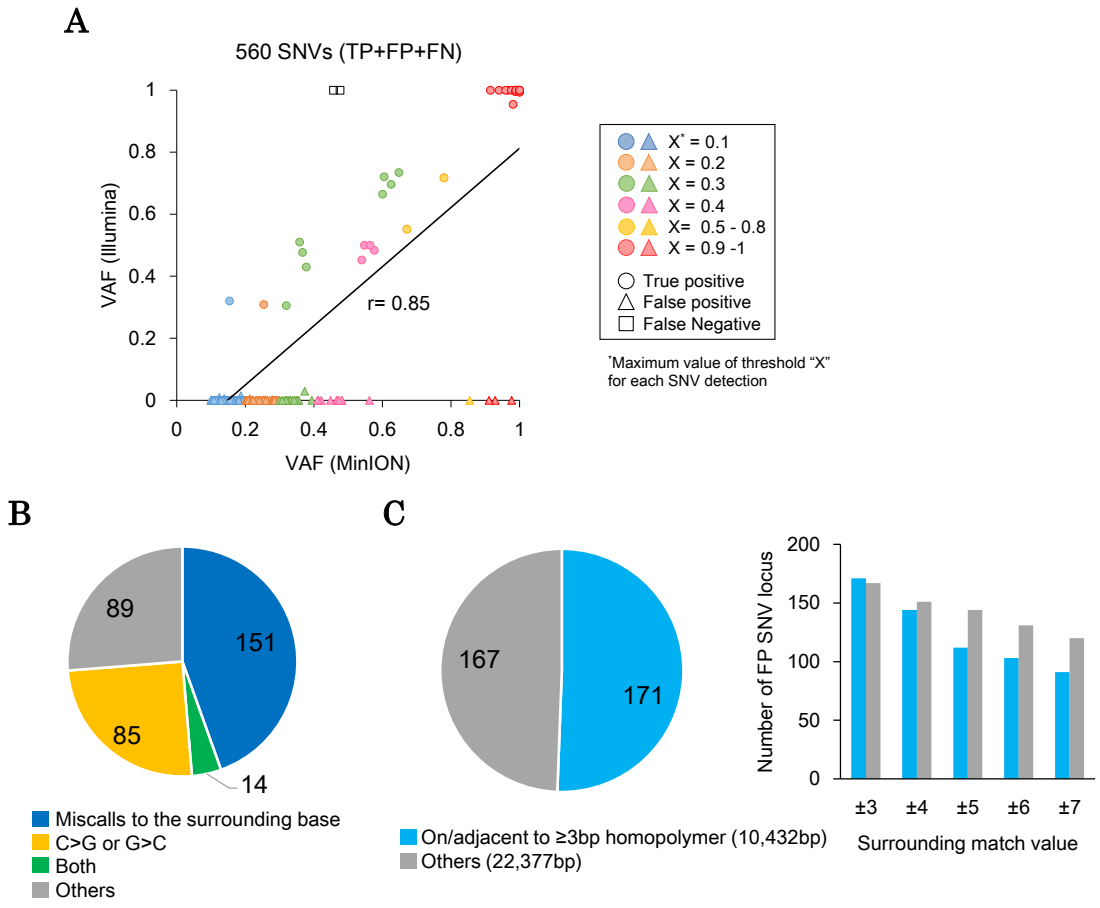
Supplementary Figure S5. Error rates in the MinION reads.

(**A**) The average error rates are shown. The error rates for mismatches, deletions and insertions are shown in the left, center and right panels, respectively. The positions of known SNPs and mutations were not included in the calculation of the error rates. (**B**) The depths of the deletion errors between homopolymers and other sites in the target regions. Approximately 21% of the homopolymer regions showed more than 20% tag deletion frequencies (the rightmost blue bar), whereas these highly frequent deletions were only observed in 4% of the other non-homopolymer regions (rightmost red bar). (**C**) Examples of deletion errors accumulated at homopolymer sites.



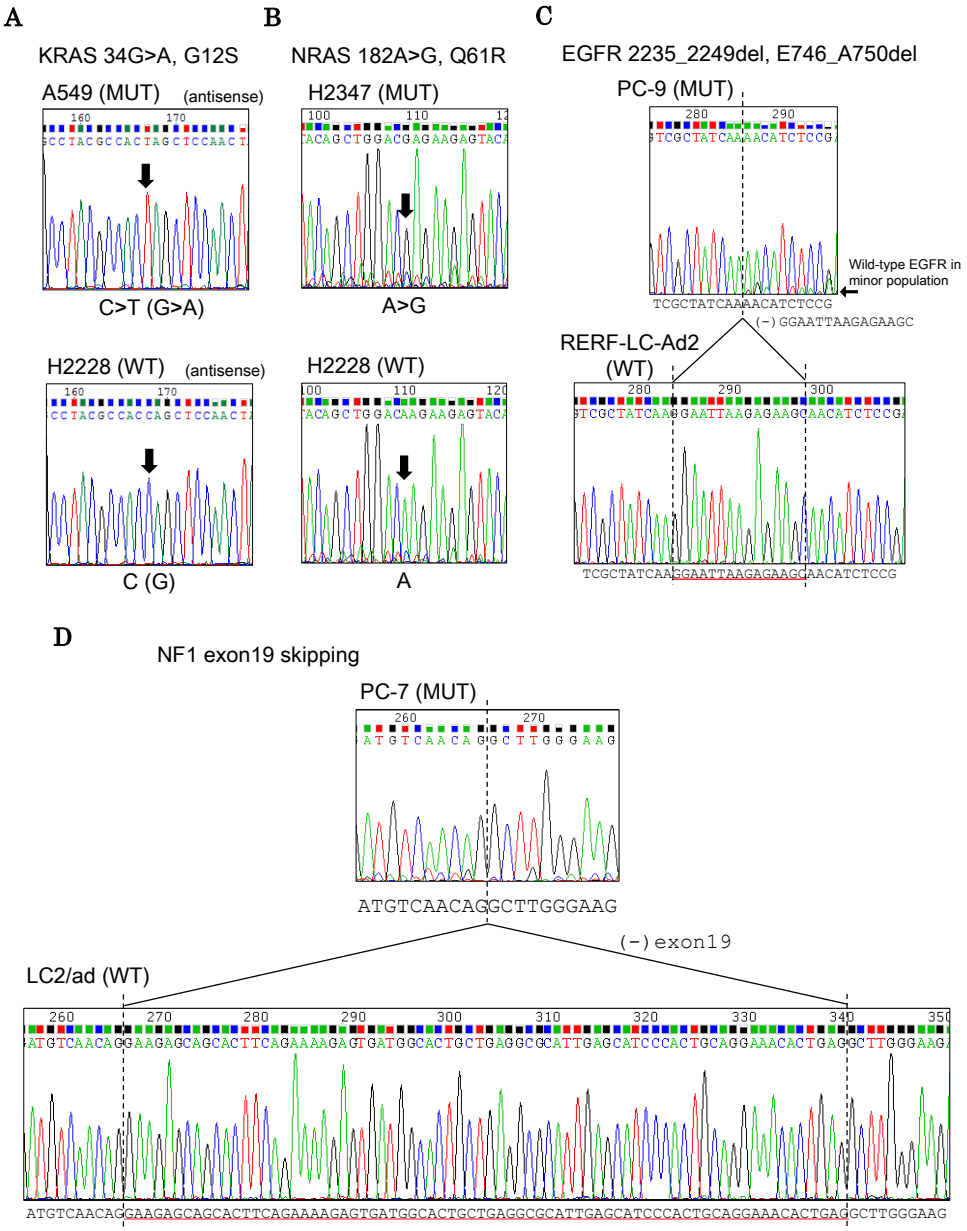
Supplementary Figure S6. False negative SNV detection.

The two false-negative SNVs detected are shown in this figure. A homozygous SNP in KRAS (*264C/T, 3'UTR) in A549 cells (**A**) and H2228 cells (**B**) was not detected because of the SNP located between the 4-mer homopolymers GGGG and TTTT.



Supplementary Figure S7. False positive SNV detection.

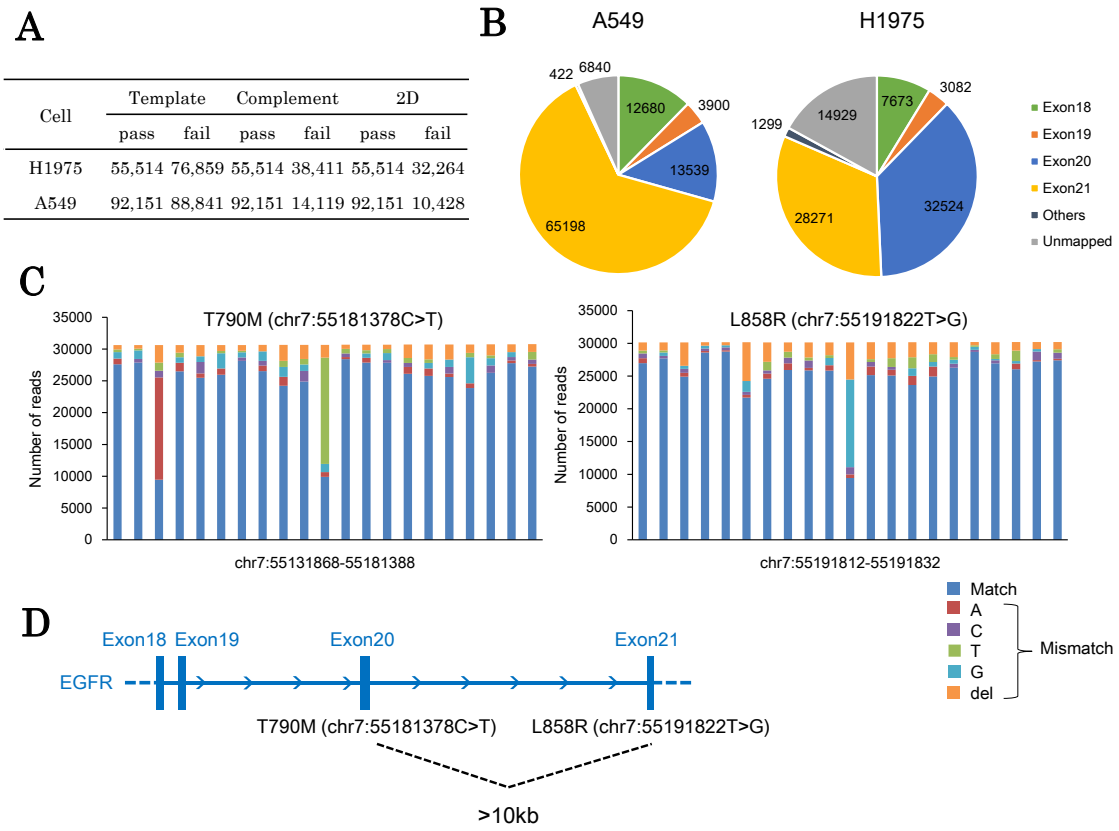
(**A**) Variant tag frequencies in the SNV detection compared between the MinION and Illumina RNA-Seq data. A similar analysis was conducted in **Fig. 2C**. False positive detection (521 SNVs) in the MinION reads, shown as triangles, are also included in the graph. The color of the triangles represents the maximum "X" (the depth threshold for MinION reads) for each false position detected. (**B**) Patterns of false positive SNVs. In total, 338 false positive SNV loci (521 false positive SNV detection) were observed when "X" was set to 0.1. One SNV locus was redundantly counted because different bases were called between the cell lines. (**C**) Number of false positives detected compared between the homopolymer and other sites (left, when the surrounding match, with one of the parameters for SNV detection, set to ±3; right, when the parameter changed from ±3 to 7).



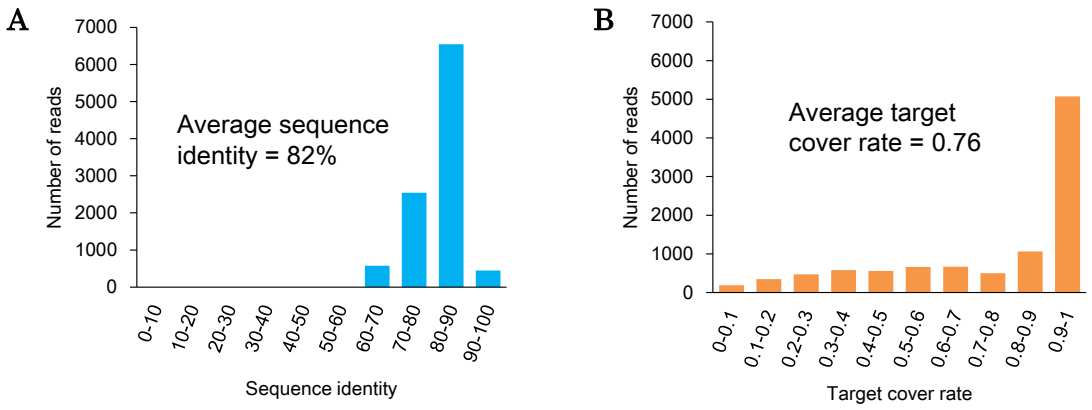
Supplementary Figure S8. Validation of the cancer-related aberrations using Sanger sequencing. The direct Sanger sequencing data of the mutations are shown for the mutant and wild-type cell lines. KRAS G12S in A549 (**A**), NRAS Q61R in H2347 (**B**), EGFR E746_A750del in PC-9 (**C**), NF1 exon 19 skipping in PC-7 (**D**), and the data in the wild-type cell lines are also represented.

A

Cell	Template		Complement		2D	
	pass	fail	pass	fail	pass	fail
H1975	55,514	76,859	55,514	38,411	55,514	32,264
A549	92,151	88,841	92,151	14,119	92,151	10,428

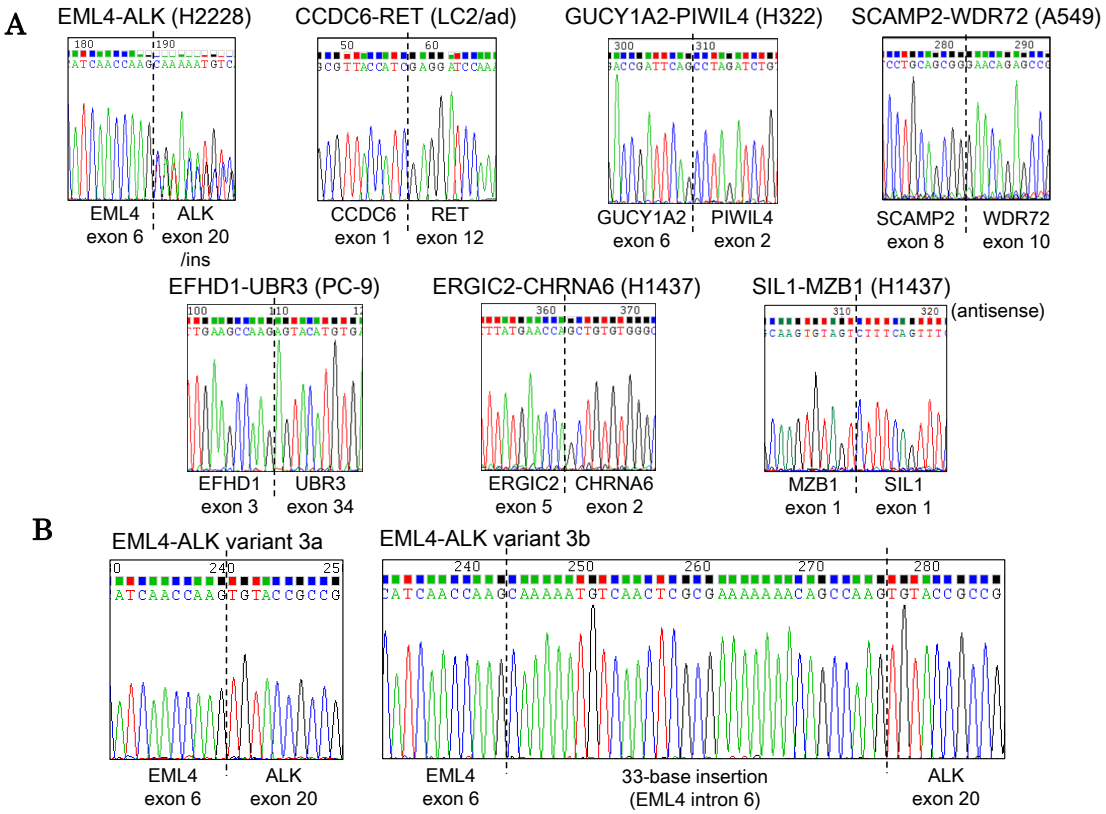


Supplementary Figure S9. Sequencing of EGFR amplicons from genomic DNAs using MinION. (**A**) The number of reads obtained using EGFR genomic DNA amplicons of H1975 and A549 using MinION sequencing. (**B**) Mapping results of EGFR amplicons. (**C**) EGFR T790M (left) and L858R (right) mutations detected in H1975 genomic DNA templates by MinION sequencing. (**D**) Locus of exon 18-21 of EGFR in the human genome UCSC hg38. The distance between T790M (exon 20) and L858R (exon 21) in the genome is greater than 10 kb.



Supplementary Figure S10. Sequence identity and coverage of MinION reads aligned to the fused RNAs.

Sequence identity (**A**) and target cover rate (**B**) of the MinION reads aligned to the seven fused RNAs. The average numbers are shown in the inset.



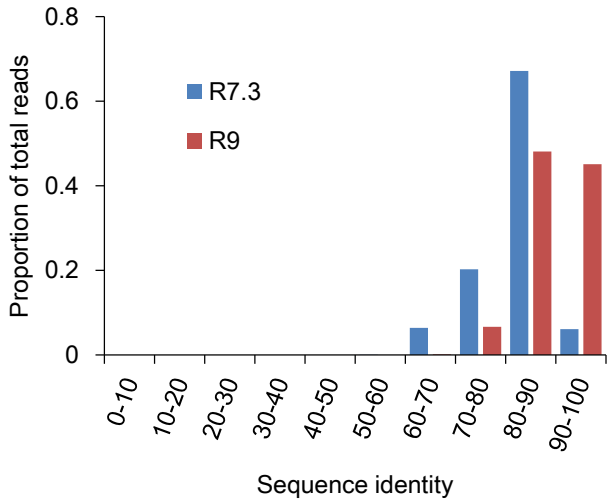
Supplementary Figure S11. Validation of the fusion transcripts.

(**A**) Validation of the junction points in the fusion transcripts. The junction points of the seven fusion transcripts were validated by direct Sanger sequencing of the PCR amplicons. (**B**) Two types of junction points were confirmed by TA cloning in the EML4-ALK in H2228 cells.

A

Cell	Fwd primer	Rev primer	Index	Fwd	Rev
H2228	*EGFR-s_fw-ind1*	*EGRF-1-1_rv*	GGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCT	CTAAGATCCCGTCCATCGCC	ACATATGGGTGGCTGAGGGA
PC-9	*EGFR-s_fw_ind2*	*EGRF-1-1_rv*	GGTGCTGTCGATTCCGTTTGTAGTCGTCTGTTTAACCT	CTAAGATCCCGTCCATCGCC	ACATATGGGTGGCTGAGGGA
H1975	*EGFR-s_fw_ind3*	*EGRF-1-1_rv*	GGTGCTGGAGTCTTGTGTCCCAGTTACCAGGTTAACCT	CTAAGATCCCGTCCATCGCC	ACATATGGGTGGCTGAGGGA

B



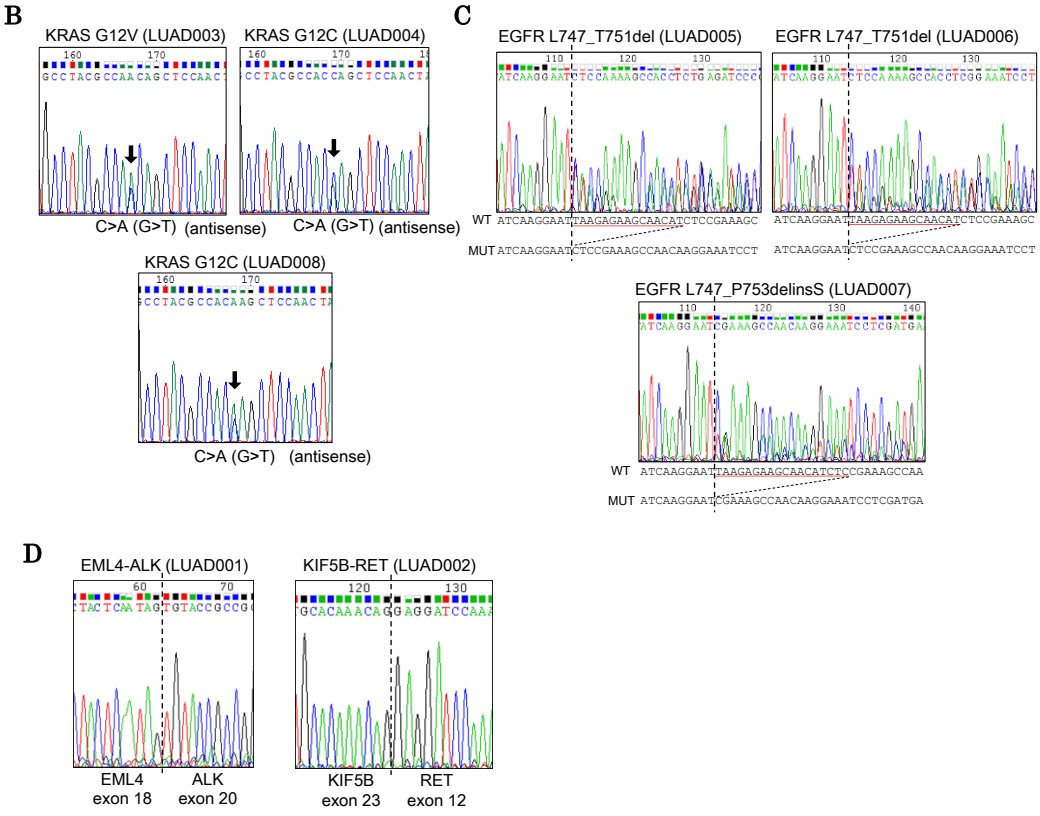
Supplementary Figure S12. MinION sequencing of cDNA amplicons using R9 flow cell.

(**A**) PCR primers for cDNA amplicons for EGFR. Barcode index sequences were used for multiplex sequencing. (**B**) Comparison of the sequence identity between R7.3 and R9 flow cells. On average, 83% and 88% sequence identity are shown in the data from the R7.3 (pass + fail) and R9 (pass only) flow cells, respectively.

Methods: The cDNA templates were amplified using 1 cycle of 20 min at 95 °C; 30 cycles of 15 sec at 95 °C, 15 sec at 55 °C and 2 min at 72 °C; and a final 2 min cycle at 72 °C with the PCR primers described above. Using the amplified templates, the MinION library preparation was conducted according to the manufacturers' protocol for the R9 flow cells (FLO-MIN104). In total, 7077, 10162 and 6422 2D pass reads were obtained from the templates of H2228, PC-9 and H1975, respectively. All reads were aligned to the mRNA sequences using LAST (v658) with the trained parameters -a12, -A15, -b4 and -B4. Alignments with the best score in each query were extracted and used to calculate the sequence identities.

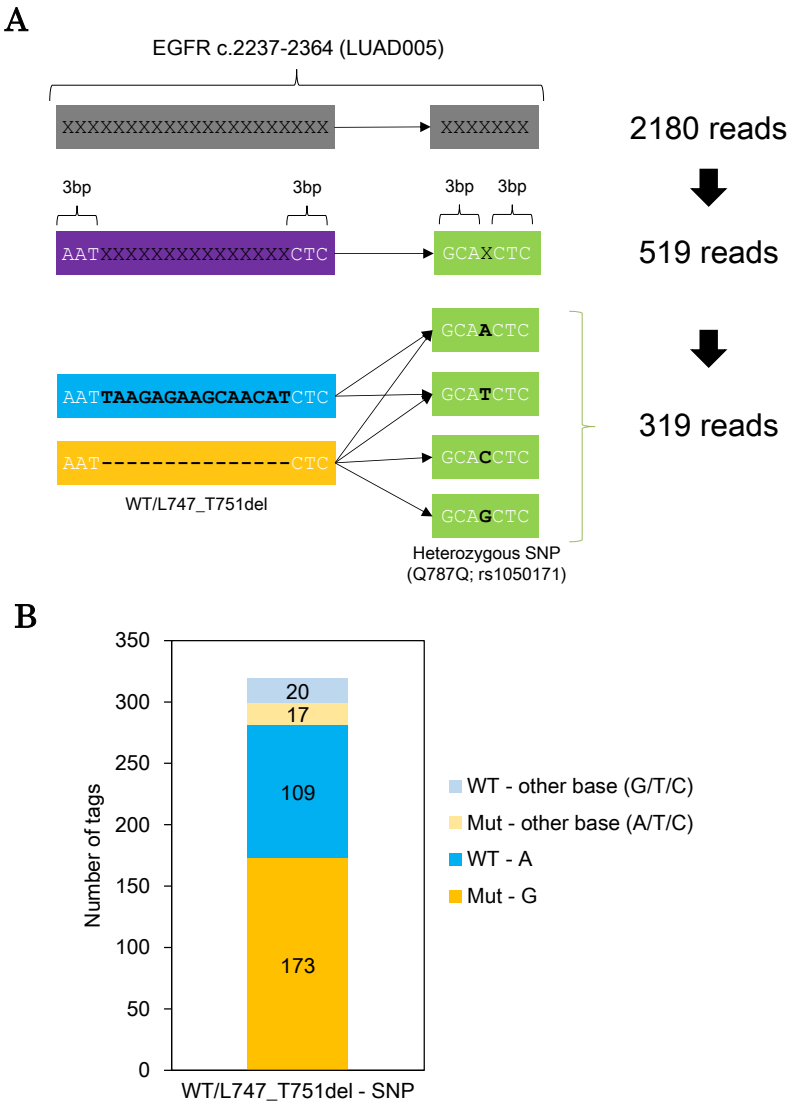
A

ID	Gender	pStage	Recurrence	Smoking history	Driver gene	RNA Integrity Number (RIN)
LUAD005	M	1A	-	-	EGFR	7.7
LUAD006	F	1A	-	-	EGFR	8.0
LUAD007	M	1B	-	+	EGFR	8.6
LUAD003	F	1B	-	+	KRAS	8.6
LUAD004	M	1A	-	+	KRAS	8.1
LUAD008	M	1B	+	+	KRAS	8.0
LUAD001	M	2A	-	+	ALK fusion	8.7
LUAD002	M	1A	-	-	RET fusion	8.0



Supplementary Figure S13. Information regarding the clinical samples and validation of their driver mutations by Sanger sequencing.

(**A**) Eight RNA samples originating from eight Japanese lung adenocarcinoma patients were used in this study. RIN values were measured using a BioAnalyzer. Direct Sanger sequencing of EGFR (**B**) and KRAS (**C**) mutations are shown for the mutant and wild-type samples. (**D**) The junction points in EML4-ALK E18;A20 and KIF5B-RET K23;R12[6] were also validated by direct Sanger sequencing.



Supplementary Figure S14. Phasing analysis of an EGFR exon 19 deletion and a SNP.

(**A**) The number of reads covering both a 15 bp-deletion (c.2240-2254del, p.L747_T751del) and a heterozygous SNP (c.2361G/A) with ±3 bp in EGFR in the case LUAD005. In total, 319 reads were used for the phasing analysis. (**B**) Allelic patterns of the deletion and SNP in the 319 reads. The 15 bp-deletion was phased with 'G' in 54% of the reads, and the wild type was associated with 'A' in 34% of the reads.

Supplementary Tables

Supplementary Table S1. Information regarding the samples used in this study.

Amplicon mixture	Cell line	Gene	Mutation	Primer name
#1	PC-9	EGFR	E746_A750del	*EGFR-i ~ ii*
	A549	KRAS	G12S	*KRAS-i ~ ii*
	H2347	NRAS	Q61R	*NRAS-i ~ ii*
	PC-7	NF1	Exon 19 skipping	*NF1-i ~ vi*
#2	H1975	EGFR	T790M, L858R	*EGFR-i ~ ii*
	H2228	KRAS	WT	*KRAS-i ~ ii*
	H2228	NRAS	WT	*NRAS-i ~ ii*
	LC2/ad	NF1	WT	*NF1-i ~ vi*
#3	LC2/ad	CCDC6-RET	Fusion	*CCDC6-RET*
	H2228	EML4-ALK	Fusion	*EML4-ALK*
	H322	GUCY1A2-PIWIL4	Fusion	*GUCY1A2-PIWIL4*
	A549	SCAMP2-WDR72	Fusion	*SCAMP2-WDR72*
	PC-9	EFHD1-UBR3	Fusion	*EFHD1-UBR3*
	H1437	ERGIC2-CHRNA6	Fusion	*ERGIC2-CHRNA6*
	H1437	SIL1-MZB1	Fusion	*SIL1-MZB1*
	RERF-LC-Ad2	EGFR	WT	*EGFR-i ~ ii*

Supplementary Table S2. PCR primers.

A. RT-PCR of cDNA templates

Primer name	Gene	Product size	Fwd sequence	Rev sequence
EGFR-i	EGFR	3230	CCCTGACTCCGTCCAGTATT	AGCTTTGCAGCCCATTTCTA
EGFR-ii	EGFR	2366	CAGCGCTACCTTGTCATTCA	GCTGTAGGGGCTCTGACTGA
KRAS-i	KRAS	2558 (2682)	AGGCCTGCTGAAAATGACTG	TTCTCTTGAGCCCTGAGGAA
KRAS-ii	KRAS	1525	TGTCATCTTGCCTCCCTACC	TCTCCCCCTTTAAAATCTCTACA
NRAS-i	NRAS	2001	GTGGAGCTTGAGGTTCTTGC	TTTCCTTGCAGGGCTAACTG
NRAS-ii	NRAS	2291	CACAGCAATAGGGGCTTGAT	AAATATCGGCCCTTCCATTT
NF1-i	NF1	1066	GGTCAAACAGTTGCTGCCAG	CGGTGCCATTCGTATTGCTG
NF1-ii	NF1	2052	GCTGCAATTGCCTGTGTCAA	CCCAAGCACACGAACATACC
NF1-iii	NF1	2685 (2748)	AGGCGAATGTCCCATGTGAG	GGTGGCAGCAGGTAGTTTCT
NF1-iv	NF1	1677	CACGGAGGTTCAAAACTGGT	TGCTAATGCCAAACAGCAAG
NF1-v	NF1	3402	AGGTCCGCTCTCCCTTAGAG	GCAGGCTGACCAGTCTTTTC
NF1-vi	NF1	2526	AGCCATTTGCACAGAGCTCT	ACCTACAAACCTGGGAGGGT
CCDC6-RET	CCDC6-RET (in-frame)	1401	GCAGCAAGAGAACAAGGTGC	ACCATCCTAAGTTGCTGGGC
EML4-ALK	EML4-ALK (in-frame)	2119 (1944)	CCGGCAGTCTCGATGATAGT	GCAACGTTAGGTGGGACAGT
GUCY1A2-PIWIL4	GUCY1A2-PIWIL4 (in-frame)	2721	CTGCGGACCTCAGAATTAGC	CAACAACCATCACGTCCTTG
SCAMP2-WDR72	SCAMP2-WDR72 (in-frame)	1080	GCGGACCCAGTGGATGTAAA	TGCCCCATCTTTAAGCCCAG
EFHD1-UBR3	EFHD1-UBR3 (in-frame)	1259	GGTCTTCAACCCCTACACGG	GCAATGCTGAGGTGGAGAGT
ERGIC2-CHRNA6	ERGIC2-CHRNA6 (frame-shift)	1761	TCGCGATATTTCCGGGTACG	ACCACCATGGCCACGTATTT
SIL1-MZB1	SIL1-MZB1 (5'UTR, in-frame)	640	GAGCCCTGGTGTGTTTCATT	TGTGGCTGACACCTTCTCTG
EGFR-iii	EGFR	1313 896 (nested)	ACAACACCCTGGTCTGGAAG CTAAGATCCCGTCCATCGCC	TGCACTCAGAGAGCTCAGGA ACATATGGGTGGCTGAGGGA
KRAS-iii	KRAS	866 (990)	GGAGAGAGGCCTGCTGAAAA	ACTGCATGCACCAAAAACCC
EML4-ALK-E18:A20	EML4-ALK (in-frame)	1038	GCAGGTGGTTTGTTCTGGAT	TGTCTTCAGGCTGATGTTGC
KIF5B-RET-K23:R12	KIF5B-RET (in-frame)	1114	ACCTGCGCAAACTCTTTGTT	AGGCCGTCGTCATAAATCAG

B. PCR of genomic DNA templates.

Primer name	Target exon (UCSC hg38)	Target region (UCSC hg38)	Product size	Fwd sequence	Rev sequence
EGFR-g1	Exon 18 (123bp; chr7:55173921-55174043)	chr7:55173522-55174302	781	CTGTGCTGGAAGCCATGTTC	CACCCCATGGCAAGGTCAAT
EGFR-g2	Exon 19 (99bp; chr7:55174722-55174820)	chr7:55174366-55175042	677	TGGGCTCATCTTCGTTTGCT	TGTTTCCAGCCTTTTGGGGT
EGFR-g3	Exon 20 (186bp; chr7:55181293-55181478)	chr7:55180840-55181827	988	GTTCCTGATGTGCAGGGTCA	GTCCTGAATGGGGGAAGCAA
EGFR-g4	Exon 21 (156bp; chr7:55191719-55191874)	chr7:55191401-55192313	913	CAGCCTGGCAAGTCCAGTAA	AGAAGGACTCCATTGCTGCC

Supplementary Table S3. PCR and sequencing primers for the Sanger sequencing validation.

A. Primers for direct Sanger sequencing

Sequencing target	Sample	PCR primer			Sequencing primer	
		Name	Fwd sequence	Rev sequence	Name	Sequence
EGFR E746_A750del	PC-9, RERF-LC-Ad2 (WT)	*EGFR-s*	CTAAGATCCCGTCCATCGCC	ACATATGGGTGGCTGAGGGA	*EGFR-s-fw*	CTAAGATCCCGTCCATCGCC
EGFR T790M / L858R	H1975	*EGFR-s*	CTAAGATCCCGTCCATCGCC	ACATATGGGTGGCTGAGGGA	*EGFR-s-rv*	ACATATGGGTGGCTGAGGGA
KRAS G12S	A549, H2228 (WT)	*KRAS-s*	ACTGAATATAAACTTGTGGTAGTTGG	CCCTCCCCAGTCCTCATGTA	*KRAS-s-rv*	CCCTCCCCAGTCCTCATGTA
NRAS Q61R	H2347, H2228 (WT)	*NRAS-s*	GTTGGGAAAAGCGCACTGAC	AGGTGTGTTTGTGCTGTGGA	*NRAS-s-fw*	GTTGGGAAAAGCGCACTGAC
NF1 exon19 skipping	PC-7, LC2/ad (WT)	*NF1-s*	CGTACTCCTGGAGCCTCTCT	GCTGACAGGTGTATCTGCGT	*NF1-fw*	CGTACTCCTGGAGCCTCTCT
EML4-ALK variant 3	H2228	*EML4-ALK-s*	GCCCTCTTCACAACCTCTCC	TCGTCCTGTTCAGAGCACAC	*EML4-ALK-s-fw*	GCCCTCTTCACAACCTCTCC
CCDC6-RET	LC2/ad	*CCDC6-RET-s*	GGTGCTGAAGATAGAGCTGGA	AGAAGGTTGAAGAGCCGCTC	*CCDC6-RET-s-fw*	GGTGCTGAAGATAGAGCTGGA
GUCY1A2-PIWIL4	H322	*GUCY1A2-PIWIL4-s*	CTGGTGATGTAGCCCAGCAA	AAAGGATGGCACCGTCGAA	*GUCY1A2-PIWIL4-s-fw*	CTGGTGATGTAGCCCAGCAA
SCAMP2-WDR72	A549	*SCAMP2-WDR72-s*	GAGTGGACTTTGGCCTCTCC	GCCCCATCTTTAAGCCCAGA	*SCAMP2-WDR72-s-fw*	GCCCCATCTTTAAGCCCAGA
EFHD1-UBR3	PC-9	*EFHD1-UBR3-s*	CCGGGAGTTCCTGCTCATTT	CGCCTAAGATCCCGGTCTTC	*EFHD1-UBR3-s-fw*	CCGGGAGTTCCTGCTCATTT
ERGIC2-CHRNA6	H1437	*ERGIC2-CHRNA6-s*	CGGTAGGCTGGGACCATAAC	TGGGTGATGGCCACTTCAAA	*ERGIC2-CHRNA6-s-fw*	CGGTAGGCTGGGACCATAAC
SIL1-MZB1	H1437	*SIL1-MZB1-s*	CCCTGGTGTGTTTCATTGGC	CTCGAACTCCGTAGTCCTGC	*SIL1-MZB1-s-rv*	CTCGAACTCCGTAGTCCTGC
EGFR ex19del	LUAD005, LUAD006, LUAD007	*EGFR-s-c*	CCAACCAAGCTCTCTTGAGG	CTGCGGTGTTTTCACCAGTA	*EGFR-s-c-fw*	CCAACCAAGCTCTCTTGAGG
KRAS G12 mutations	LUAD003, LUAD004, LUAD008	*KRAS-s-c*	ACTGAATATAAACTTGTGGTAGTTGG	CCCTCCCCAGTCCTCATGTA	*KRAS-s-c-rv*	CCCTCCCCAGTCCTCATGTA
EML4-ALK E18:A20	LUAD001	*EML4-ALK-s-c*	GCAGGTGGTTTGTTCTGGAT	TGTCTTCAGGCTGATGTTGC	*EML4-ALK-s-c-fw*	GCAGGTGGTTTGTTCTGGAT
KIF5B-RET K23:R12	LUAD002	*KIF5B-RET-s-c*	ACCTGCGCAAACTCTTTGTT	AGGCCGTCGTCATAAATCAG	*KIF5B-RET-s-c-fw*	ACCTGCGCAAACTCTTTGTT

B. Primers for TA cloning and Sanger sequencing

Sequencing target	Clinical sample	PCR primer			Sequencing primer	
		Name	Fwd sequence	Rev sequence	Name	Sequence
EGFR T790M / L858R phasing	H1975	*EGFR-s*	CTAAGATCCCGTCCATCGCC	ACATATGGGTGGCTGAGGGA	*EGFR-s-rv*	ACATATGGGTGGCTGAGGGA
EML4-ALK variant 3a/b	H2228	*EML4-ALK-s*	GCCCTCTTCACAACCTCTCC	TCGTCCTGTTCAGAGCACAC	*M13-fw*	GTAAAACGACGGCCAGT

Supplementary Table S4. MinION sequencing statistics.

Amplicon mixture	Sequencing run	Number of reads			Average read length (2D)	Average QV (2D)
		Template	Complement	2D		
#1	#1	69,715	44,599	38,432	1,762	10.7
#2	#2	6,632	4,040	3,334	1,544	10.9
	#3	111,871	57,698	49,433	1,857	10.8
#3	#4	11,692	5934	4,716	1,910	10.0
	#5	36,620	17,470	13,145	2,104	10.3
Average		47,306	25,948	21,812	1,835	10.5

Supplementary Table S5. SNV detection concordance analysis.

A. All (39 SNVs; VAF > 10% in Illumina)

X* (Threshold of VAF)	TP	FP	FN		Precision	Recall
			Miscall	Unk†		
0.1	37	521	0	2	0.066	0.949
0.2	36	103	1	2	0.259	0.923
0.3	35	34	2	2	0.507	0.897
0.4	27	15	3	9	0.643	0.692
0.5	23	4	3	13	0.852	0.590
0.6	23	4	3	13	0.852	0.590
0.7	22	4	2	15	0.846	0.564
0.8	21	4	1	17	0.840	0.538
0.9	21	3	0	18	0.875	0.538
0.95	19	1	0	20	0.950	0.487
1.0	4	0	0	35	1.000	0.103

TP: true positive; FP: false positive; FN: false negative.

*Thresholds of variant allele frequencies in the MinION reads. The SNVs with VAF ≤ X in Illumina escaped from false positive detection.

†Unknown.

*†Please see the Method section and **Supplementary Fig. S2**.

B. Heterozygous SNVs (32 SNVs; VAF > 50% in Illumina)

X (Threshold of VAF)	TP	FP	FN		Precision	Recall
			Miscall	Unk		
0.1	30	521	0	2	0.054	0.938
0.2	30	103	0	2	0.226	0.938
0.3	30	34	0	2	0.469	0.938
0.4	25	15	0	7	0.625	0.781
0.5	23	4	0	9	0.852	0.719
0.6	23	4	0	9	0.852	0.719
0.7	22	4	0	10	0.846	0.688
0.8	21	4	0	11	0.840	0.656
0.9	21	3	0	11	0.875	0.656
0.95	19	1	0	13	0.950	0.594
1.0	4	0	0	28	1.000	0.125

C. Homozygous SNVs (23 SNVs; VAF > 75% in Illumina)

X (Threshold of VAF)	TP	FP	FN		Precision	Recall
			Miscall	Unk		
0.1	21	521	0	2	0.039	0.913
0.2	21	103	0	2	0.169	0.913
0.3	21	34	0	2	0.382	0.913
0.4	21	15	0	2	0.583	0.913
0.5	21	4	0	2	0.840	0.913
0.6	21	4	0	2	0.840	0.913
0.7	21	4	0	2	0.840	0.913
0.8	21	4	0	2	0.840	0.913
0.9	21	3	0	2	0.875	0.913
0.95	19	1	0	4	0.950	0.826
1.0	4	0	0	19	1.000	0.174

Supplementary Table S6. List of true positive SNPs and mutations.

Cell line	Primer	Chromosome	Position (UCSC hg19)	Base (genome)		Base (RNA)		Amino acid changes	Illumina RNA-Seq (true positive)			MinION				
									Number of reads		VAF (%)	Number of reads (without mismatches ±3bp)		VAF (%)	False negative detection	Low Depth
				Ref	Var	Ref	Var		Total	Variant		Total (A + T + C + G)	Variant			
PC-9	*EGFR-i*	chr7	55214348	C	T	C	T	N158N	346	346	100	1244	1244	100		
PC-9	*EGFR-i*	chr7	55266417	T	C	T	C	T903T	422	422	100	1321	1301	98		
PC-9	*EGFR-ii*	chr7	55273609	A	T	A	T	(3'UTR)	343	64	19	80	8	10		y
PC-9	*EGFR-ii*	chr7	55274084	T	C	T	C	(3'UTR)	550	549	100	56	55	98		y
H1975	*EGFR-i*	chr7	55233089	C	T	C	T	A613A	298	298	100	2250	2242	100		
H1975	*EGFR-i*	chr7	55238874	T	A	T	A	T629T	335	335	100	1916	1900	99		
H1975	*EGFR-i*	chr7	55249063	G	A	G	A	Q787Q	354	260	73	1911	1238	65		
H1975	*EGFR-i*	chr7	55249071	C	T	C	T	T790M	347	249	72	1934	1507	78		
H1975	*EGFR-i*	chr7	55259515	T	G	T	G	L858R	408	294	72	1790	1083	61		
H1975	*EGFR-i*	chr7	55266417	T	C	T	C	T903T	257	179	70	1904	1191	63		
RERF-LC-Ad2	*EGFR-i*	chr7	55214348	C	T	C	T	N158N	903	388	43	765	289	38		
RERF-LC-Ad2	*EGFR-i*	chr7	55229255	G	A	G	A	R521K	994	481	48	675	389	58		
RERF-LC-Ad2	*EGFR-i*	chr7	55238874	T	A	T	A	T629T	960	458	48	676	248	37		
RERF-LC-Ad2	*EGFR-i*	chr7	55266417	T	C	T	C	T903T	765	765	100	1033	1022	99		
RERF-LC-Ad2	*EGFR-i*	chr7	55268916	C	T	C	T	D994D	1321	661	50	1038	568	55		
RERF-LC-Ad2	*EGFR-ii*											282	159	56		
RERF-LC-Ad2	*EGFR-ii*	chr7	55274084	T	C	T	C	(3'UTR)	1148	586	51	201	72	36		
A549	*KRAS-ii*	chr12	25359841	T	C	A	G	(3'UTR)	27	27	100	233	219	94		
A549	*KRAS-ii*	chr12	25360138	T	C	A	G	(3'UTR)	48	48	100	758	755	100		
A549	*KRAS-i*	chr12	25361091	T	C	A	G	(3'UTR)	16	16	100	405	391	97		
A549	*KRAS-i*	chr12	25361756	C	A	G	T	(3'UTR)	36	36	100	440	435	99		
A549	*KRAS-i*	chr12	25362217	A	G	T	C	(3'UTR)	31	31	100	536	536	100		
A549	*KRAS-i*	chr12	25362465	G	A	C	T	(3'UTR)	35	35	100	168	80	48	y	
A549	*KRAS-i*	chr12	25362552	A	C	T	G	(3'UTR)	57	57	100	416	415	100		
A549	*KRAS-i*	chr12	25398285	C	T	G	A	G12S	22	21	95	262	257	98		
H2228	*KRAS-ii*	chr12	25359841	T	C	A	G	(3'UTR)	75	75	100	633	579	91		
H2228	*KRAS-ii*	chr12	25360138	T	C	A	G	(3'UTR)	150	150	100	1483	1468	99		
H2228	*KRAS-i*	chr12	25361091	T	C	A	G	(3'UTR)	59	59	100	604	594	98		
H2228	*KRAS-i*	chr12	25361685	G	A	C	T	(3'UTR)	179	81	45	556	300	54		
H2228	*KRAS-i*	chr12	25362217	A	G	T	C	(3'UTR)	106	106	100	791	791	100		
H2228	*KRAS-i*	chr12	25362465	G	A	C	T	(3'UTR)	141	141	100	234	107	46	y	
H2228	*KRAS-i*	chr12	25362552	A	C	T	G	(3'UTR)	147	146	99	552	552	100		
H2347	*NRAS-i*	chr1	115256529	T	C	A	G	Q61R	226	225	100	623	615	99		
PC-7	*NF1-ii*	chr17	29553485	G	A	G	A	P678P	20	20	100	1754	1682	96		
PC-7	*NF1-vi*	chr17	29703374	G	A	G	A	(3'UTR)	28	28	100	1556	1536	99		
PC-7	*NF1-vi*	chr17	29703438	C	G	C	G	(3'UTR)	20	20	100	1371	1338	98		
LC2/ad	*NF1-i*	chr17	29508775	G	A	G	A	L234L	96	53	55	921	618	67		
LC2/ad	*NF1-ii*	chr17	29553485	G	A	G	A	P678P	110	34	31	1187	301	25		
LC2/ad	*NF1-vi*	chr17	29703374	G	A	G	A	(3'UTR)	144	44	31	1644	525	32		
LC2/ad	*NF1-vi*	chr17	29703438	C	G	C	G	(3'UTR)	128	41	32	1978	304	15		
LC2/ad	*NF1-vi*	chr17	29704002	T	C	T	C	(3'UTR)	170	113	66	1273	764	60		

Supplementary Table S7. Number of MinION reads aligned to fusion genes.

Fusion transcript	Alignment to mRNAs		Alignment to fusion RNAs		
	Total*	Split alignment†	Total	Junction covered	
				±10 bp	±50 bp
CCDC6-RET	782	341 (44%)	718	464 (65%)	451 (63%)
EML4-ALK	742	540 (73%)	558	544 (97%)	533 (96%)
GUCY1A2-PIWIL4	6401	4433 (69%)	6621	5249 (79%)	5185 (78%)
SCAMP2-WDR72	316	244 (77%)	311	262 (84%)	256 (82%)
EFHD1-UBR3	562	460 (82%)	559	521 (93%)	507 (91%)
ERGIC2-CHRNA6	820	529 (65%)	786	538 (68%)	530 (67%)
SIL1-MZB1	620	191 (31%)	556	411 (74%)	129 (23%)

*We selected reads aligned to either of the fusion partners.

†We selected split alignments to two regions and discarded reads aligned to three or more regions.

Supplementary Table S8. Summary of the EGFR-diluted sequencing analysis.

Sequencing run	MUT : WT*	Number of 2D reads			Sequence identity (avg.)	%VAF‡			Expected % VAF**
		Total	Aligned	On-target		T790M	L858R	Double	
#D1	1: 1	23,948	18,163 (76%)	16,650 (70%)	85%	22.3%	17.9%	22.5%	35.8%
#D2	1: 4	36,038	30,636 (85%)	29,025 (81%)	83%	9.9%	8.4%	8.7%	14.3%
#D3	1: 9	22,417	14,445 (64%)	13,804 (62%)	80%	4.1%	4.2%	7.6%	7.2%
#D4	1: 19	43,328	35,807 (83%)	34,680 (80%)	82%	4.9%	4.3%	4.1%	3.6%
#D5†	1: 99	59,907	48,117 (80%)	46,057 (77%)	85%	0.09%	0.13%	0.06%	0.7%

*MUT: H1975, WT: RERF-LC-Ad2.

†Three sequencing runs were merged.

‡The reads without mismatches ±3 bp of the SNV were used in the calculation of the VAFs.

**H1975 transcribed mutant alleles in 71.6% of the EGFR RNAs (calculated using the MinION sequencing result of H1975 *EGFR-i*).

Supplementary Table S9. Summary of the fusion transcripts in the clinical samples.

A. Analysis of the fusion transcripts.

Fusion transcript	Alignment to mRNAs		Alignment to fusion RNAs		
	Total*	Split alignment†	Total	Junction covered	
				±10 bp	±50 bp
EML4-ALK E18;A20	2,296	1,219 (53%)	2,165	1,803 (83%)	1,718 (79%)
KIF5B-RET K23;R12	44,994	26,023 (58%)	44,412	31,331 (71%)	30,524 (69%)

*We selected reads aligned to either of the fusion partners.

†We selected split alignments to two regions and discarded reads aligned to three or more regions.

B. Concordance of SNV detection in the fusion transcripts.

X (Threshold of VAF)	TP	FP	FN		Precision	Recall
			Miscall	Unk		
0.1	2	4	0	0	0.333	1.000
0.2 - 0.95	2	0	0	0	1.000	1.000
1.0	1	0	1	0	1.000	0.500

Note: Similarly to **Supplementary Fig. S2** and **Supplementary Table S5**, six SNV candidates were detected by the MinION sequencing of EML4-ALK and KIF5B-RET when the VAF threshold "X" set to 0.1. Two SNVs were verified as true positive SNVs using Illumina RNA-Seq or direct Sanger sequencing. One of the two SNVs was detected only in Sanger sequencing because the Illumina RNA-Seq had small depths in the position of the SNV.

References

1. Suzuki, A., Makinoshima, H., Wakaguri, H., et al. 2014, Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic Acids Res*, **42**, 13557-13572.
2. Frith, M. C., Park, Y., Sheetlin, S. L. and Spouge, J. L. 2008, The whole alignment and nothing but the alignment: the problem of spurious alignment flanks. *Nucleic Acids Res*, **36**, 5863-5871.
3. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
4. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. and Frith, M. C. 2011, Adaptive seeds tame genomic sequence comparison. *Genome Res*, **21**, 487-493.
5. Hamada, M., Ono, Y., Asai, K. and Frith, M. C. 2016, Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics*.
6. Kohno, T., Ichikawa, H., Totoki, Y., et al. 2012, KIF5B-RET fusions in lung adenocarcinoma. *Nat Med*, **18**, 375-377.