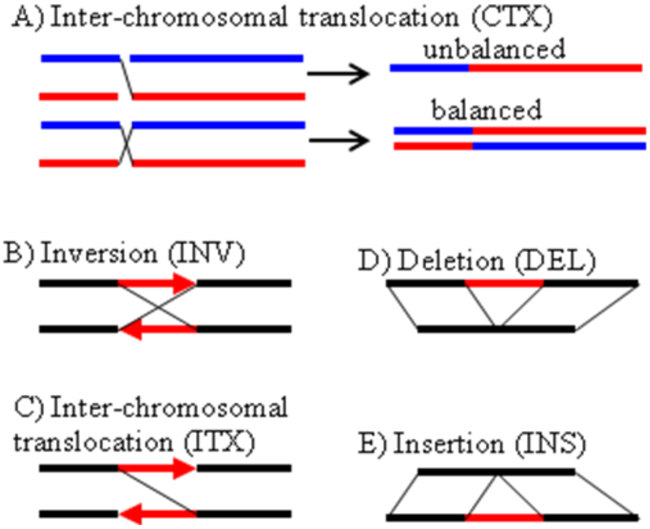
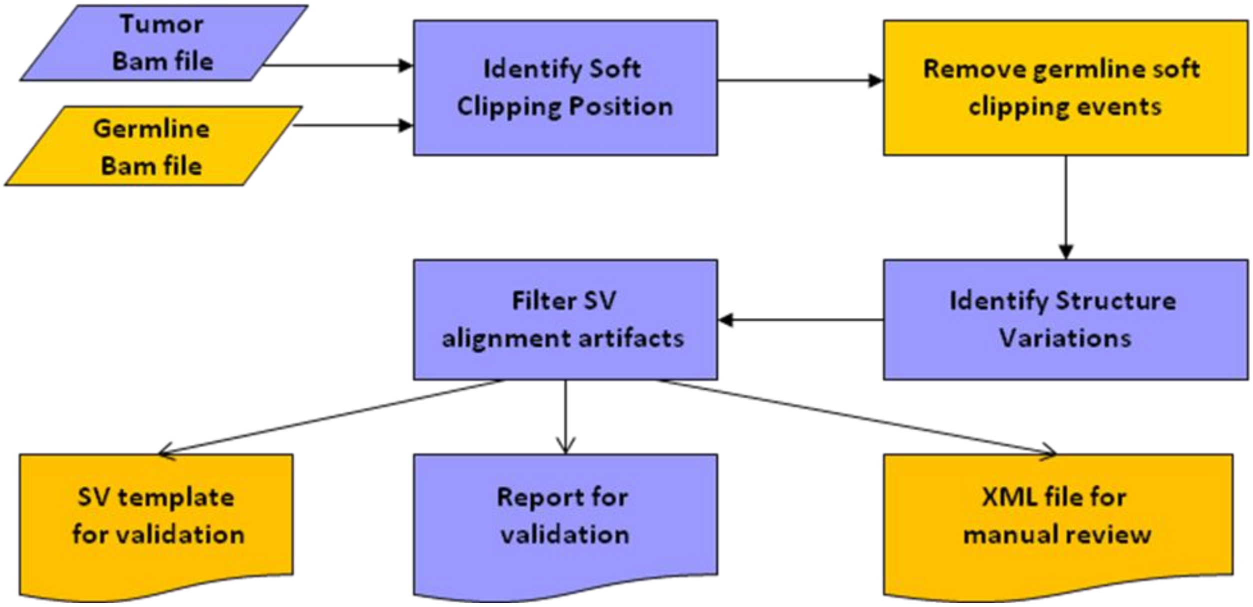
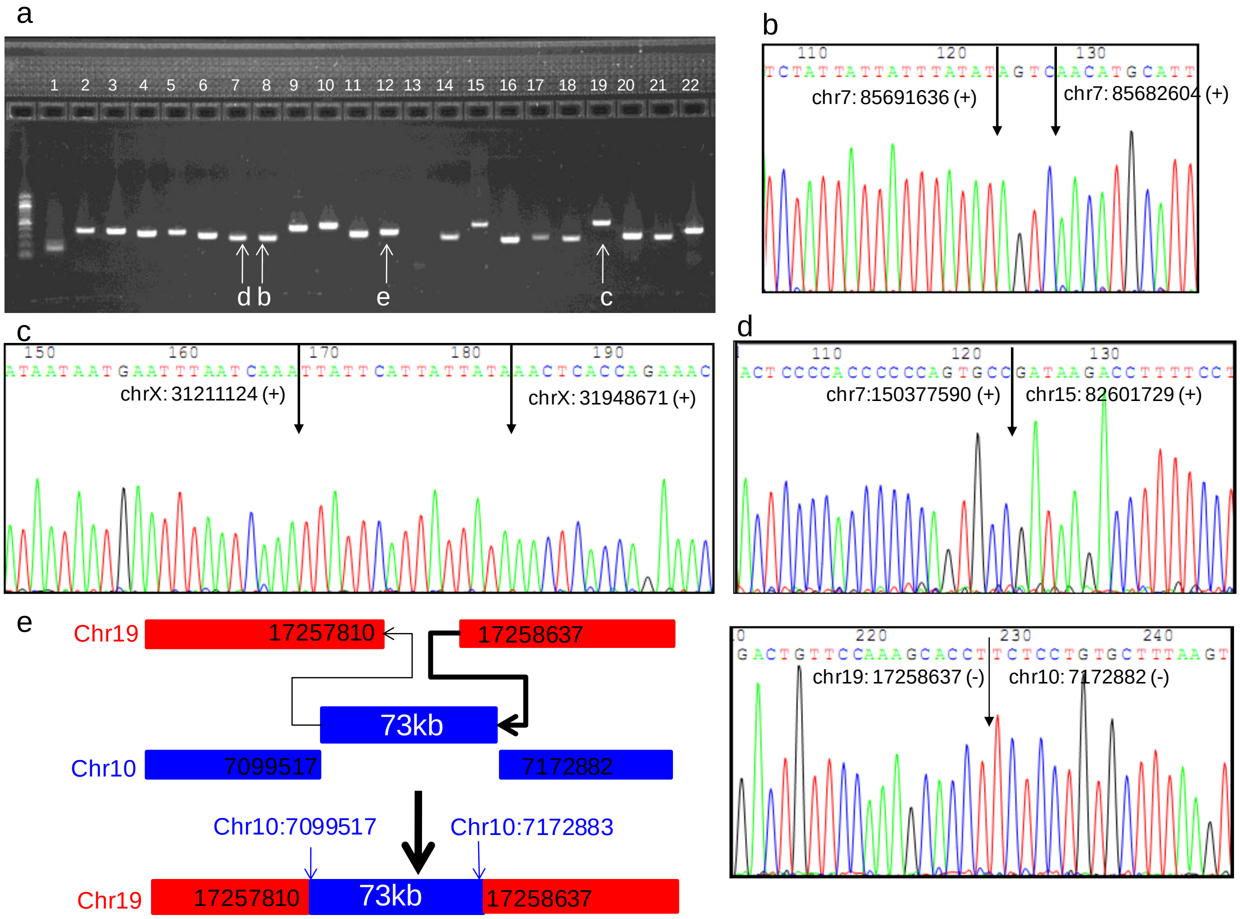
Supplementary Figure 1. Definition of structural variations in CREST analysis. A) Inter-chromosomal translocation (CTX) has two breakpoints located on two different chromosomes (represented in red and blue). A balanced translocation generates two products with reciprocal pattern of translocation while an unbalanced translocation only shows one product. B-D) Intra-chromosomal structural variations where the two breakpoints are located on the same chromosome. The reference genome representing the wild-type is shown at the top while the altered genome found in a sample is shown at the bottom. B) Inversion (INV) has reciprocal join in opposite orientations. C) Intra-chromosome translocation (ITX) has unilateral join in opposite orientation. D) Deletion (DEL) has two breakpoints joined in ascending order of genomic coordinates in the same orientation. E) Insertion (INS) has two breakpoints joined in descending order of genomic coordinates in the same orientation.


A) Inter-chromosomal translocation (CTX)
unbalanced
balanced
B) Inversion (INV)
D) Deletion (DEL)
C) Inter-chromosomal translocation (ITX)
E) Insertion (INS)

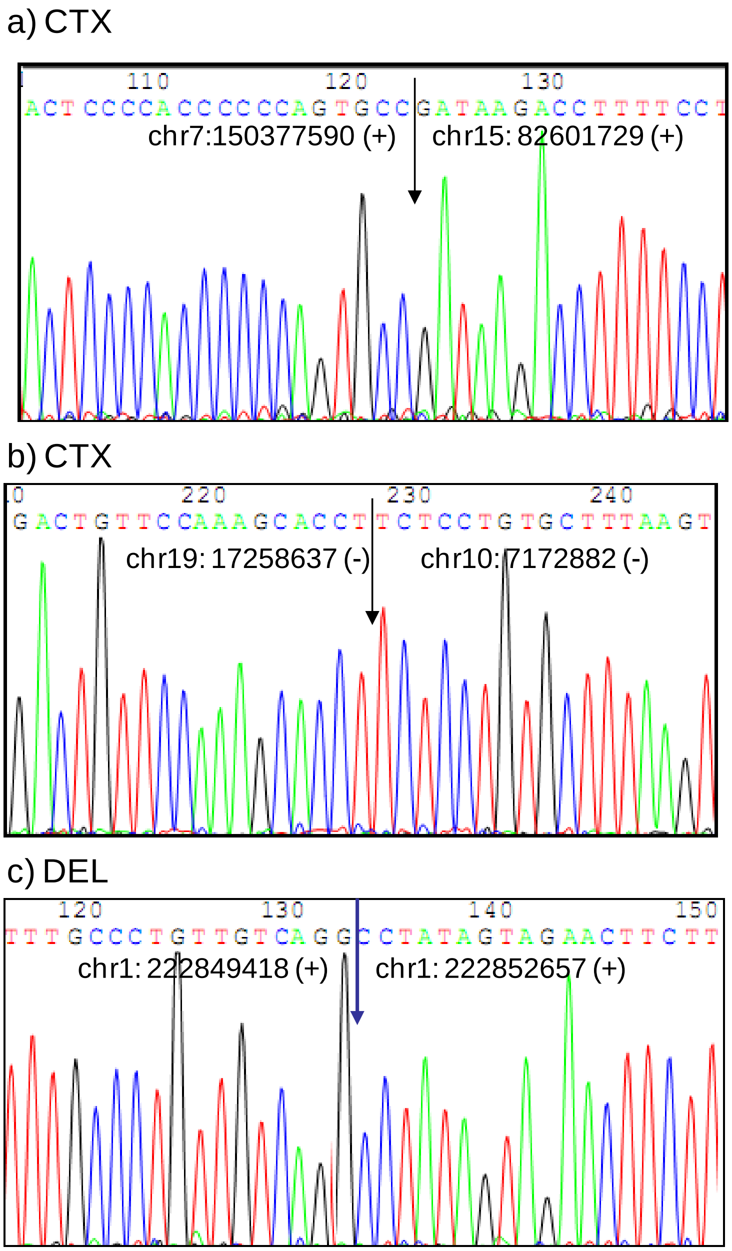
Supplementary Figure 2. Process flow of CREST. The objects in yellow are optional. Specifically, yellow parallelogram, yellow rectangle and yellow box represent optional input, process and output files respectively. Removal of germline soft-clipping events is an optional process triggered only when the input consists of two paired bam files. Three report files are generated including one XML file for manual review. Supplementary Figure 5 shows an example of XML file. A user may decide to remove a SV if it is found by low-complexity soft-clipped reads as the majority of false positive SVs appear to be in regions of low-complexity based on our experimental validation results.


Tumor Bam file
Germline Bam file
Identify Soft Clipping Position
Remove germline soft clipping events
Filter SV alignment artifacts
Identify Structure Variations
SV template for validation
Report for validation
XML file for manual review

Supplementary Figure 3 Validation of 20 putative SVs detected only by CREST but not reported previously[1] in the melanoma cancer cell line COLO-829. (a) PCR amplification of the 20 SVs (lane 1-20) and 2 SVs identified by Pleasance et al, lane 21 and 22. PCR products of the expected size were not generated for only two of the analyzed SV (lanes 1 and 13). The arrows point to the SVs listed in (b) to (e). (b) An INS SV on chromosome 7 resulting from a tandem duplication of 9,032bp. The arrows define the exact breakpoints while the 4 bases between the two arrows are non-template sequence inserted in the re-arrangement. (c) A 737,547bp deletion on chromosome X with a 15bp of non-template sequence. (d) A CTX between chromosomes 7 and 15. (e) An inter-chromosomal insertion revealed by combining a novel SV (right, connected with thick line) with a known SV (left, connected with thin line). The Sanger sequencing data that confirm the novel SV is shown at the right.



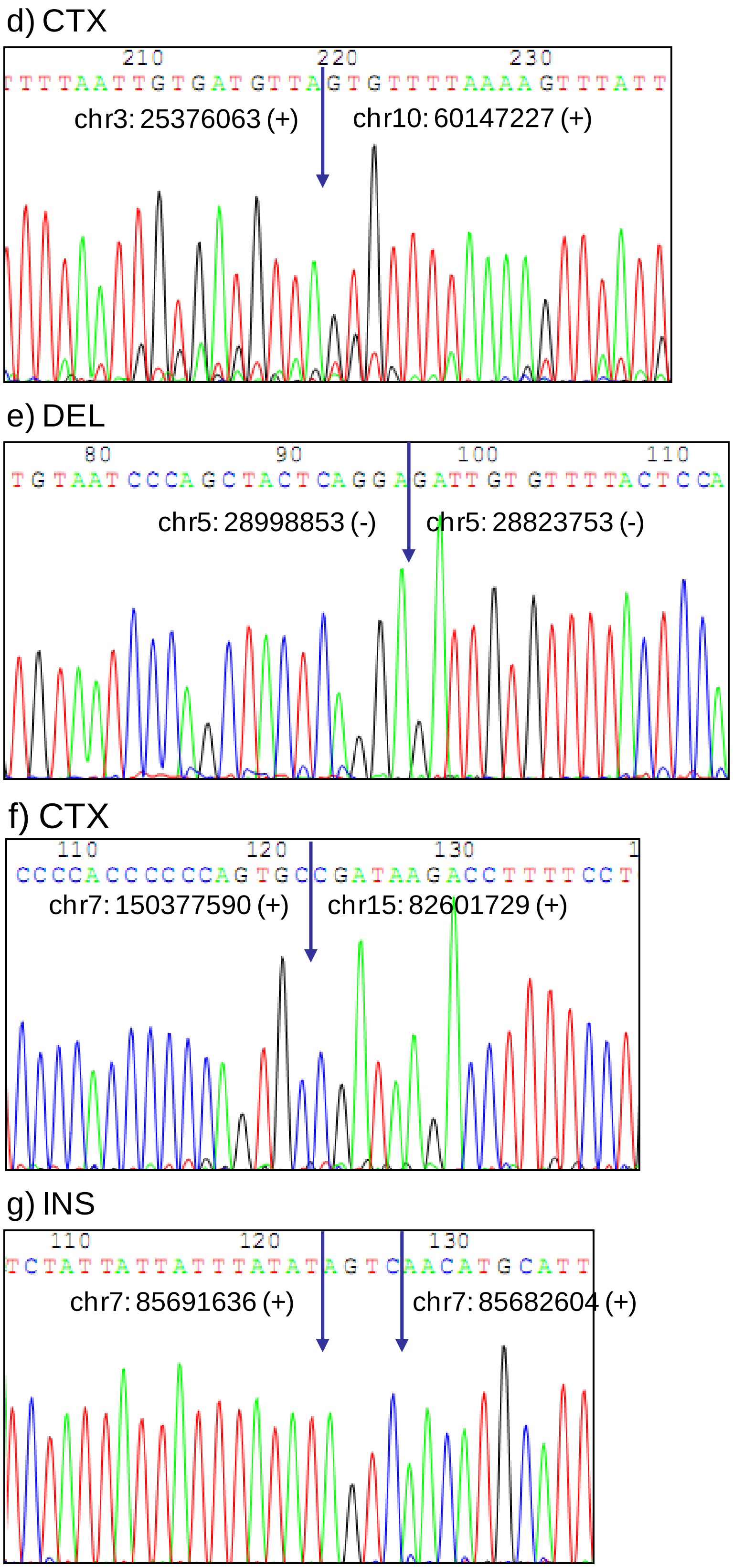
Supplementary Figure 4 Sanger sequencing data for the 18 validated novel SVs in COLO-829 cell line



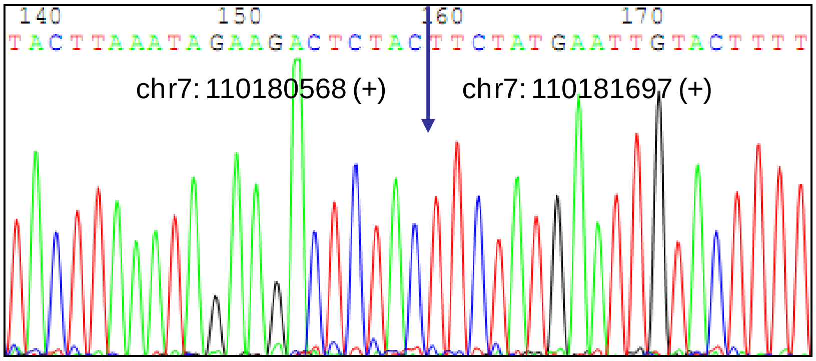
a) CTX

b) CTX

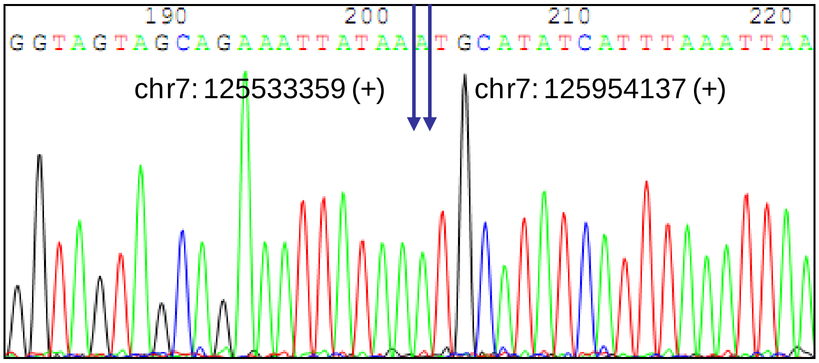
c) DEL


d) CTX
210
220
230
TTTTAATTGTGATGTTAGTGTTTTAAAAGTTTATT
chr3: 25376063 (+)
chr10: 60147227 (+)
e) DEL
80
90
100
110
TGTAATCCCAGCTACTCAGGAGATTGTGTTTTACTCCA
chr5: 28998853 (-)
chr5: 28823753 (-)
f) CTX
110
120
130
1
CCCCACCCCCCAGTGCCGATAAGACCTTTTCCT
chr7: 150377590 (+)
chr15: 82601729 (+)
g) INS
110
120
130
TCTATTATTATTTATATAGTCAACATGCATT
chr7: 85691636 (+)
chr7: 85682604 (+)

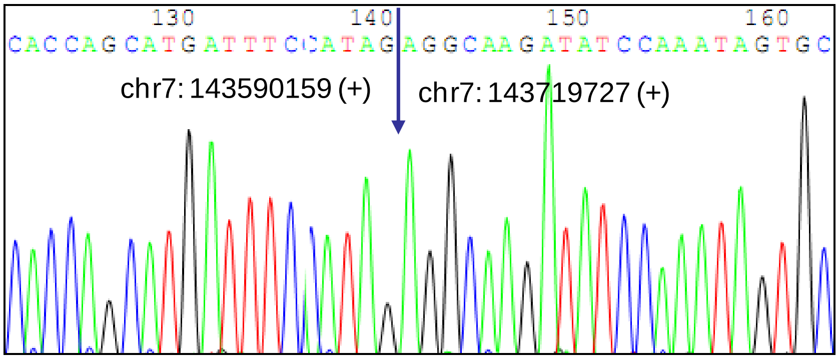
h) DEL



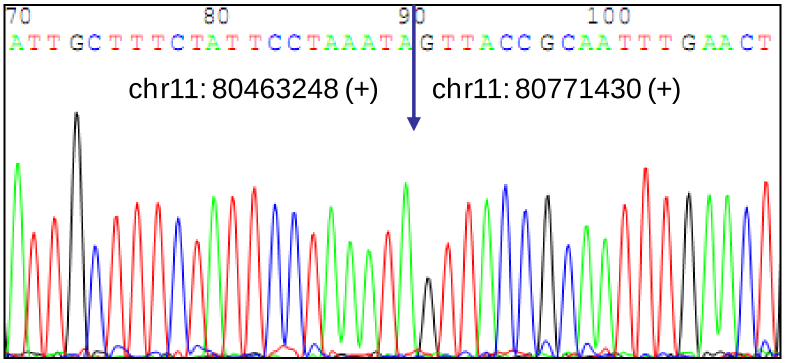
i) DEL



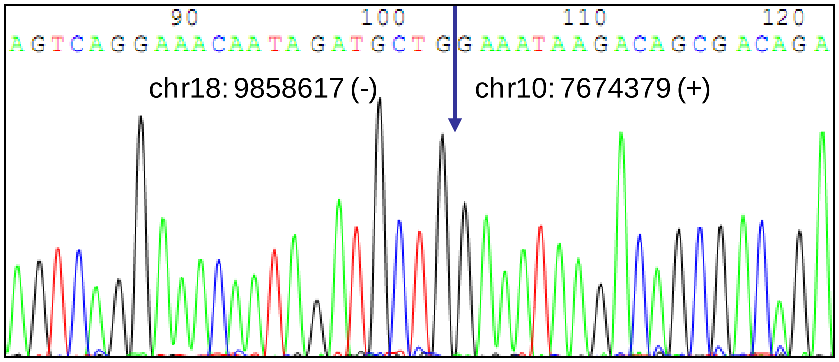
j) DEL



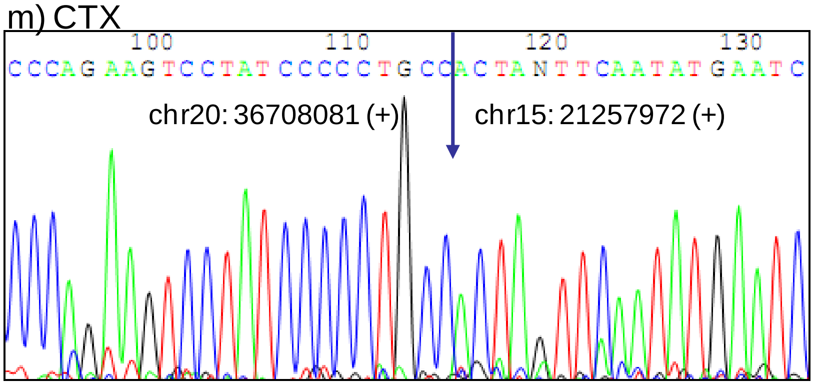
k) DEL



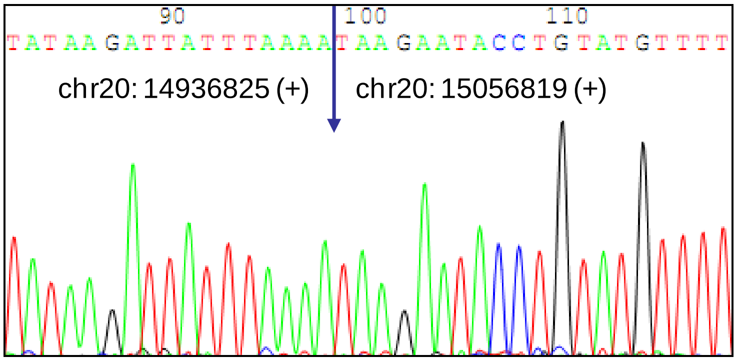
l) CTX



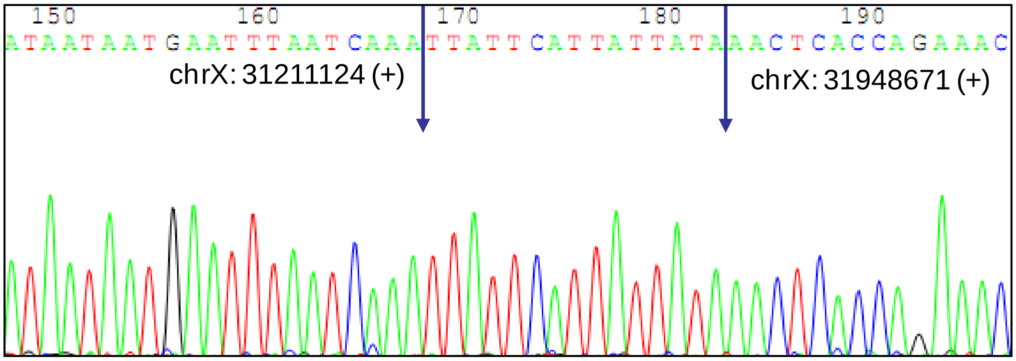
m) CTX



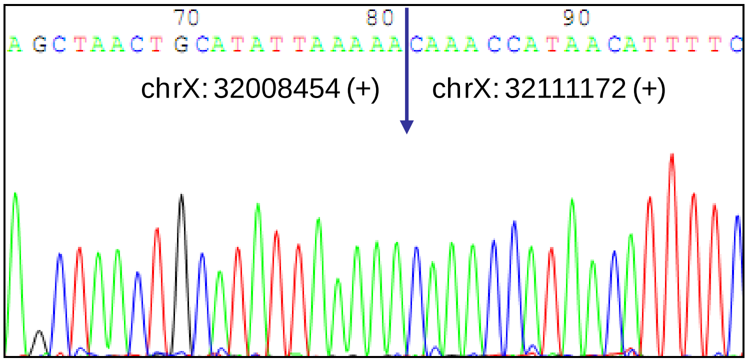
n) DEL



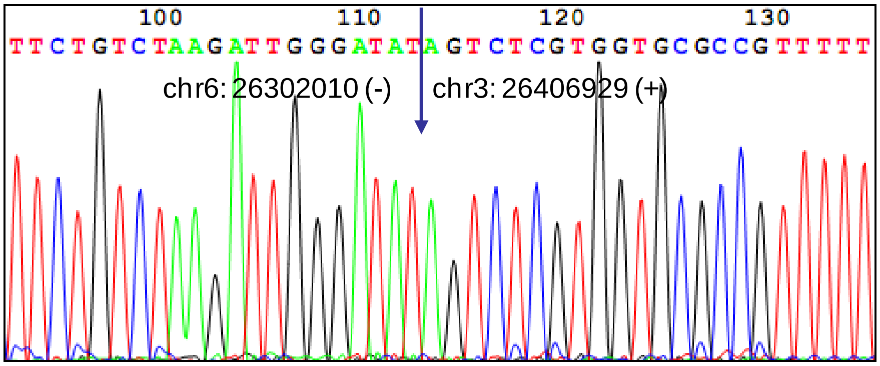
o) DEL



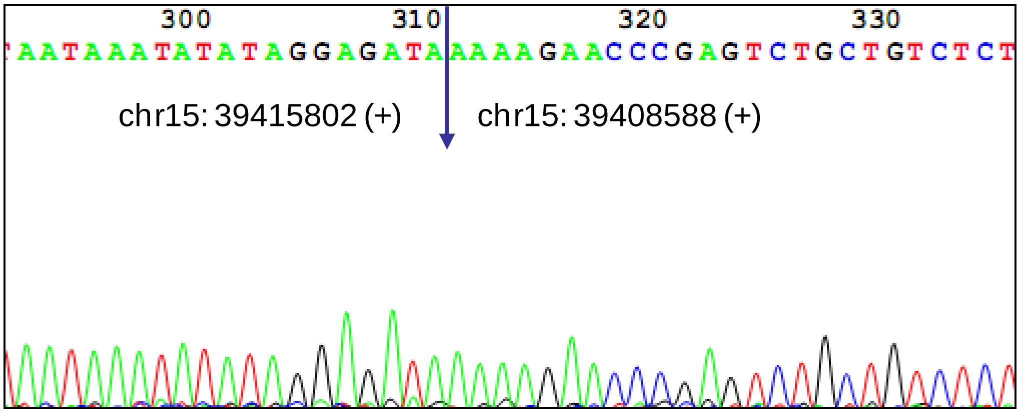
p) DEL



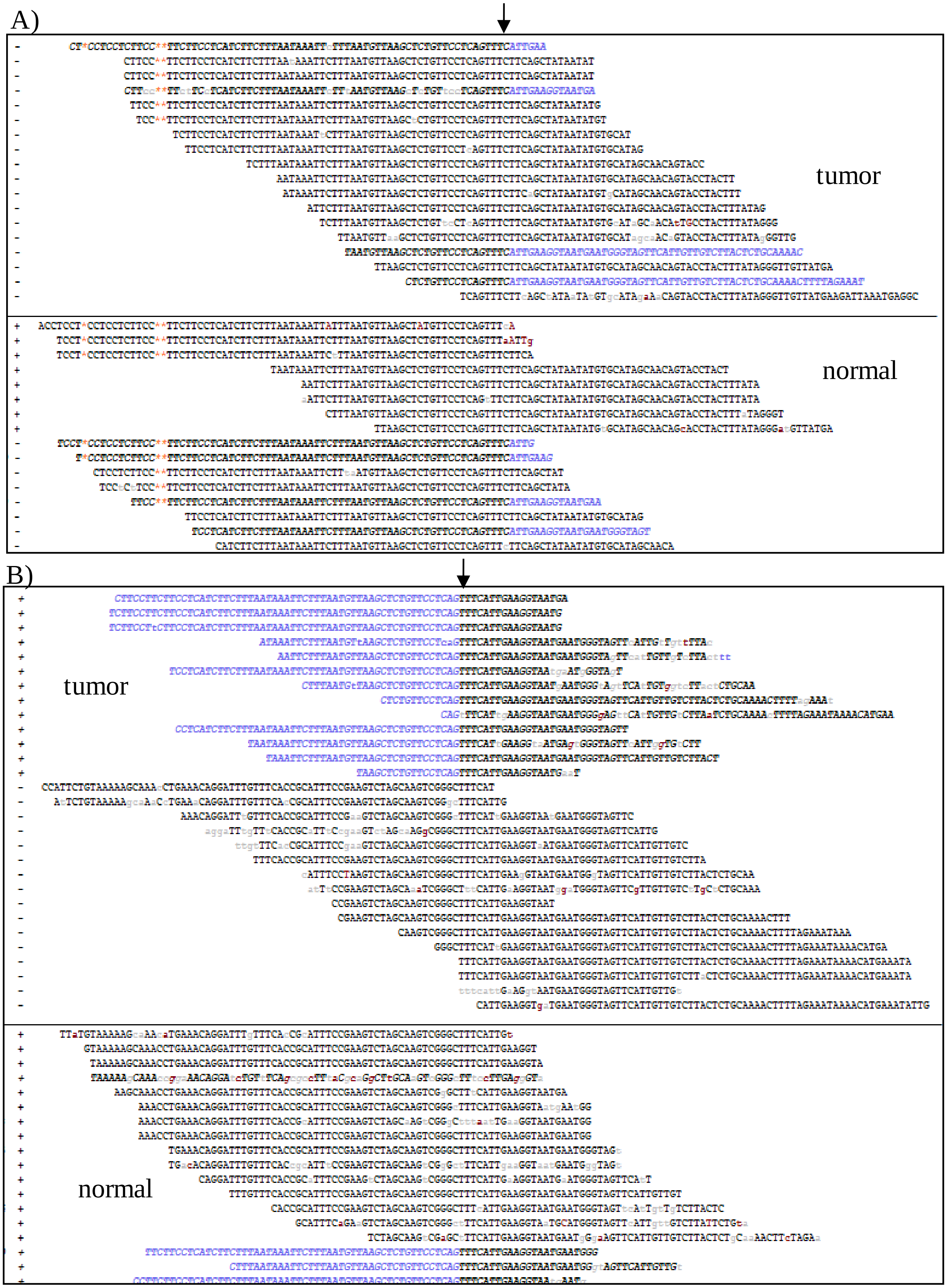
q) CTX



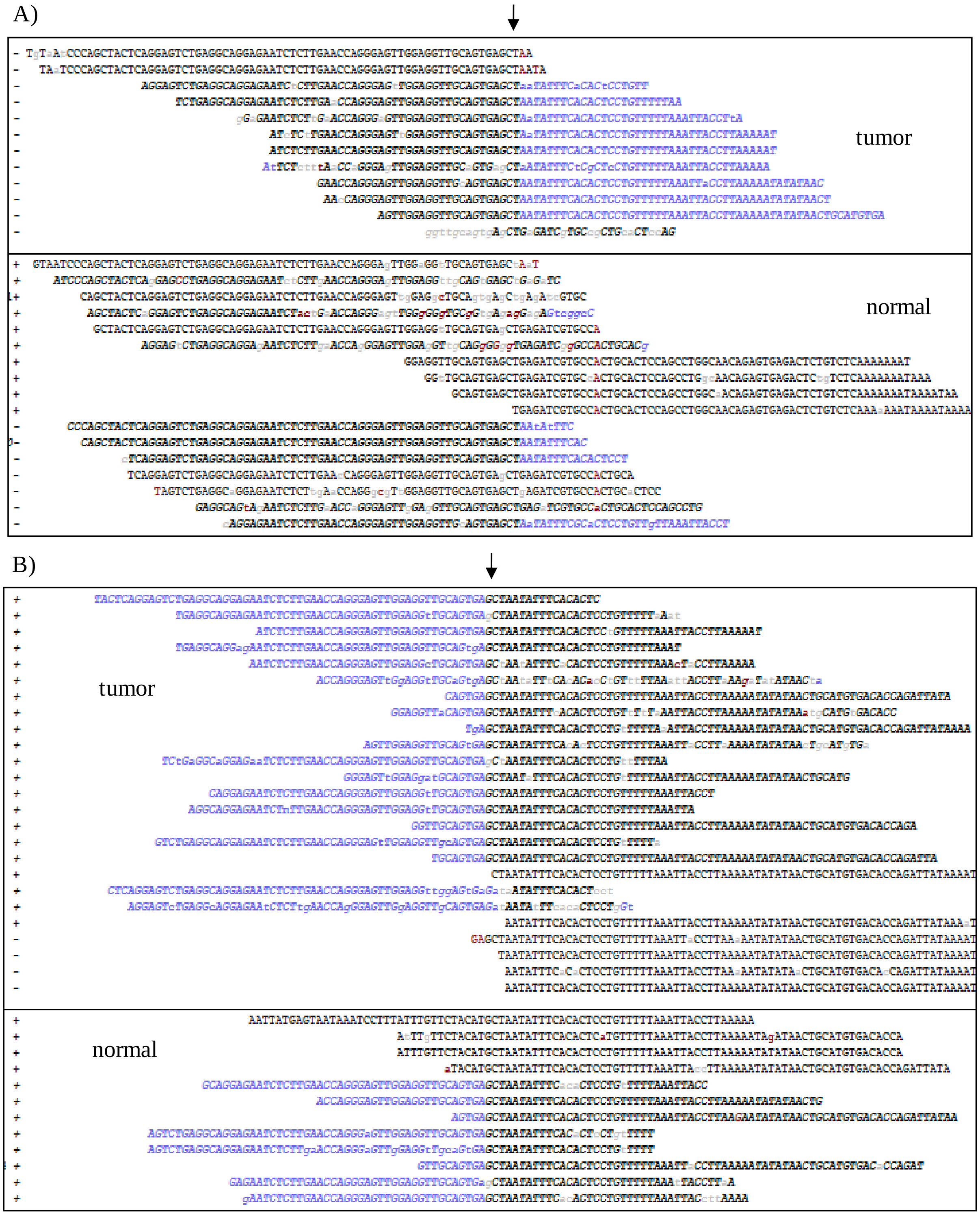
r) INS



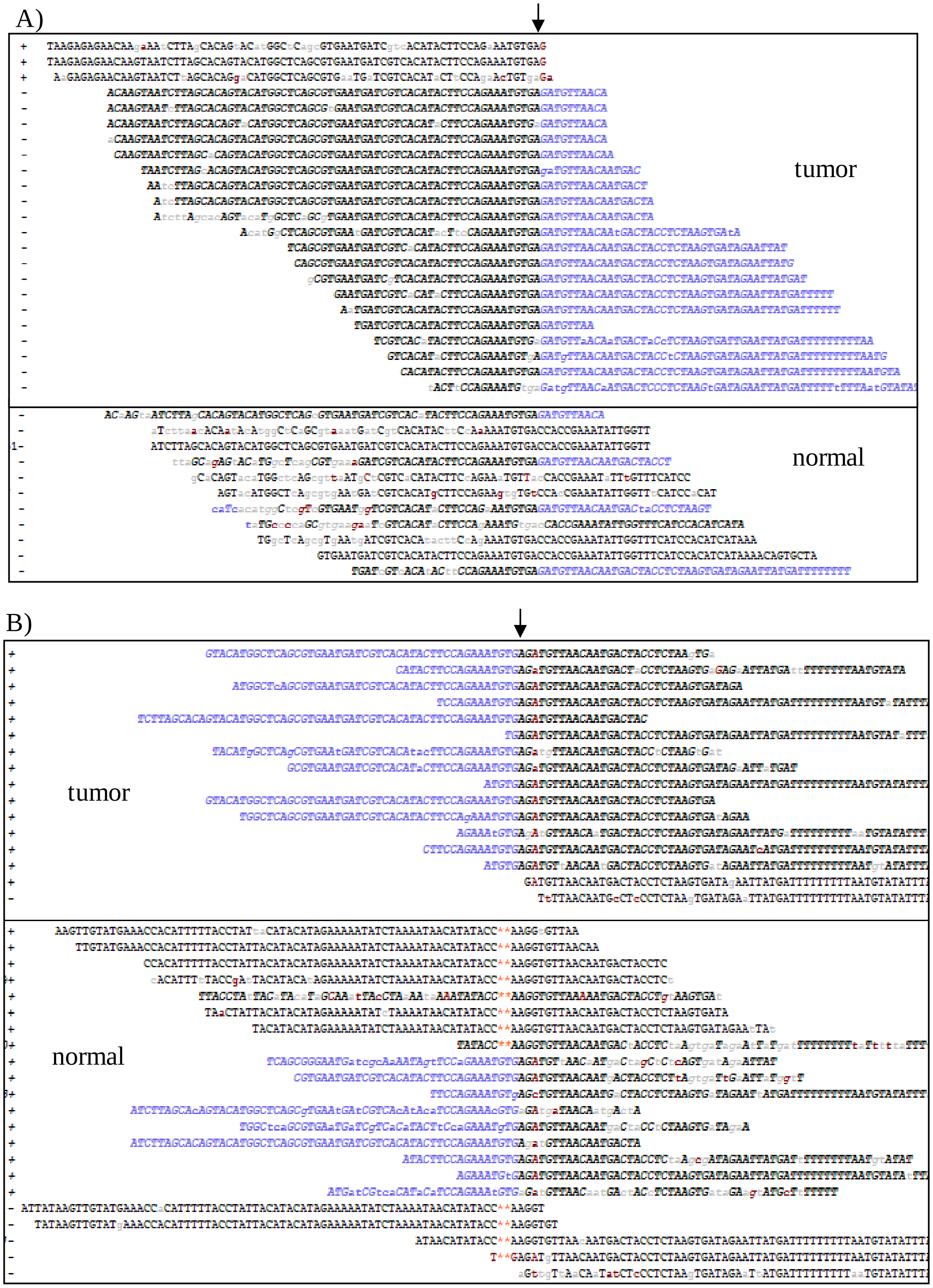
Supplementary Figure 5 Alignment of next-gen reads in both tumor (top panel) and normal (bottom panel) of cell line COLO-829 at deletion spanning chr6:51307435-51308086 reported by Pleasance et al. The center of the alignment is indicated by an arrow. The “+” or “-“ sign at the beginning indicates the sequence alignment orientation for each read. Mismatches or indels to the reference genome are shown in brown. Soft-clipping reads are displayed in two segments: the segment that matches the reference genome is shown in black and bold letters while the soft-clipping segment is shown in blue and italic letters. Low-quality bases (<20 phred score) are shown in lower case and in gray if they match the reference genome. A) Alignment centered at the first breakpoint chr6: 1307435. The soft-clipping segments found in both tumor and normal match the sequence at the second breakpoint. B) Alignment centered at the second breakpoint chr6:51308086. The soft-clipping segments found in both tumor and normal match the sequence at the first breakpoint indicating that the SV is of germline origin.


A)
tumor
normal
B)
tumor
normal

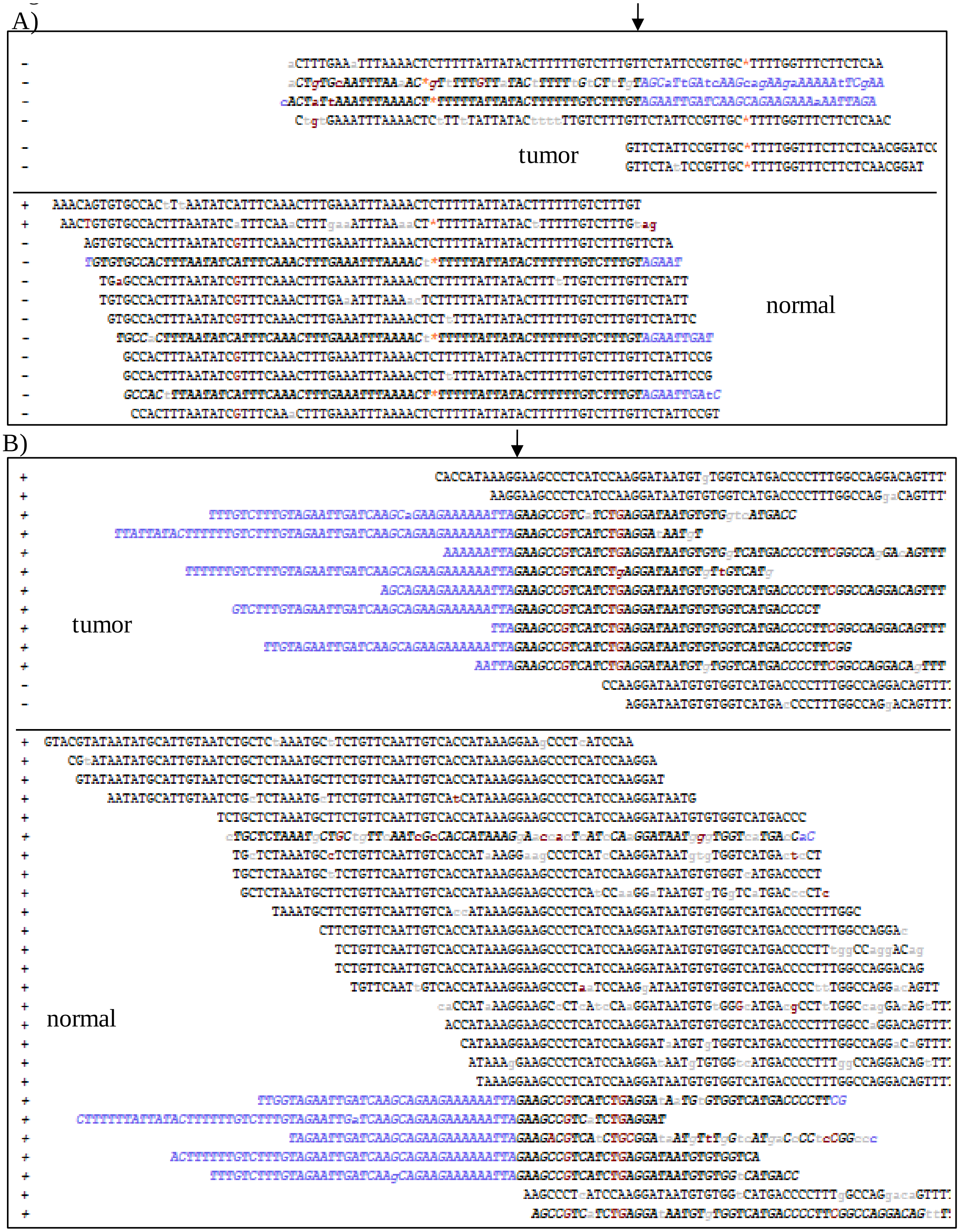
Supplementary Figure 6 Alignment of next-gen reads in both tumor (top panel) and normal (bottom panel) of cell line COLO-829 at deletion spanning chr14:48401118-48403688 reported by Pleasance et al. A) Alignment centered at the first breakpoint chr14:48401118. B) Alignment centered at the second breakpoint chr14:48403688.



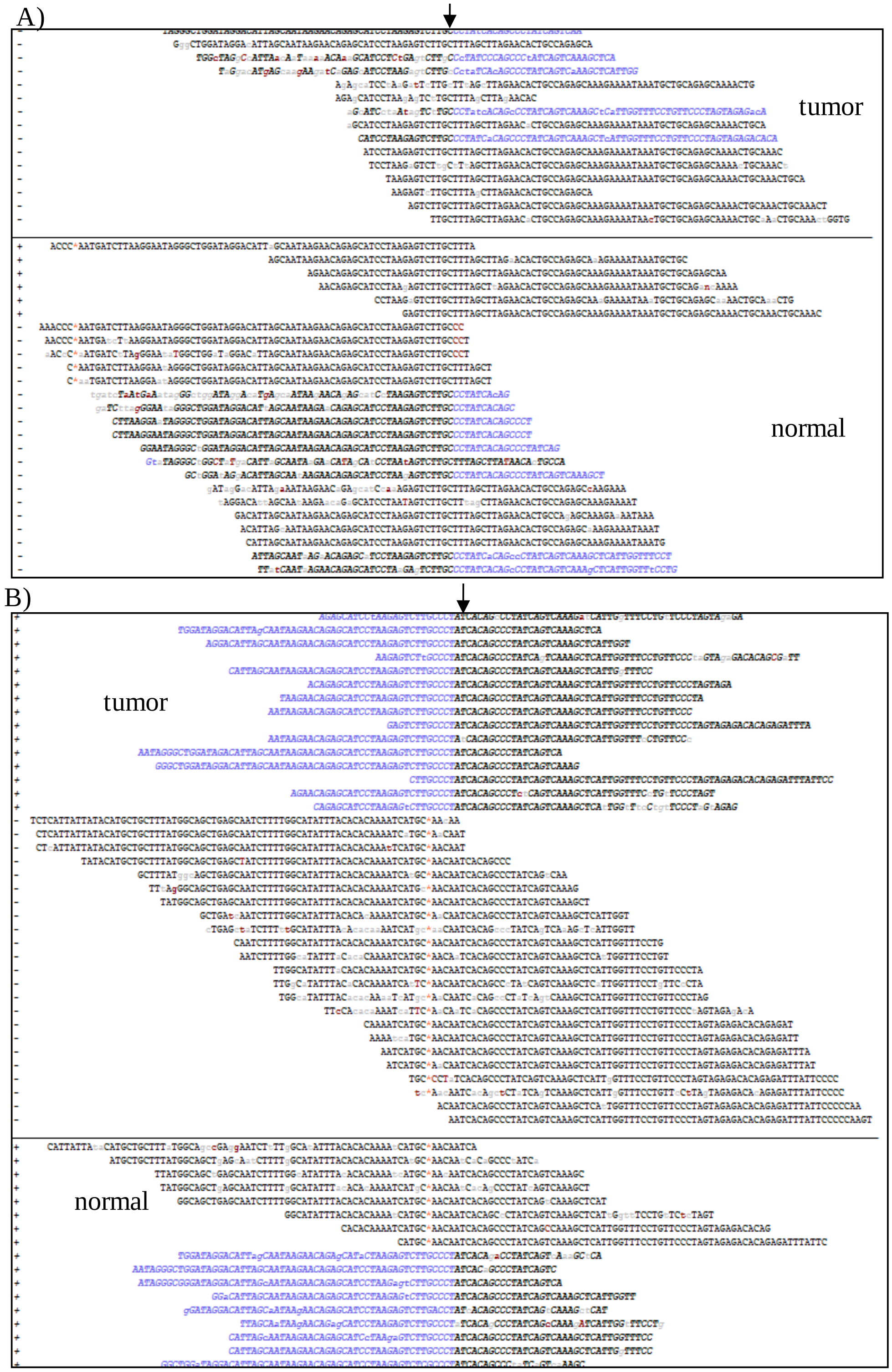
Supplementary Figure 7 Alignment of next-gen reads in both tumor (top panel) and normal (bottom panel) of cell line COLO-829 at deletion spanning chr14:33583777-33584588 reported by Pleasance et al. The deletion was also found in dbSNP (rs72415809). A) Alignment centered at the first breakpoint chr14:33583777. B) Alignment centered at the second breakpoint chr14:33584588



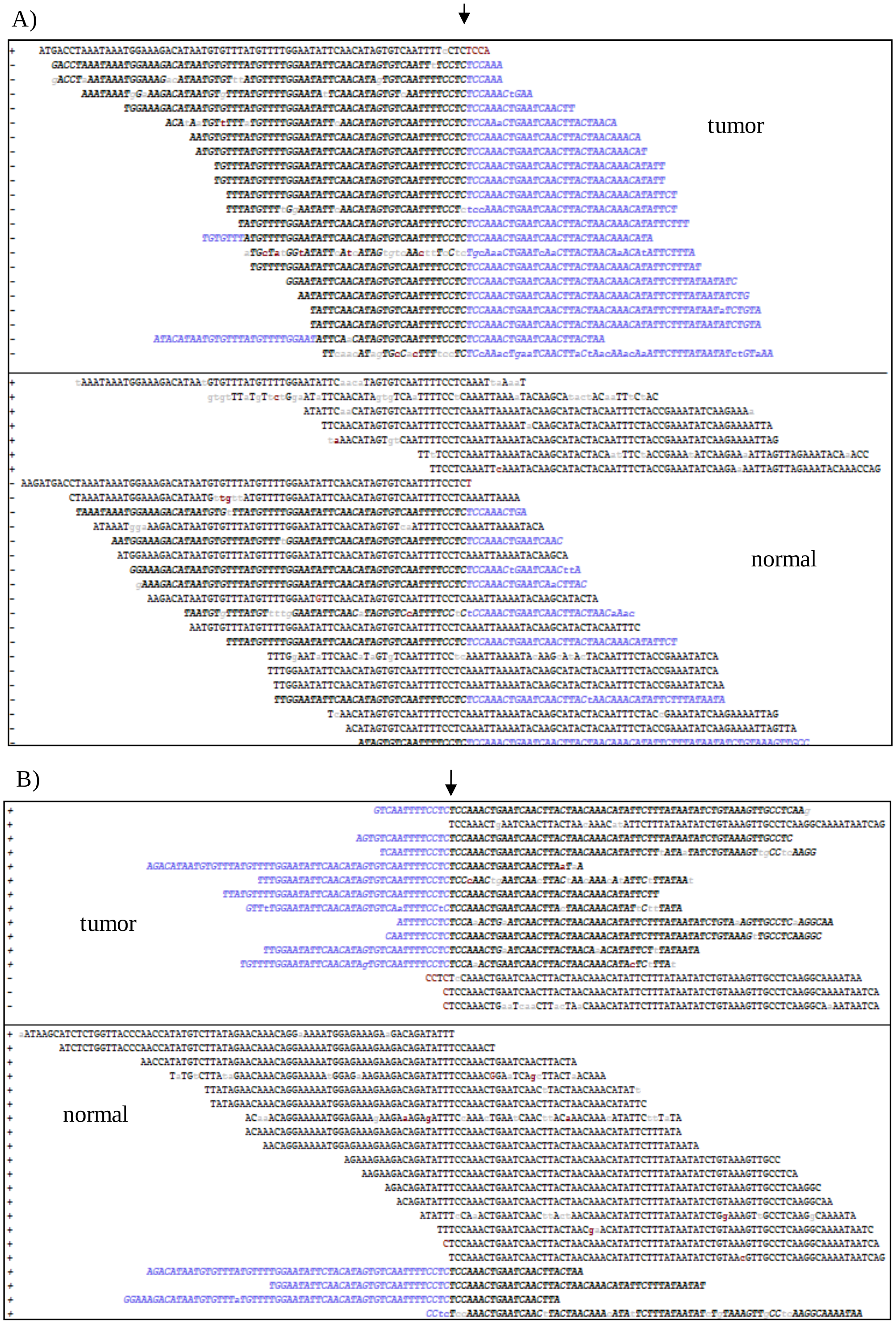
Supplementary Figure 8 Alignment of next-gen reads in both tumor (top panel) and normal (bottom panel) of cell line COLO-829 at inter-chromosomal alteration spanning chr1:16797227-145856276 reported by Pleasance et al. A) Alignment centered at the first breakpoint chr1:16797227. B) Alignment centered at the second breakpoint chr1:145856276



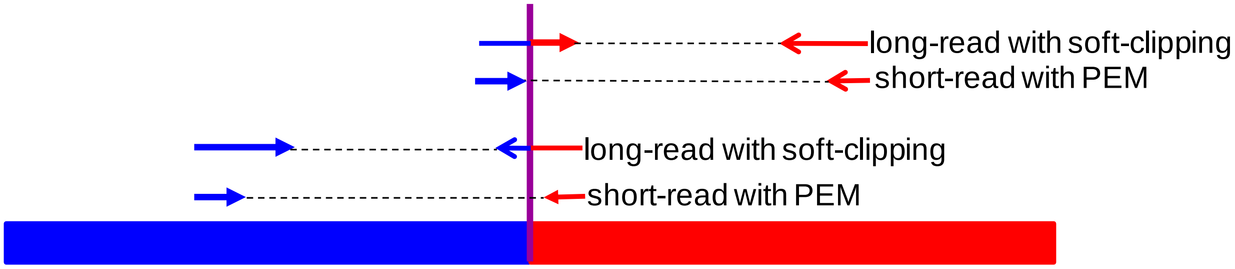
Supplementary Figure 9 Alignment of next-gen reads in both tumor (top panel) and normal (bottom panel) of cell line COLO-829 at deletion spanning chr4:131181342-131224000 reported by Pleasance et al. A) Alignment centered at the first breakpoint chr4:131181342. B) Alignment centered at the second breakpoint chr4:131224000.



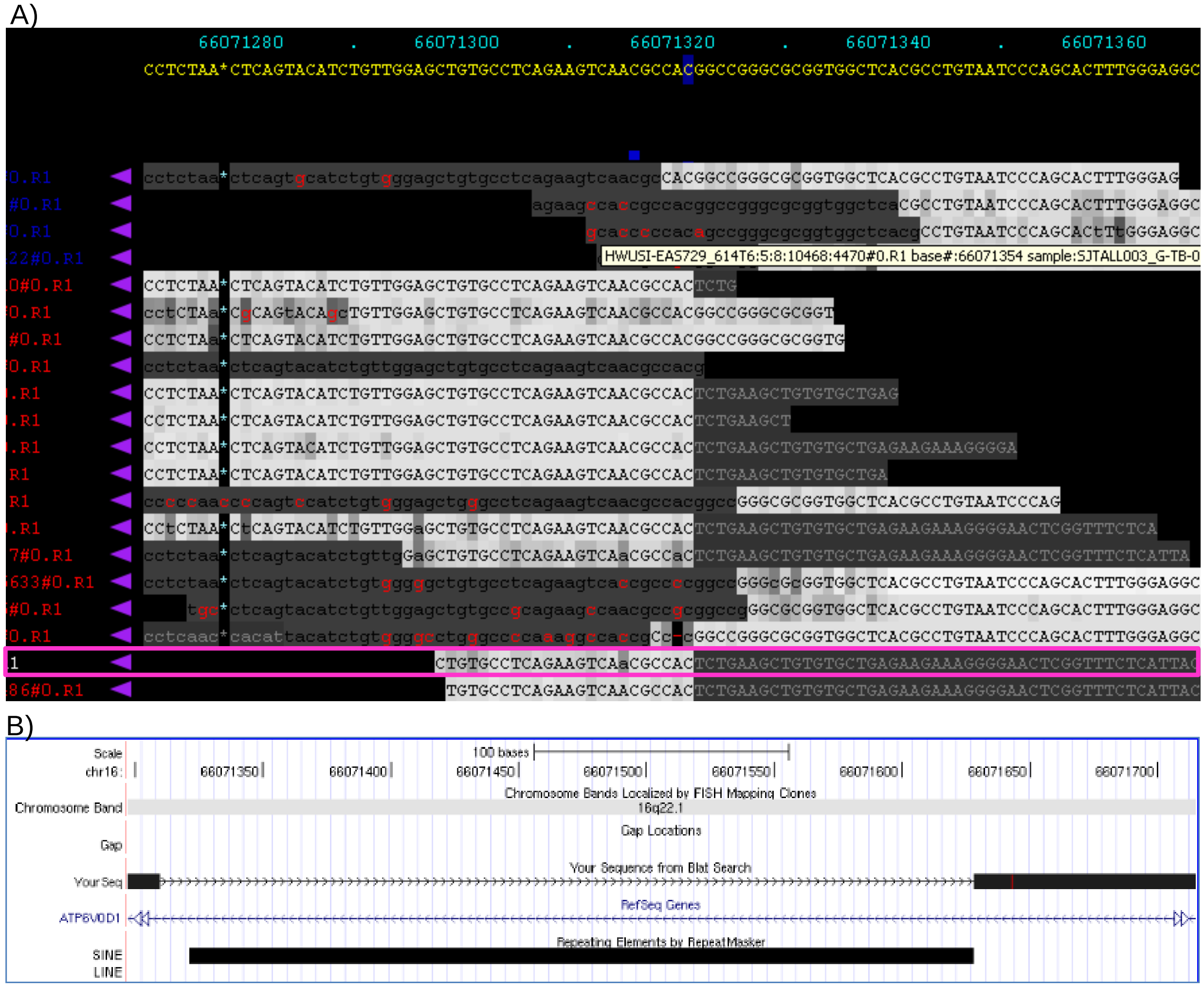
Supplementary Figure 10 Alignment of next-gen reads in both tumor (top panel) and normal (bottom panel) of cell line COLO-829 at deletion spanning chr10:85512695-85513886 reported by Pleasance et al. The deletion was also found in dbSNP (rs71822308). A) Alignment centered at the first breakpoint chr10:85512695. B) Alignment centered at the second breakpoint chr10: 85513886



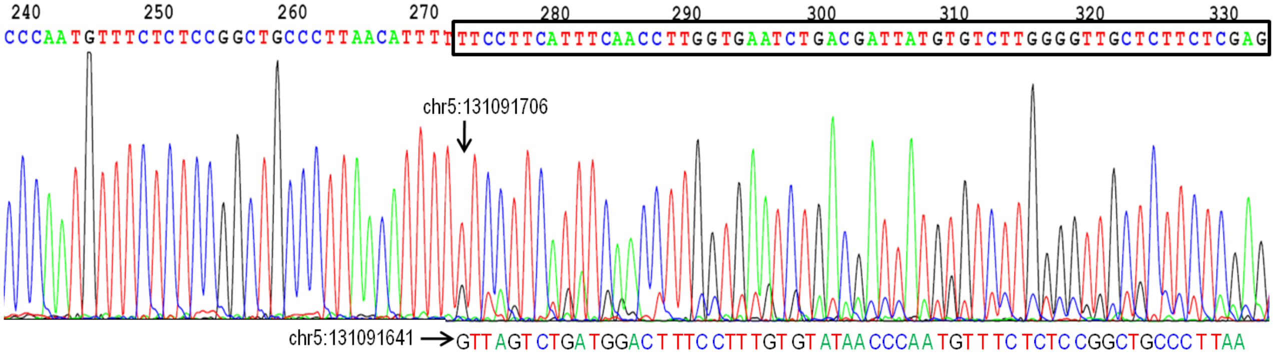
Supplementary Figure 11 Illustration of soft-clipped and paired-end discordant mapping signatures across SV breakpoints in short and long NGS reads.



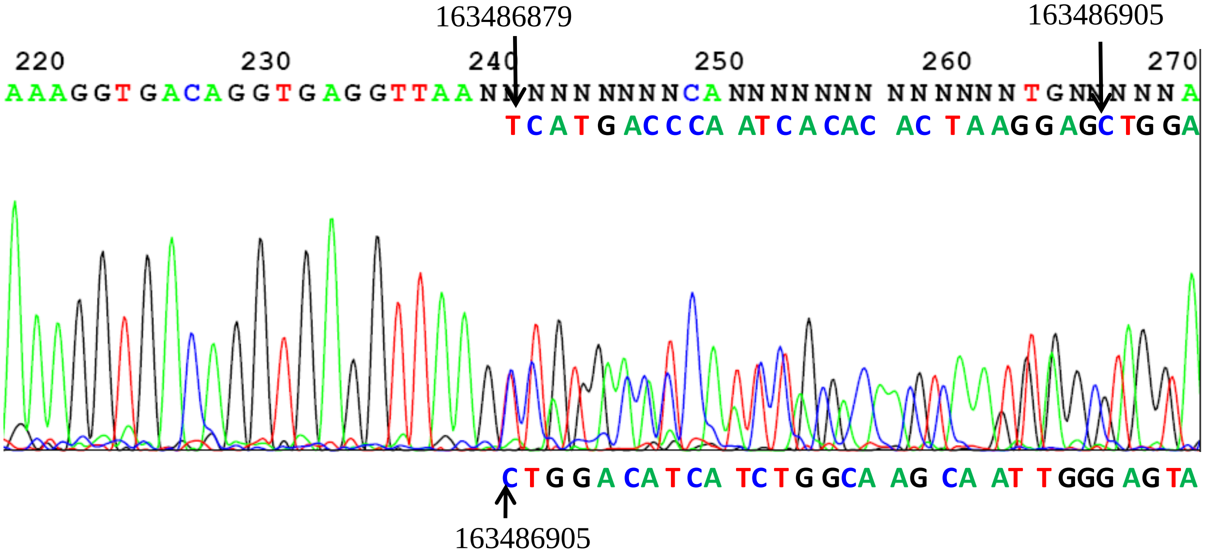
Supplementary Figure 12 An example of a validated somatic deletion in a repetitive region detected by CREST. The 305bp somatic deletion (chr6: 66071321- 66071625) was found in SJTALL003. A) Soft-clipping pattern in whole-genome sequencing reads. The display uses the same style as figure 1 in the main manuscript. B) Alignment of one of the soft-clipped reads in A) (shown in magenta color) to the reference human genome


A)
66071280
66071300
66071320
66071340
66071360
CCTCTAA*CTCAGTACATCTGTTGGAGCTGTGCCTCAGAAGTCAACGCCACGGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGC
HWUSI-EAS729_614T6:5:8:10468:4470#0.R1 base#:66071354 sample:SJTALL003_G-TB-0
B)
Scale
chr16:
100 bases
66071350
66071400
66071450
66071500
66071550
66071600
66071650
66071700
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
16q22.1
Gap Locations
Gap
Your Sequence from Blat Search
Your Seq
RefSeq Genes
ATP6V0D1
Repeating Elements by RepeatMasker
SINE
LINE

Supplementary Figure 13 An example of a low-frequency 65bp somatic insertion that was co-amplified with wild-type in sample SJTALL012. The insertion was caused by replication of 131,091,641bp to 131,091,705bp on chromosome 5. The residues labeled at the bottom were base calls from the secondary peak in the chromatogram. The arrows indicate the start position of the insertion.



Supplementary Figure 14 An example of a small somatic deletion detected by CREST. The 26bp somatic deletion (chr1:163486879-163486905) was found in SJTALL003. The bases pointed by the arrows were manually decoded from the double-peak region. The bases labeled at the top match the reference sequence while those at the bottom match the deletion 26bp away.



Supplementary Table 1

Summary of CREST SV analysis results and validation in five T-ALL samples.

Sample	CREST Prediction						Assayed	Validated SVs						Success
	CTX	ITX	DEL	INS	INV	Total		CTX	ITX	DEL	INS	INV	Total	
SJTALL002	6	9	8	6	0	29	28	6	7	8	4	0	25	89%
SJTALL003	15	3	8	4	0	30	28	13	3	7	3	0	26	93%
SJTALL011	6	5	2	3	1	17	18*	5	5	1	3	1	15	83%
SJTALL012	5	6	4	2	0	17	17	5	3	3	2	0	13	76%
SJTALL013	4	2	4	7	0	17	16	2	1	3	4	0	10	63%
Total	36	25	26	22	1	110	89	31	19	22	16	1	89	83%

*****Two pairs of breakpoints matching the two ends of an inversion were assayed

Supplementary Table 2

Summary of CREST SV analysis results for COLO-829.

The sheet labeled SV includes all inter-chromosomal re-arrangements or intra-chromosomal SVs spanning more than 500bp. The second sheet "Small_indels" include all indels less than 500bp. The first four columns (ChrA, PosA , OrtA, #SC) show the chromosome, position, orientation and number of soft-clipped reads observed at the first breakpoint. Column 5-8 (ChrB, PosB, OrtB, #SC) show the chromosome, position, orientation and number of soft-clipped reads observed at the second breakpoint. Column 9 (Type) defines the SV Type. Column 10,11 show GSAV score and predicted SV type. 0 means not predicted by GSAV. Column 12, 13 shows the gene names where the first and the second breakpoint is located, respectively. The entire genomic regions are used to define overlap between gene and a SV.

Table 3. Experimental validation results of 20 novel SVs in COLO-829 identified by CREST.

ChrA, PosA and OrtA refer the chromosome, position, and orientation of the first breakpoint. ChrB, PosB and OrtB refer the chromosome, position and orientation of the second breakpoint. The Lane# listed here corresponds to the lane# marked in Supplementary Figure 3. The experiment also includes two known SVs as positive controls.

ChrA	PosA	OrtA	ChrB	PosB	OrtB	Type	Verified	Lane#	Forward Primer	Reverse Primer
Novel SVs by CREST										
1	172599726	+	4	96089478	+	CTX	no	1	TTCAGCATTAATTGAAATGCTCATATGGTT	AAGCAGAGCTTGCAGTGAGCCGAGATT
1	222849418	+	1	222852657	+	DEL	yes	2	TGCACTTTGCTTTATTGTGCCTTGCAGA	TCTGTAGCATGCAATGCTGTTTGATAGC
3	25376069	+	10	60147227	+	CTX	yes	3	GTTCAGTGGTGATCTCCCCTTTATCA	TGCAATGGTATTTCATGCAGCTACAAA
5	28823753	+	5	28998853	+	DEL	yes	4	GGACCCAGGGAATGAAGCATCAGGT	CCCAGGAGCGCGGATGACTTTGA
6	26302010	-	3	26406921	+	CTX	yes	5	AAAGTCTCTCCCTTCAGGCCTGTCCTC	GGGTTCATCTCACTAGGGAGTGCCAGA
6	65355103	+	15	19202028	-	CTX	yes	6	TCTTCCCGATCTTCTTTCCACCCTTG	AGTCAAGCAGGGGGATCCCTTGAGA
7	150377590	+	15	82601729	+	CTX	yes	7	CACGGTGAGCTGACCTCCCTACCTAT	TTTGGATCCTGAGAAAGGAAGGCTTAT
7	85691636	+	7	85682604	+	INS	yes	8	TGATCAATAGTGACAATCAGGTGTTCCA	TGGTGAGTGCTCCATGGTTTCTACACT
7	110180568	+	7	110181697	+	DEL	yes	9	TGCTTTATAGGAAATGCTTGGTTGATGCTG	GCGGTAAGCAGATGTGTATCTGTTCACTTA
7	125533359	+	7	125954137	+	DEL	yes	10	GAACCAAATGTCCCACCTCAAGGGACT	CACCGCGCCCAGCCAACTATATAACA
7	143590165	+	7	143719727	+	DEL	yes	11	GGGAACTCTGCTGCCCATCTGAGAA	CCCCAGAGCCCCATCCTGTCAATAA
10	7172882	+	19	17258637	+	CTX	yes	12	TGGAAATGTTTCATGCGTGACCATCAG	GAGGTCACTCTTGTCGCCATCTTGGTA
10	60147428	+	12	70953341	-	CTX	no	13	TGAGTGGCAAGCCTGGTATGCTGTGTCT	AACCACCACGCAGGCGGGGACTC
11	80463250	+	11	80771428	+	DEL	yes	14	ACTTTGGGTTTGTTGTGCAGATGTCA	TTTGGGATCAAGGGCATTGCCTAAACT
15	39415802	+	15	39408584	+	INS	yes	15	TAAATTTATGGCCGGGCATGGTG	GGGTTATGAAATGGCACAGCTCAAGACA
18	9858617	-	10	7674379	+	CTX	yes	16	ATCTGCACACTGCATTGCCTCCCAATA	CCTCCCCTCTATTTTCCTCTTTCCCCTAC
20	36708081	+	15	21257972	+	CTX	yes	17	CTCCTCATTCCACCAAACCAGTGCT	TTTTCAAAACCATGAAGGAAAAGGAAAC
20	14936825	+	20	15056818	+	DEL	yes	18	TCAAATTCCCCTAGAAAAGAACTGGCTTTC	AACAATTCTCGAAAATACCAGCCATCA
X	31211124	+	X	31948671	+	DEL	yes	19	GGTGTTCTTTGAACCAAGTGGAGTCTGA	TGGGTCAATTCGGTTGATTAACTTTGGA
X	32008454	+	X	32111172	+	DEL	yes	20	GGCACCTCCTTCTGTAATCACAGTGTTGCT	CACCAGGGAACTTCAACACCATCCAAG
Known SVs by Pleasance et al										
10	7099517	-	19	17257810	-	CTX	yes	21	ACTGGGCATTCCAAACCTGCAAAGAG	ATGAAAGGCATTTGGCTTCCCAGTTCC
17	7277162	+	17	7278457	+	DEL	yes	22	AAAATGTATATTAGGCCTGGTGCAGTGG	CTGAGTCGTGGCAGGGAAAACACTTT

Supplementary Table 4

Performance of CREST, BreakDancer GSAV and Pindel on detecting structural variations found in NA12878 using simulated WGS data.

SV Type	**High Quality Simulation**				**Normal Quality Simulation**			
	CREST	**BreakDancer**	**GSAV**	**Pindel**	**CREST**	**BreakDancer**	**GSAV**	**Pindel**
DEL(n=642)	0.794	0.768	0.919	0.202	0.743	0.729	0.911	0.115
DUP (n=271)	0.675	0	0.712	0.155	0.613	0.011	0.701	0.085
All (n=913)	0.759	0.54	0.858	0.188	0.704	0.516	0.849	0.106
False Positives	22	361	3,389,524	41	24	4,407	3,225,728	107

Supplementary Data 1

Comparison of CREST performance with other SV detection methods

To compare the performance of CREST with programs that implement the paired-end discordant mapping (PEM) approach, we re-analyzed the T-ALL DNA sequencing data using two programs that implement the PEM algorithm: BreakDancer[2] and GSAV[3]. Using the default parameters and retaining tumor-only SVs with scores ≥ 30 and number of supporting read pairs ≥ 3, BreakDancer predicts a total of 1,064 SVs, 27 (2.7%) of which are also found by CREST. GSAV predicts a total of 5,880,492 SVs, 91 (0.0015%) of which overlap with the CREST prediction including 76 validated SVs. For the 1,064 SVs predicted by BreakDancer, we ran a post-process to evaluate the possibility that valid SVs detected by BreakDancer might have been missed by CREST. The analysis first removes non-specific reads and then assembles the remaining reads mapped to the breakpoint interval predicted by BreakDancer (details in **Online Methods**). Aside from the 27 validated SVs, no additional SVs predicted by BreakDancer survived this post-process. This comparison demonstrates not only the high false positive rate of paired-end discordant-based methods, but also their inability to detect true SVs that can be picked up using CREST. We speculate that longer read length (75bp to 100bp) coupled with the ability to compute local alignments using mapping tools such as BWA may cause loss of the paired-end discordance mapping signature for some of the SVs. **Supplementary Fig. 11** illustrates this possibility by showing that PEM signature obtained from a short-read can be replaced by soft-clipping signature of a long read for a DNA fragment cross a structural variation.

The concept of using sequences that span breakpoints has been previously used to identify insertion/deletion (indel) variations by the program Pindel[4]. To compare the performance of Pindel with CREST, we re-analyzed the 5 T-ALL cases using a modified version of Pindel which can directly use BAM files as input data. For each case, tumor and matched normal data were analyzed together. Indels absent in the germline data were retained as putative somatic events. A total of 425,963 putative somatic indels were identified, only 5 of which were among the validated somatic SVs found by CREST. These five validated SVs included four deletions ranging from 26bp to 1,398bp and one 26bp insertion, one of which is a 305bp deletion that removes a repetitive ALU retrotransposon on chromosome 6 (**Supplementary Fig. 12).**

The lack of the consistency between Pindel and CREST is expected. Even though both programs use reads across the breakpoints for variation detection, Pindel was designed for identifying insertion/deletion variations while CREST was designed to detect gross structural variations including inter-chromosomal translocations as well as non-indel intra-chromosomal alterations such as inter-chromosomal translocations and inversions. The majority (58%) of the validated SVs in the 5-TALL cases are these non-indel events which cannot be detected by Pindel. Furthermore the majority of the indels found by CREST are gross insertions or deletions: among the 36 validated indels, only 9 are less than 10kb in size, while 17 are larger than 1Mb, and the longest is a 57Mb deletion. CREST was not designed for small indel detection because mapping algorithms like BWA[5] are able to compute gapped alignments for the majority of NGS reads with small indels up to 50bp, resulting in lack of soft-clipping signature for small indels that are <50bp. On the other hand, Pindel is only able to compute indels within a specified length. Increasing this length is expected to increase the probability of finding random hits that are likely to be false positive. Application of Pindel on finding indels >10kb was not reported by Ye *et al*[4].

In addition to small indel size, the >400K somatic indels predicted by Pindel of the five T-ALL tumors far exceed the estimated 562 to 1,125 somatic indels ascertained from the background mutation rate of pediatric cancer (**details in Online Methods**). This suggests that the vast majority of the predictions are false positives. Some of false positive somatic indels are caused by false negative prediction of the matching germline sample because approximately 12% of the "somatic" indels predicted by Pindels match the germline indels detected by our indel analysis pipeline (data not shown). To verify the projected error rate, we reviewed the first 100 >100bp deletions of chromosome 1 from the sample SJTALL002 from a total of 13,401 >100bp somatic deletions predicted by Pindel across the five tumors. The deletions range in size from 101bp to 8,120bp. 37 deletions are of germline origin as soft-clipped reads are present in both tumor and matching normal. The remaining deletions do not have soft-clipped reads. They include a) 22 overlapping deletions ranging in size from 339 to 1,019bp located within a 1,233bp centromeric region of chr1:121,186,883-121,185,650 with >10,000 fold coverage in both tumor and normal; b) 38 deletions in regions with multiple simple tandem repeats (STR) and/or polynucleotide repeats which are prone to alignment artifacts; and c) 3 deletions predicted to be of size 669bp, 4,395bp and 6,778bp but have 1-2bp small indel at one of the breakpoints (in both tumor and normal) instead with no evidence for the predicted larger deletions in either tumor or normal based on SNP array or NGS sequence coverage. In summary, false negative germline prediction, false positive calls in repetitive regions such as STR, polynucleotide repeats, centromere and alignment error of small indels attributed to the false positive call of these 100

>100bp somatic deletions predicted by Pindel, confirming the high error rate projected from the background somatic mutation rate in pediatric cancer.

The concept of using sequences that span SV breakpoints has been previously used to identify altered mRNAs. Maher et al[6] demonstrated the feasibility of identifying chimeric transcripts through an integrated analysis of long (>200bp) and short (36bp) reads where the long reads serve as a template for SV breakpoint discovery. However, this group later[7] considered paired-end mapping for chimeric transcript detection superior to the single-end, long-read (100bp) approach because the split-read method generated a higher number of SVs that could not be validated by paired-end mapping. The experience by Maher et al[6] shows that identification of soft-clipping signature alone is not sufficient for accurate detection of structure variations. To filter false positive soft-clipping signature, CREST requires presence of soft-clipping signatures across both sides of a SV detected through an iterative approach of assembly-mapping-searching-assembly-alignment. In fact, over 99% of the soft-clipped sites identified in the first step of the algorithm get filtered by this process.

Supplementary Data 2

SV analysis in melanoma cancer cell line COLO-829 by CREST

To further assess the performance of CREST, we applied it to a published whole-genome sequencing dataset from the metastatic melanoma cancer cell line COLO-829[1]. Using a paired-end discordant mapping method[8] the published analysis reported 37 validated SVs[1]. By comparison, CREST identified 76 SVs (**Supplementary Table 2**) including 26 of the 37 reported SVs. Of the 11 reported SVs that were not identified by CREST, 6 were found to have soft-clipped reads in the matching normal sample COLO-829BL, including 2 that have been reported as germline deletions in dbSNP (rs72415809 and rs71822308). Importantly, with the exception of one deletion that had a very low frequency (1.7%) in both tumor and normal, the frequency of the other soft-clipping reads in the germline sample ranged from 17% to 42% (**Supplementary Table 2 and Supplementary Figs. 5-10**), strongly supporting the interpretation that these SVs are germline variants.

CREST identified 50 additional SVs that were not reported by Pleasance et al[1], 34 of which were either inter-chromosomal translocations, or ≥500bp intra-chromosomal alterations (**Supplementary Table 2**). We selected 20 of the novel SVs and 2 known SVs for direct validation using PCR amplification of DNA extracted from the COLO-829 cell line followed by Sanger sequencing. Eighteen of the 20 novel SVs had PCR product of the predicted size (**Supplementary Fig. 3a**), and were confirmed to represent the SV breakpoints by Sanger sequencing (**Supplementary Fig. 3b-d, Supplementary Fig. 4**). The validated SVs include 7 CTX, 9 DEL and 2 INS. Interestingly, one validated novel SV has breakpoints of chr10:7172882(+), chr19:17258637(+) (**Supplementary Fig. 3e**) and upon initial inspection appears to form a reciprocal translocation with a known SV with the two breakpoints of chr10:7099517(-), chr19:17257810(-). However, the layout and the orientation of the breakpoints of these two SVs show that this rearrangement will result in an insertion of a 73kb segment on chromosome 10 between 7099517-7172882bp to chromosome 19. Therefore, it is an inter-chromosomal insertion rather than a reciprocal translocation. The remaining six validated CTX are unbalanced inter-chromosomal translocations based on the current analysis.

Supplementary Data 3

Simulation study for assessing CREST false negative rate in finding germline indels

To provide an assessment of false negative rate of CREST on identifying germline SV polymorphisms, simulated whole-genome sequencing data were generated using validated copy number variations (i.e. deletions, duplications and insertions) compiled as a gold standard data set for NA12878, one of the individuals whose germline structural variations were characterized extensively by the 1000 Genomes Project (Mills et al). NA12878 was sequenced at an average of 42x coverage using three sequencing platforms (Illumia/Solexa, Roche/454 and Life Technologies/SOLiD) and analyzed by 19 SV detection methods, 12 of which were evaluated for their sensitivity in detecting deletion polymorphisms. Two sets of whole-genome, 40x simulation data were generated to model the library construction, error distribution and mapping rate of the empirical Illumina sequencing data of the 10 ALL whole-genome sequencing data reported in this study. Details are described in preparation of simulated whole-genome sequencing data for NA12878 in **Online Methods**.

We ran CREST, BreakDancer, GSAV on the simulated WGS data sets using the default parameters. Pindel was downloaded from https://trac.nbic.nl/pindel/wiki/UserManual. The search range for Pindel was set to 9 (e.g. maximum SV size to be 2,071,552) because none of the germline SVs in NA12878 is longer than 2MB. BreakDancer output was supplied as one of the input parameters.

Since pair-end mapping can only infer the approximate location of the SV breakpoints, a SV predicted by BreakDancer or GSAV within 220bp of a known SV was considered a match. A much more stringent criterion (+/-20bp) was used to define a match between a SV predicted by CREST and a known SV because CREST is expected to generate SV breakpoints at base-pair resolution. The results (**Supplementary Table 4**) show that CREST was able to predict 70-76% of the SVs identified in NA12878 compared to 53-56% of the prediction rate by BreakDancer. Furthermore, the false positive rate in CREST is very low (3% of the total calls) and do not vary between the high-quality and normal quality simulation. By contrast, BreakDancer has a very high false positive rate (43%) even using high-quality simulation data. In normal-quality simulation, the number of false positive calls increased by 10-fold resulting in a 91% false positive rate. GSAV has the highest sensitivity of 85-86%; however, the over 3 million false SVs in each simulation indicate an extremely high false positive rate. Pindel has the highest false negative rate as it only found 19% and 11% of the known SVs while the number of false positive calls 2-4 times of that of CREST.

Supplementary Discussion

Limitations of CREST and Comparison of CREST with Genome STRiP

Although CREST provides a significant improvement over standard paired-end approaches for identifying SVs, it is unable to robustly identify all SVs under the following three circumstances. a) SV breakpoints located at highly repetitive DNA sequences or within tandem duplications of the reference genome. This will result in loss of both the soft-clipping and mate-pair discordant mapping signature. b) Rearrangements that have target site duplications (also known as microhomology) or non-template insertions at the breakpoints that are of similar or longer length than a single NGS read. Identification of non-template insertions longer than the NGS reads will require de-novo assembly for constructing contigs that include both the insertions as well as their flanking reference sequences. CREST is able to find soft-clipping signatures at the flanking reference sequences but the two breakpoints cannot be connected by a single NGS read. c) SV breakpoints in regions with low quality sequences or low sequence coverage. False negative rate for CREST is dependent on how often structural variations occur in these regions which may vary dramatically from one sample to the other. One potential improvement for CREST is to combine soft-clipped reads with PEM reads for SV detection, which may increase its sensitivity in regions with low or poor sequence coverage.

Compared with the other three samples, SJTALL012 and SJTALL013 had a lower validation rate (**Supplementary Table 1**). In these two samples, over 50% of un-validated SVs are INS or ITX that have breakpoints within 500bp. We speculate that the decreased validation rate may be related to the following two reasons: a) failure to obtain a mutant-only PCR product since both the wild-type and the mutant allele are amplified by the primers and a low-frequency mutant allele may not be detectable by Sanger sequencing (**Supplementary Fig. 13**); and b) ITX within a small region may be caused by artifact during library preparation. In this study, the majority (85%) of the 55 validated inter-chromosomal SVs are gross alterations spanning longer than 500bp. The longest is a 57Mb deletion in SJTALL013 which was also detected as a somatic deletion on analysis of single nucleotide polymorphism (SNP) microarray data as a somatic deletion (data not shown). Though small indels below 100bp can be found by CREST (**Supplementary Figs.13** and **14**), it is not a tool designed for such analysis as alignment artifacts in small indels usually require a more accurate alignment algorithm (such as the Smith-Waterman algorithm[9]) for correction. Furthermore, new mapping algorithms such as BWA[5] can compute gapped alignments for indels up to 53bp, resulting in loss of the soft-clipping signature in these regions.

Recently Handsaker et al[10] developed a new method, Genome STRiP, which uses population footprint to reduce false discovery of germline polymorphic deletion caused by chimeric clones, read depth variation, and alignment artifact in repetitive regions of Next-generation sequencing. Genome STRiP is able to achieve high accuracy based on three population signatures: a) coherence around shared alleles; b) heterogeneity in population and c) allele substitution. Its sensitivity, however, is highly dependent on allele frequency in a population which ranges from a low of 30% for low-frequency (1%) deletions to a high of 80% for high-frequency (>15%) events

Genome STRiP is optimized for germline polymorphism analysis; however, these optimizations are unlikely to be applicable for detecting somatically acquired structural variations in cancer genome. Specifically, almost all somatically acquired structural variations are "incoherent", i.e. SV breakpoints are unique in each individual tumor even though the same gene is targeted in multiple tumors. For examples, the breakpoints for the highly recurrent BCR-ABL1 rearrangement in chronic myeloid leukaemia are dispersed[11] and none of 89 somatic SVs detected in the five T-ALL cases share the same breakpoint. Similarly, allele substitution assumes allele segregation in a population which does not apply for a cancer genome. On the other hand, a highly selected oncogenic mutant allele in a specific tumor may have the same mutant allele present in all tumors; thereby violating the rule of population heterogeneity. Although there has not been such a report for somatic structural variation, the result derived from single-nucleotide variations shows the most important oncogenic event can be missed if we rule out such a possibility. For example, amino acid G12 of the oncogene KRAS is mutated almost universally in all (95%) pancreatic tumors[12]. Furthermore, Genome STRiP was designed for finding only deletions. It is not able to find SVs that are inter-chromosomal translocations, intra-chromosomal translocations, inversions and amplifications; thereby can miss critical oncogenic fusion proteins such as BCR-ABL and ETV6-RUNX1 that are caused by non-deletion events.

In contrast, CREST does not rely on population signature for reducing false positive calls. The iterative approach of assembly-mapping-searching-assembly-alignment that requires consistent signature at the two breakpoints removes the artifacts associated with the soft-clipping signature. In addition, the option for running paired tumor-normal analysis not only identifies somatically acquired SVs, it also filters systematic errors in library construction and alignment artifacts that are present in both tumor and matching normal.

Supplementary References

1. Pleasance, E.D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196.
2. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681 (2009).
3. Sindi, S., Helman, E., Bashir, A. & Raphael, B.J. A geometric approach for classification and comparison of structural variants. *Bioinformatics* **25**, i222-230 (2009).
4. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
5. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
6. Maher, C.A. et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101 (2009).
7. Maher, C.A. et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* **106**, 12353-12358 (2009).
8. Campbell, P.J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729 (2008).
9. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
10. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-276.
11. Score, J. et al. Analysis of genomic breakpoints in p190 and p210 BCR-ABL indicate distinct mechanisms of formation. *Leukemia* **24**, 1742-1750.
12. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-1806 (2008).