Supplementary Table 1. SV breakpoint calling results from the top performers in the ICGC-TCGA DREAM 8.5 Somatic Mutation Calling Challenge subchallenges in silico datasets 1, 2, 3 and 4.

Synthetic datasets	Mutation types	Purity	Subclones	Top performers	Sensitivity/Precision	Balanced accuracy
in silico 1	SNV & SV(deletions, duplications, inversions)	100%	N/A	Delly	0.841/0.987	0.914
				Manta	0.822/0.987	0.905
				Meerkat	0.838/0.931	0.885
				novoBreak	0.811/0.932	0.872
in silico 2	SNV & SV (deletions, duplications, inversions, insertions)	80%	N/A	novoBreak	0.794/0.983	0.888
				Manta	0.754/0.984	0.869
				Delly	0.739/0.980	0.859
in silico 3	SNV & SV (deletions, duplications, inversions, insertions) & INDEL	100%	50%, 33%, 20%	novoBreak	0.801/0.984	0.892
				Manta	0.768/0.989	0.879
				Delly	0.783/0.972	0.878
in silico 4	SNV & SV (deletions, duplications, inversions) & INDEL	80%	50%, 35%	novoBreak	0.849/0.990	0.920
				Grigoriev_lab	0.853/0.985	0.919
				CASbreak	0.810/0.994	0.902

Supplementary Table 2.INDEL calling results from the top performers in the ICGC-TCGA DREAM 8.5 Somatic Mutation Calling Challenge subchallenges in silico datasets 3 and 4.

Synthetic datasets	Mutation types	Purity	Subclones	Top performers	Sensitivity/Precision	Balanced accuracy
in silico 3	SNV & SV (deletions, duplications, inversions, insertions) & INDEL	100%	50%, 33%, 20%	Pindel	0.875/0.976	0.926
				novoBreak-indel	0.746/0.908	0.827
				Manta	0.623/0.940	0.781
in silico 4	SNV & SV (deletions, duplications, inversions) & INDEL	80%	50%, 35%	novoBreak-indel	0.788/0.928	0.858
				EmToo_Broad	0.772/0.921	0.846
				AstraZeneca_VarDict	0.750/0.835	0.793

Supplementary Table 6. Summary of novoBreak calls from the whole-genome sequencing data of 22 Breast Cancer patients from TCGA.

sample	#filtered_calls	DEL	DUP	INV	TRA
TCGA-A1-A0SM	111	19	16	27	49
TCGA-A2-A04P	801	68	105	402	226
TCGA-A2-A0D1	252	57	35	88	72
TCGA-A2-A0D4	211	34	42	72	63
TCGA-A2-A0YG	209	40	52	76	41
TCGA-A2-A25B	453	73	185	84	111
TCGA-A8-A08B	945	72	71	685	117
TCGA-A8-A08L	320	201	21	41	57
TCGA-A8-A08S	173	42	13	71	47
TCGA-A8-A092	44	17	7	9	11
TCGA-A8-A09I	394	119	83	142	50
TCGA-A8-A09X	82	13	10	13	46
TCGA-AN-A04D	172	98	14	12	48
TCGA-AN-A0AT	647	93	350	36	168
TCGA-AO-A0JM	301	48	35	120	98
TCGA-AO-A124	747	170	351	50	176
TCGA-AR-A0TX	331	56	41	80	154
TCGA-AR-A24Z	625	113	114	208	190
TCGA-AR-A256	638	254	154	78	152
TCGA-BH-A0H0	56	8	7	12	29
TCGA-BH-A0H6	65	6	6	11	42
TCGA-E2-A152	92	27	12	18	35

Supplementary Note 1 Evaluation of *k*-mer size, minimal count, low-quality base trimming and error correction on the performance of novoBreak

We tested the performance of novoBreak under a range of *k*-mer sizes, minimal *k*-mer count, with and without trimming low-quality read ends, and with and without correcting errors using the DREAM in silico 2 (IS2) data. We recorded the sensitivity, peak memory consumption and runtime under different settings. Because the insertions simulated in IS2 are template insertions, which may introduce confusions in evaluation, we excluded them from these experiments and performed evaluation based on deletions (DELs), inversions (INVs) and duplications (DUPs). The *k*-mer size and the minimal *k*-mer count are parameters of novoBreak, which can be easily set on the command lines. We modified the source code to generate a version of novoBreak that does not trim low-quality ends and a version that does not correct errors in BAM files. Below are the results from these settings:

k-mer size	Min *k*-mer count	Trimming low-quality ends?	Correcting errors?	Sensitivity (N=491)	Peak memory Consumption	Time with 16 cores
23	3	Y	Y	435 (88.6%)	37.37GB	8.6h
25	3	Y	Y	436 (88.8%)	37.34GB	9.1h
27	3	Y	Y	437 (89.0%)	37.25GB	10.1h
29	3	Y	Y	437 (89.0%)	37.25GB	10.7h
31	3	Y	Y	437 (89.0%)	37.27GB	10.0h
31	1	Y	Y	NA	Out of 200GB Memory cap	NA
31	2	Y	Y	442 (90.0%)	37.28GB	12.3h
31	3	N	Y	437 (89.0%)	43.28GB	13.5h
31	3	Y	N	437 (89.0%)	46.69GB	14.1h

Note that these experiments were performed using a network-based storage cluster with relatively slower IO operations than using a standard storage. Therefore, the reported runtimes were bigger than what one might observe when using standard hardware. However, these differences do not affect the fairness of our comparison. From this result, it is clear that choosing different *k*-mer size had little effect on the sensitivity. The sensitivity from a smaller *k*-mer size was only slightly lower than those from larger *k*-mer sizes because junction sequences in these sizes appeared largely unique in the reference genome. Setting the minimal *k*-mer count at 1 retained lots of sequencing errors, which caused novoBreak to run out of memory. Setting minimal *K*-mer count at 2 resulted in a slightly better sensitivity at the cost of more memory and ~25% more runtime than setting minimal *k*-mer count at 3. Trimming low-quality ends or correcting errors had almost no effect on the sensitivity, but the peak memory consumptions and runtimes were notably increased when these operations were not performed. To achieve a balanced performance across sensitivity, peak memory and runtime and to optimize implementation, we chose 3 as the minimal *k*-mer count and 31 as the *k*-mer size and implemented the low-quality end trimming and error correction for the default setting of novoBreak.

Supplementary Note 2 Assembly and alignment tools used in novoBreak

The default assembler used in novoBreak is SSAKE[1] and the default aligner is BWA-MEM[2]. During the development of novoBreak, we tested quite a few assemblers and aligners. For assemblers, we tested SOAPdenovo[3], velvet[4], phrap[5], CAP3[6], SGA[7], celera[8], SSAKE[1] and our own assemblers. For aligners, we tested BWA-SW[9], BWA-MEM[2], BLAT[10], LAST[11], LASTZ[12] and BLASTZ[13]. For the aligners, we tested them under the default settings. But for the assemblers, we tried quite a few settings based on the documentations and our understanding of the algorithms. For example, for the *de Bruijn* graph assemblers, we tested a range of *k*-mers to get the optimal assembly results. In the end, we found BWA-MEM and SSAKE were the best choices in achieving a high balanced accuracy. As the reads become longer and more accurate, other aligners or assemblers may become better choices. The modular design of novoBreak makes it easy to swap in alternative aligners or assemblers into novoBreak workflow. Thus, improvement of novoBreak can be relatively easily achieved in the future versions.

Supplementary Note 3 novoBreak performance in low coverage regions

In the DREAM challenge in silico 3 (IS3) data, the average read depth is about 40X. There are 79 unique true positive calls detected by novoBreak but neither discovered by DELLY nor Manta. Sixty-six (83.5%) of the 79 calls are covered by less than 40 reads, 50 (63.2%) are covered by less than 30 reads, and 23 (29.1%) are covered by less than 20 reads. A comparison of the coverage of these SV regions between the tumor and the normal genomes shows that there is significantly less coverage (mean 27.9X) in the tumor genome than in the normal genome (mean 33.5X) (P-value = 0.0037 by Student's T-test). A further look of the BWA alignment shows that 69 (87.3%) of the calls are not supported by any split read, 5 (6.3%) by only 1 split read. Only 5 (6.3%) are supported by more than 1 split read. Additionally, no discordant read pairs are found in 69 (87.3%) of the 79 calls. These statistics indicate that lack of coverage is the reason why DELLY and Manta failed to detect SVs in these regions.

Supplementary Note 4 Evaluation of novoBreak using data from a patient with low-grade glioma

We also benchmarked novoBreak on data produced from a low-grade glioma (LGG) patient (SJLGG039)[14]. Nineteen (19) SVs have been previously discovered and experimentally validated in this patient using whole genome sequencing (analyzed using CREST[15]) and whole transcriptome sequencing data (analyzed using deFuse[16]). We downloaded the BAM files containing the only whole genome sequencing data from the European Bioinformatics Institute (accession id: EGAS00001000255). Under the standard setting applied to analyze the DREAM challenge, TCGA and the COLO829 data, novoBreak detected only 5 of the 19 SVs. However, this can be explained by a potential lack of purity of the tumor data. We noticed that 31 out of the 38 expected breakpoints (81.6%) had 3 or fewer supporting reads (mutA and mutB columns in Supplementary table 7[14]), which were considerably less than what would be expected from a genome of 45X average coverage, Thus, we adjusted the last quality-control step of novoBreak and filtered the raw calls using a less stringent filter (released in novoBreak v1.1.3rc). We were able to identify 16/19 (84.2%) of the SVs, as summarized in the following table.

Sample	ChrA	PosA	Orient_A	ChrB	PosB	Orient_B	Type	CREST	novoBreak
SJLGG039	1	3643330	-	11	115411715	+	CTX	NO	NO
SJLGG039	1	3643581	+	11	69538645	+	CTX	YES	YES
SJLGG039	1	3649438	+	22	47557135	-	CTX	YES	YES
SJLGG039	1	12582471	+	12	103110023	+	CTX	YES	YES
SJLGG039	1	12584121	+	11	118491727	-	CTX	YES	YES
SJLGG039	3	12649234	+	3	46001511	+	DEL	NO	NO
SJLGG039	3	49168971	+	11	66966686	+	CTX	NO	YES
SJLGG039	3	186716601	+	4	1899271	+	CTX	YES	YES
SJLGG039	4	5039611	+	3	183723574	+	CTX	NO	YES
SJLGG039	7	9539557	+	7	6300098	+	INS	YES	YES
SJLGG039	10	98887882	+	16	2158836	+	CTX	NO	YES
SJLGG039	11	69540077	+	1	9555182	+	CTX	NO	NO
SJLGG039	11	69541460	+	12	103111315	+	CTX	YES	YES
SJLGG039	12	103110079	-	10	95588803	+	CTX	YES	YES
SJLGG039	12	103112184	+	3	9662023	+	CTX	NO	YES
SJLGG039	12	106295961	+	1	12581667	+	CTX	YES	YES
SJLGG039	16	410315	-	4	5037703	+	CTX	NO	YES
SJLGG039	16	2170637	+	12	103110956	-	CTX	NO	YES
SJLGG039	16	2172065	+	1	1196437	+	CTX	YES	YES

For comparison, we also ran CREST (under default settings) on the downloaded BAM files. Among the 19 validated SVs, CREST detected 10 (52.6%). In our further investigation, we found that 3 breakpoints (chr1:3643330, chr11:115411715 and chr3:46001511) that novoBreak missed were in the unfiltered results (very low-confidence). The remaining breakpoints had neither discordant read pairs nor split reads in the downloaded BAM files supporting their presence (they were also missed by CREST). This suggested that they were likely detected in the whole transcriptome sequencing data and therefore should not be regarded as false negatives of novoBreak or CREST.

References

1. Warren, R.L., Sutton, G.G., Jones, S.J. & Holt, R.A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500-501 (2007).
2. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
3. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**, 265-272 (2010).
4. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829 (2008).
5. de la Bastide, M. & McCombie, W.R. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics* **Chapter 11**, Unit11 14 (2007).
6. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome research* **9**, 868-877 (1999).
7. Simpson, J.T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome research* **22**, 549-556 (2012).
8. Myers, E.W. et al. A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).
9. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
10. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664 (2002).
11. Kielbasa, S.M., Wan, R., Sato, K., Horton, P. & Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome research* **21**, 487-493 (2011).
12. Harris, R.S. Improved pairwise alignment of genomic DNA. (ProQuest, 2007).
13. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome research* **13**, 103-107 (2003).
14. Zhang, J. et al. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nature genetics* **45**, 602-612 (2013).
15. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature methods* **8**, 652-654 (2011).
16. McPherson, A. et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7**, e1001138 (2011).