A benchmark of structural variation detection by long reads through a realistic simulated model.

Nicolas Dierckxsens[1,2]*, Tong Li [1], Joris R. Vermeesch [2] and Zhi Xie [1*]

[1] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, 510060, China [2] Center for Human Genetics, University Hospital Leuven and KU Leuven, Leuven, Belgium

*To whom correspondence should be addressed. Email: nicolasdierckxsens@hotmail.com ; xiezhi@gmail.com

1. Parameters for the structural variation callers

1. Sniffles (v1.0.11)

We aligned the simulated reads to GRCh38 using Minimap2 (v2.17-r941). We changed the parameters of Minimap2 according to the type of simulated reads, ‘minimap2 -ax map-ont’ corresponds to Nanopore simulated reads and ‘minimap2 -ax map-pb’ corresponds to PacBio simulated reads.
The parameters of Sniffles is ‘sniffles -n -1 -s 3 --genotype’.

2. SVIM (v1.3.1)

The same alignment as for Sniffles. We used SVIM to call structural variants with default parameters.

3. NanoSV (v1.2.4)

The same alignment as for Sniffles. We used NanoSV to call structural variants without ‘depth_support’ mode in its config file. Other parameters are default.

4. Picky (v0.2.a)

The same alignment as for Sniffles. We used the pipeline provided by the github page of Picky to call SVs (https://github.com/TheJacksonLaboratory/Picky/wiki/Using-an-Alternative-Aligner).

5. NanoVar (v1.3.8)

We changed the parameters of NanoVar according to the type of simulated reads, ‘nanovar -l 50 -x ont’ corresponds to Nanopore simulated reads, ‘nanovar -l 50 -x pacbio-clr’ corresponds to PacBio simulated reads and ‘nanovar -l 50 -x pacbio-ccs’ corresponds to PacBio HiFi simulated reads.

6. pbsv (v2.3.0)

We aligned simulated reads to GRCh38 using pbmm2 (v1.3.0) with default parameters. The parameters of pbsv is ‘pbsv call -m 50’.

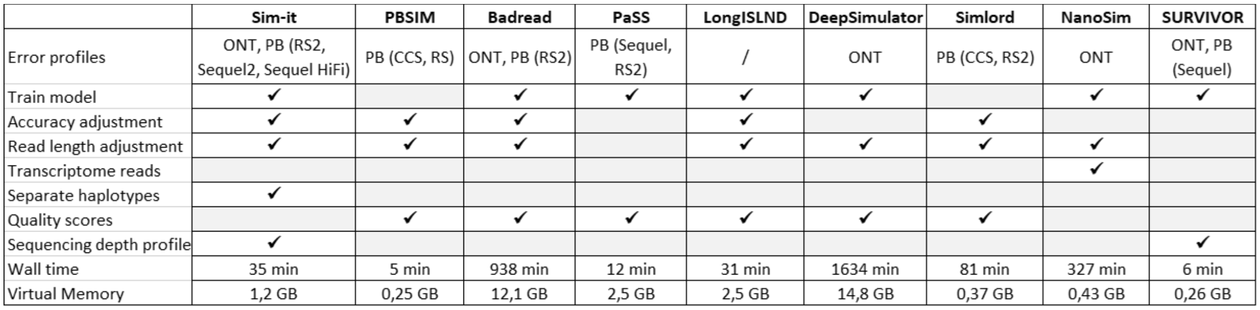
7. cuteSV (v1.0.10)

The same alignment as for Sniffles. The parameters are ‘--max_cluster_bias_INS 100 --diff_ratio_merging_INS 0.3 – max_cluster_bias_DEL 100 – diff_ratio_merging_DEL 0.3 – genotype -s 3

2. Long read simulators benchmark

We compared the features of 9 different long read simulators. We tested the wall time and memory consumption for the simulation of 15x coverage Nanopore or PacBio (when no Nanopore available) long reads for the human chromosome 1 (GRCh38).

Table S1 | Comparison of the features of each long read simulator and a system requirements benchmark.



	Sim-it	PBSIM	Badread	PaSS	LongISLND	DeepSimulator	Simlord	NanoSim	SURVIVOR
Error profiles	ONT, PB (RS2, Sequel2, Sequel HiFi)	PB (CCS, RS)	ONT, PB (RS2)	PB (Sequel, RS2)	/	ONT	PB (CCS, RS2)	ONT	ONT, PB (Sequel)
Train model	✓		✓	✓	✓	✓		✓	✓
Accuracy adjustment	✓	✓	✓		✓		✓		
Read length adjustment	✓	✓	✓		✓	✓	✓	✓	
Transcriptome reads								✓	
Separate haplotypes	✓								
Quality scores		✓	✓	✓	✓	✓	✓		
Sequencing depth profile	✓								✓
Wall time	35 min	5 min	938 min	12 min	31 min	1634 min	81 min	327 min	6 min
Virtual Memory	1,2 GB	0,25 GB	12,1 GB	2,5 GB	2,5 GB	14,8 GB	0,37 GB	0,43 GB	0,26 GB

3. Error profiles of long read simulators

The error profile of PBSIM (v1.0.4) is located in the installation folder and the relative path is ‘PBSIM/data/model_qc_clr’. The error profile of Badread (v0.1.5) was set ‘--error_model nanopore’ and ‘--error_model pacbio’ for the respective simulations. The error profile of PaSS is located in the installation folder (‘PaSS/E.coli/ecoli.config’). We used ‘-m pacbio_sequel -c PaSS/E.coli/ecoli.config’ for the simulation of PaSS. LongISLND (v0.9.5) does not provide an error profile so we trained error profiles using the same datasets as with Sim-it (v1.0). DeepSimulator (v1.5) provides a Nanopore error profile in the installation folder. Simlord (v1.0.3) does not provide error profiles so we changed the parameters ‘Probability for insertions, deletions, substitutions’ according to the observed values of real PacBio datasets. The command is ‘simlord -ps 0.0312 -pd 0.0309 -pi 0.0433’. We downloaded the NanoSim (v2.6.0) error profile named ‘human_NA12878_DNA_FAB49712_albacore’ from its website. We downloaded the SURVIVOR (v1.0.7) error profile named ‘NA12878_nano_error_profile_bwa.txt.zip’ and ‘HG002_Pac_error_profile_bwa.txt.zip’ from its website.

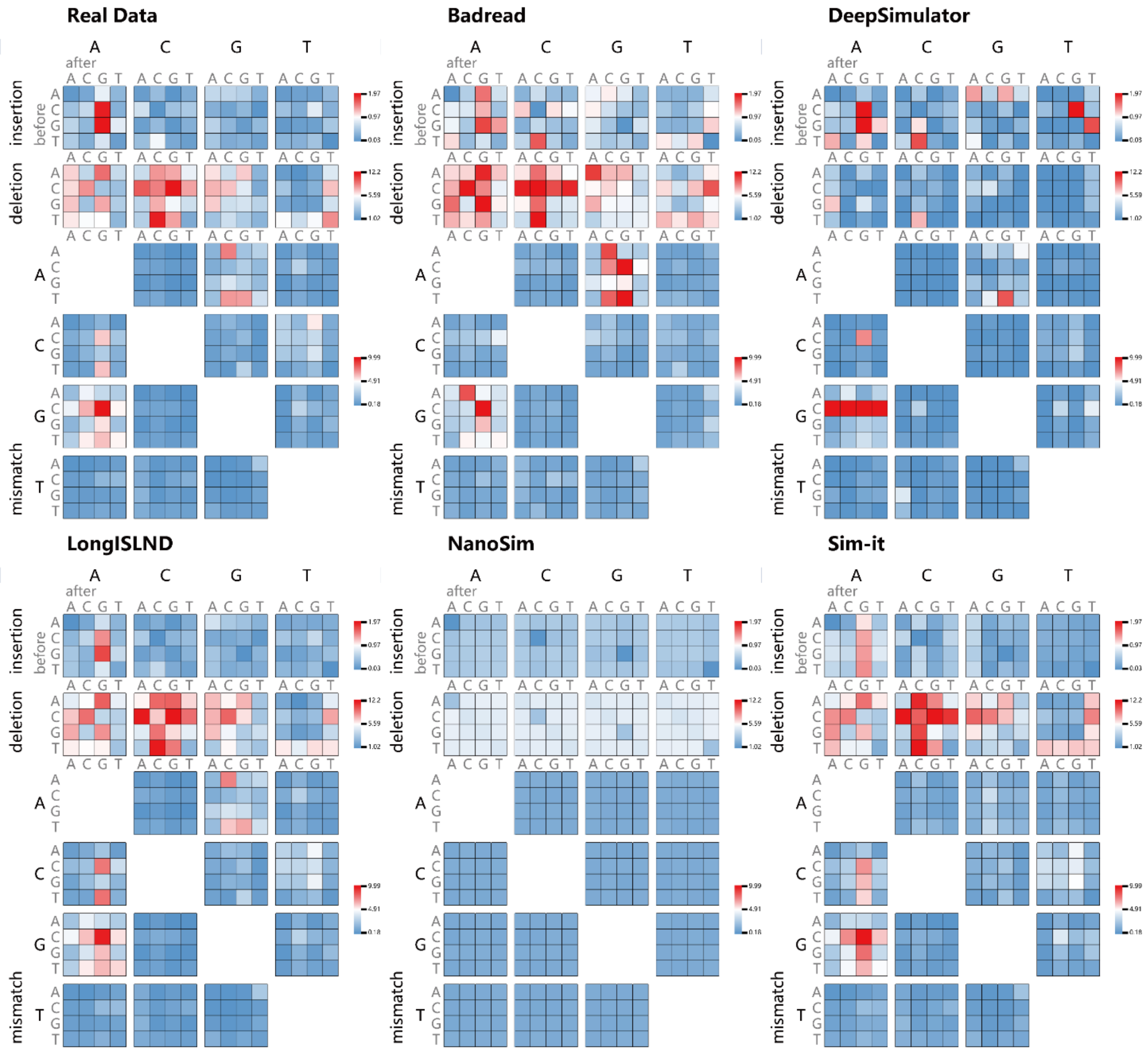


Figure S1 | Error profiles of simulated Nanopore reads from 5 different long read simulators.

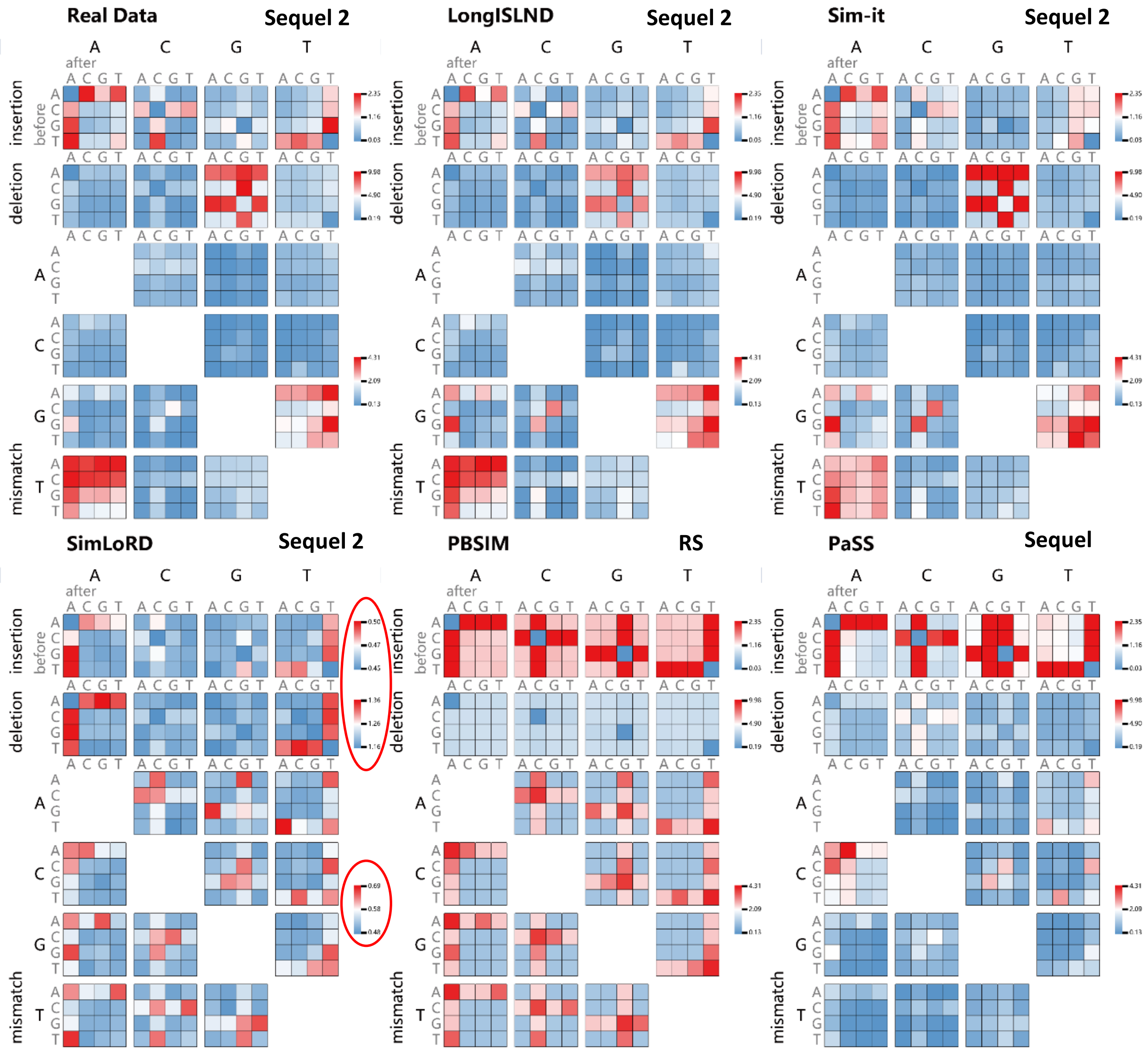


Figure S2 | Error profiles of simulated PacBio reads from 5 different long read simulators. The error rate for Simlord is much lower compared to the real data, we therefore adjusted the ratios to visualize the error profile.

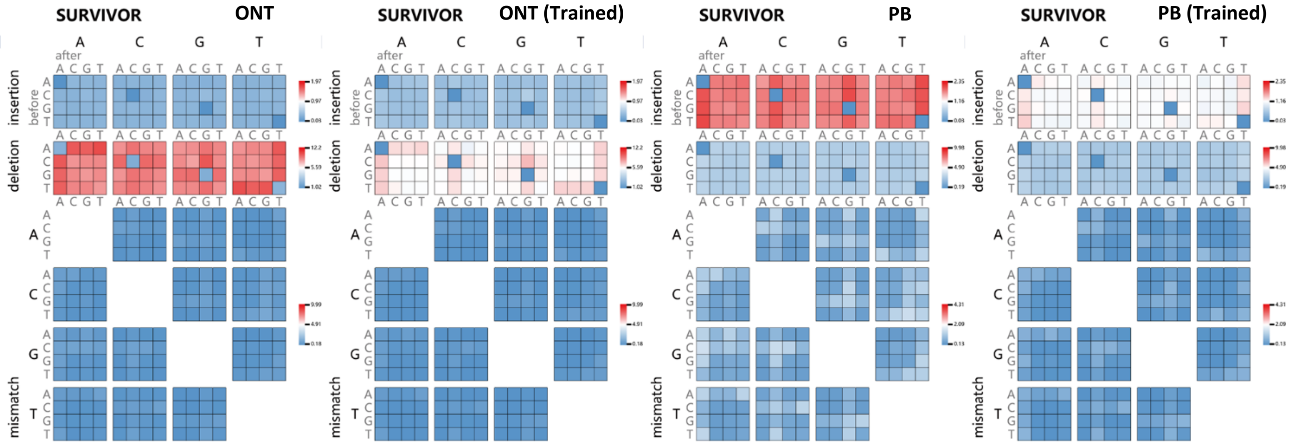


Figure S3 | Error profiles of simulated ONT and PacBio reads from the SURVIVOR simulator. For both ONT and PacBio, we used both the provided error profile and a trained error profile from the data sets we used to train Sim-it.

Table S2 | Statistics of the simulated reads for Nanopore for chromosome 1 of GRCh38.

	Match rate (%)	Insertion rate (%)	Deletion rate (%)	Substitution rate (%)	Totel error rate (%)	Average length (bp)
Real data	88,21	1,95	5,62	4,22	11,79	18.269
Sim-it	87,63	2,07	5,83	4,47	12,37	17.785
Badread	86,84	2,12	6,70	4,34	13,16	14.222
LongISLND	88,15	1,81	5,73	4,31	11,85	9.129
DeepSimulator	93,43	2,20	1,79	2,58	6,57	7.986
SURVIVOR	88,51	1,32	8,94	1,23	11,49	5.639
NanoSim	88,70	2,60	4,78	3,92	11,30	13.102

Table S3 | Statistics of the simulated reads of PacBio Sequel II for E. coli K12 substrain MG1655.

	Match rate (%)	Insertion rate (%)	Deletion rate (%)	Substitution rate (%)	Totel error rate (%)	Average length (bp)
Real data	89,50	4,29	3,09	3,12	10,50	11.478
Sim-it	89,06	4,29	3,00	3,65	10,94	11.839
PBSIM	82,02	9,37	3,02	5,59	17,98	3.035
Badread	87,25	3,67	5,34	3,74	12,75	14.939
PaSS	86,30	8,69	2,27	2,74	13,70	9.294
LongISLND	89,91	3,88	2,64	3,57	10,09	10.100
Simlord	95,52	1,76	1,11	1,61	4,48	8.155
SURVIVOR	89,31	6,57	2,48	1,64	10,69	6.435

Table S4 | Comparison of insertion and deletion lengths, and context-specific error patterns for mismatches and indels between simulated reads and the real sequencing data. Euclidean distances were calculated to compare simulated data for different simulators and real sequencing data. For each column and each dataset a color heatmap was adopted with blue indicating the most accurate caller.

	Methods	Mismatch patterns	Deletion patterns	Insertion patterns	Deletion lengths	Insertion lengths
		Euclidean distance	Euclidean distance	Euclidean distance	Euclidean distance	Euclidean distance
Nanopore (9.4.1) chromosome 1 of GRCh38	Sim-it	11.82	9.02	1.8	137.54	37.04
	Badread	15.11	13.35	1.92	497.99	337.3
	LongISLND	2.92	5.97	0.72	939.08	44.26
	DeepSimulator	24.5	34.31	6.48	973.54	107.58
	SURVIVOR	28.73	35.51	2.67	2574.67	143.59
	NanoSim	23.8	23.5	2.68	326.32	137.54
PacBio Sequel II E. coli K12 substrain MG1655	Sim-it	8.2	12.43	3.25	303.76	36.87
	PBSIM	20.04	20.5	12	101.3	2398.47
	Badread	16.69	32.78	3.15	852.44	299.44
	PaSS	16.39	22.1	9.77	551.74	1560.25
	LongISLND	4.04	5.39	1.06	180.57	147.26
	Simlord	15.33	25.64	5.82	1005.05	780.34
	SURVIVOR	14.88	20.37	7.73	279.73	1939.52

Table S5 | System requirements for each of the 6 tested SV callers. Each tool was used on the simulated dataset of Nanopore with 20x coverage and the GIAB dataset. SVIM cannot be run with multiple threads. Other SV callers are run with 24 threads

		Sniffles	cuteSV	pbsv	NanoVar	NanoSV	SVIM	combiSV	SURVIVOR
ONT simulted (20x)	Runtime (min)	64.0	5.0	172.0	62.0	66.0	31.0	<1	<1
	Peak memory (GB)	3.3	15.6	13,3	4.0	8.5	1.0	<1	<1
ONT GIAB (45x)	Runtime (min)	288.0	54.0	201.0	169.0	2913.0	199.0	2.0	<1
	Peak memory (GB)	10.9	23.1	65.4	53.2	39.9	2.8	<1	<1

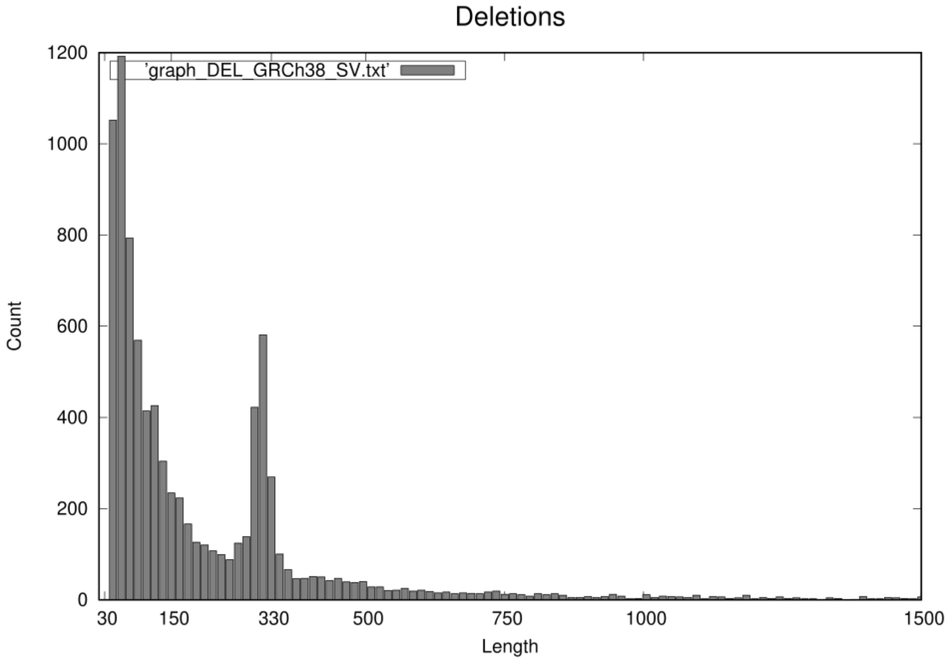


Figure S4 | Graphical output of the deletion length distribution from the structural variation simulation of the 24,600 SVs derived from sample NA19240 of dbVAR nstd152.

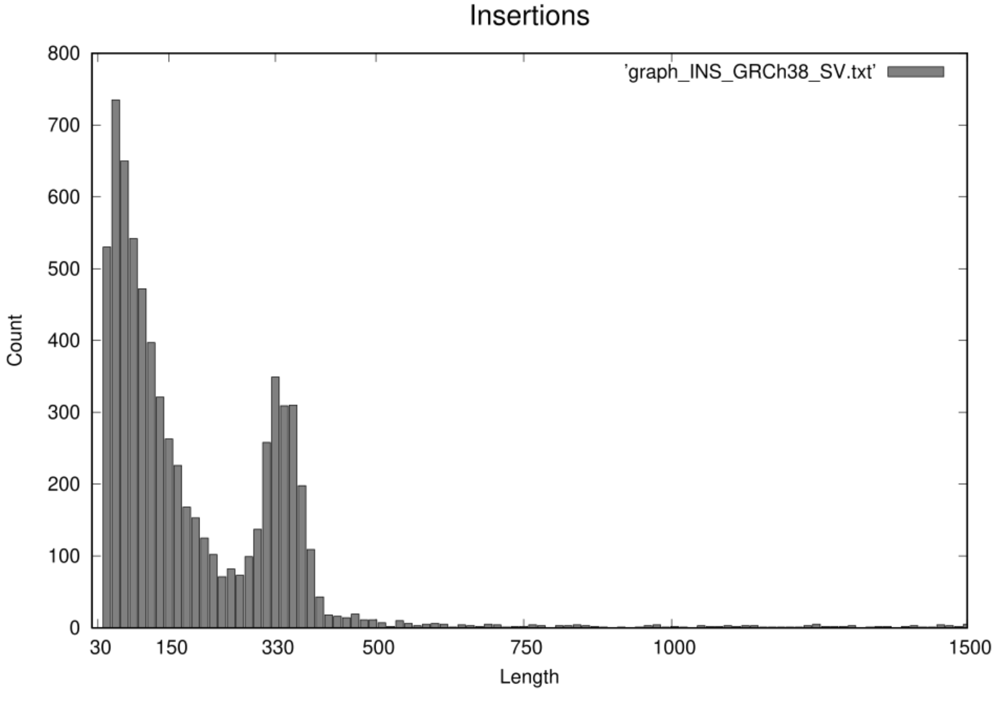


Figure S5 | Graphical output of the insertion length distribution from the structural variation simulation of the 24,600 SVs derived from sample NA19240 of dbVAR nstd152.

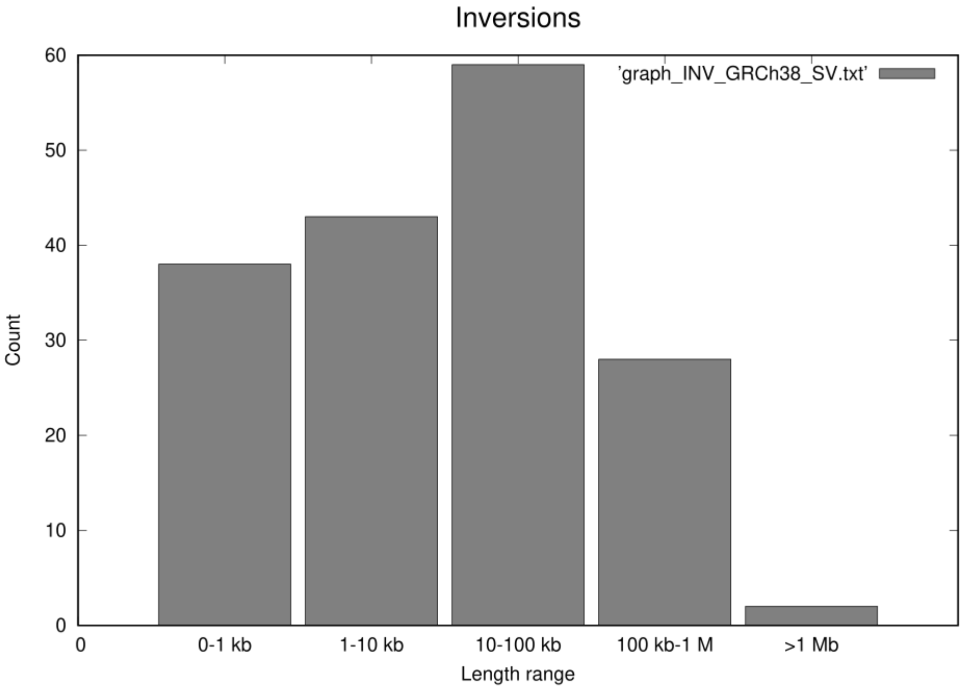


Figure S6 | Graphical output of the inversion length distribution from the structural variation simulation of the 24,600 SVs derived from sample NA19240 of dbVAR nstd152.

Table S6 | Comparison between combiSV and SURVIVOR for 9 combinations of three SV callers on a simulated Nanopore dataset of 20x and the GIAB reference dataset (Nanopore). The highest scores between combiSV and SURVIVOR are indicated in grey.

			cuteSV **Sniffles** **NanoSV**	**cuteSV** **Sniffles** **NanoVar**	**cuteSV** **Sniffles** **SVIM**	**cuteSV** **pbsv** **NanoSV**	**cuteSV** **pbsv** **NanoVar**	**cuteSV** **pbsv** **SVIM**	**cuteSV** **pbsv** **Sniffles**	**cuteSV** **NanoSV** **SVIM**	**SVIM** **NanoSV** **NanoVar**
Simulation (Nanopore)	**combiSV**	Recall	81.3%	80.7%	80.6%	79.7%	79.4%	79.8%	80.4%	79.3%	80.1%
		Precision	98.0%	98.6%	98.4%	98.7%	98.8%	98.7%	98.5%	98.7%	97.5%
		F-score	**88.9%**	**88.8%**	**88.6%**	**88.2%**	**88.0%**	**88.2%**	**88.6%**	**87.9%**	**87.9%**
		Perfect matches	8.0%	7.9%	8.0%	31.6%	31.6%	31.6%	31.7%	8.0%	3.6%
		Position score	85.2%	85.2%	85.2%	88.0%	88.0%	88.1%	88.6%	84.7%	85.4%
		Length score	87.9%	87.9%	87.9%	91.7%	91.8%	91.8%	92.0%	87.9%	87.4%
		Type score	92.7%	93.8%	93.8%	94.2%	94.3%	94.3%	93.7%	94.4%	94.2%
		Genotype score	95.2%	94.0%	94.7%	94.9%	95.2%	95.2%	95.3%	95.8%	95.1%
		Total score	**72.6%**	**72.0%**	**72.0%**	**71.8%**	**71.7%**	**71.9%**	**72.5%**	**69.9%**	**69.6%**
	SURVIVOR	Recall	79.3%	75.2%	77.8%	77.9%	75.7%	77.9%	72.0%	78.5%	78.7%
		Precision	97.9%	98.4%	97.7%	98.4%	98.8%	97.8%	98.4%	97.5%	97.7%
		F-score	**87.6%**	**85.2%**	**86.6%**	**87.0%**	**85.7%**	**86.7%**	**83.2%**	**87.0%**	**87.1%**
		Perfect matches	6.0%	4.9%	5.3%	6.0%	5.0%	5.4%	18.6%	0.4%	2.8%
		Position score	81.2%	78.5%	78.6%	80.4%	78.3%	78.4%	88.0%	85.0%	85.0%
		Length score	86.7%	88.6%	86.8%	87.5%	89.7%	88.2%	88.4%	86.2%	87.4%
		Type score	79.9%	94.6%	94.2%	81.0%	94.3%	94.2%	93.6%	49.8%	50.4%
		Genotype score	80.7%	92.6%	93.2%	82.3%	94.6%	94.5%	92.5%	51.2%	51.8%
		Total score	**63.3%**	**63.8%**	**65.3%**	**62.9%**	**65.0%**	**65.8%**	**63.7%**	**54.1%**	**54.8%**
GIAB (Nanopore)	**combiSV**	Recall	93.5%	93.6%	93.6%	94.4%	95.1%	95.4%	95.4%	93.8%	85.0%
		Precision	93.9%	93.6%	93.3%	92.8%	92.1%	91.6%	92.7%	90.1%	91.7%
		F-score	**93.7%**	**93.6%**	**93.5%**	**93.6%**	**93.6%**	**93.4%**	**94.0%**	**91.9%**	**88.2%**
		Perfect matches	0.4%	0.4%	0.4%	25.6%	25.4%	25.4%	25.4%	6.6%	5.2%
		Position score	66.3%	66.2%	66.2%	72.7%	72.3%	72.3%	72.3%	66.3%	70.1%
		Length score	79.7%	79.7%	79.7%	86.6%	86.2%	86.0%	85.5%	79.3%	80.7%
		Type score	98.8%	98.7%	98.8%	99.2%	98.7%	98.7%	98.7%	98.7%	95.6%
		Genotype score	97.4%	91.7%	97.1%	97.2%	96.0%	96.4%	96.9%	94.1%	92.5%
		Total score	**70.3%**	**69.0%**	**69.7%**	**73.6%**	**72.8%**	**72.4%**	**73.7%**	**65.6%**	**61.8%**
	SURVIVOR	Recall	80.9%	91.6%	93.8%	81.3%	91.0%	94.2%	93.5%	93.8%	90.8%
		Precision	94.1%	92.6%	77.4%	95.3%	92.9%	77.8%	93.2%	78.1%	90.4%
		F-score	**87.0%**	**92.1%**	**84.8%**	**87.8%**	**92.0%**	**85.2%**	**93.3%**	**85.2%**	**90.6%**
		Perfect matches	2.9%	1.7%	1.8%	2.6%	1.6%	1.8%	2.3%	1.7%	1.0%
		Position score	64.6%	60.0%	57.2%	64.6%	60.7%	57.4%	60.1%	58.6%	60.8%
		Length score	79.5%	81.5%	78.5%	81.9%	83.1%	80.8%	82.7%	78.3%	79.5%
		Type score	97.9%	97.9%	98.6%	98.6%	94.6%	98.8%	98.7%	98.5%	98.0%
		Genotype score	86.6%	77.1%	79.8%	90.8%	90.1%	91.0%	88.4%	89.1%	84.7%
		Total score	**58.6%**	**61.7%**	**42.3%**	**61.1%**	**63.9%**	**45.7%**	**66.1%**	**45.5%**	**60.0%**

4. Complex substitutions in NA19240

The recall of complex substitutions (CSUB) are significantly higher for the real PacBio dataset (60%) of NA19240 than for our simulated datasets (1-20%). Because we expected a drop in recall we examined the alignment of 27 CSUBs manually with IGV. We selected only homozygous CSUBs to simplify the interpretation of the alignments. CSUBs were not selected on any other criteria, we selected 12 random homozygous CSUBs from chromosome 1 and 15 random homozygous CSUBs of chromosome 2. For the simulated CSUBs, the length of the deleted sequences were always the same as the length of the inserted sequences. This is not necessarily the case for the real CSUBs, which could partially explain the discrepancy between the recall values and different alignment patterns. Nevertheless, several of the CSUBs we examined were in fact deletions or insertions that were incorrectly categorized as a CSUB. From only examining the alignments, we could only confirm one CSUB out of the 27 potential CSUBs as a true CSUB. For 6 presumed CSUBs, we aligned and inspected several individual PacBio reads separately. When the alignment does not show any SV at the given position, it also possible that called position was inaccurate. For each of the screenshots of IGV, the top alignment is the simulation and the bottom the true dataset of NA19240.

Figure 7: We did not observe a SV at this position by inspecting the alignments.

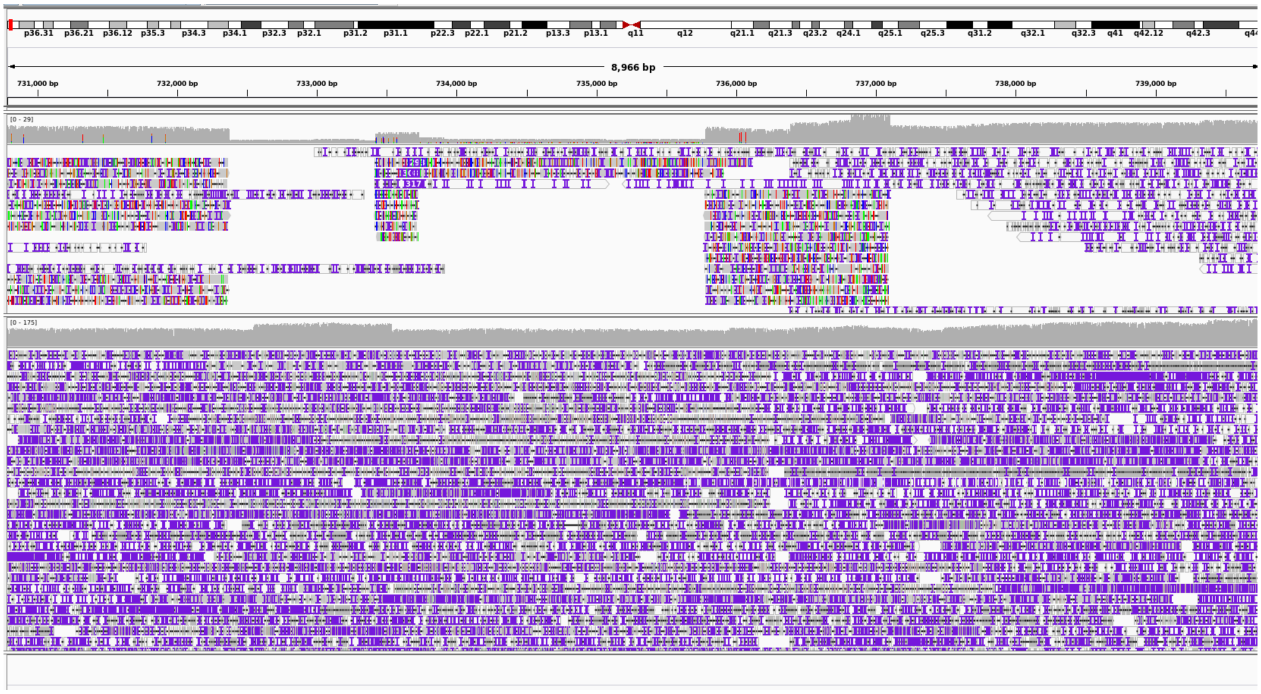

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3
8,966 bp
731,000 bp
732,000 bp
733,000 bp
734,000 bp
735,000 bp
736,000 bp
737,000 bp
738,000 bp
739,000 bp
[0 - 29]
[0 - 175]

Figure S7 | CSUB chromosome 1, position 732377, length 3402 bp

Figure 8: This is the only alignment that visually resembles a theoretical CSUB. SV callset ‘nstd152’ called 3 CSUBs of 61 bp within a region of 400 bp. This region is a tandem repeat region and when we individually aligned several PacBio reads, we found that there are two different haplotypes. Both haplotypes have a shorter tandem repeat region compared to the reference and for one haplotype we also observed an insertion of around 100 bp and can there be categorized as a heterozygous CSUB.

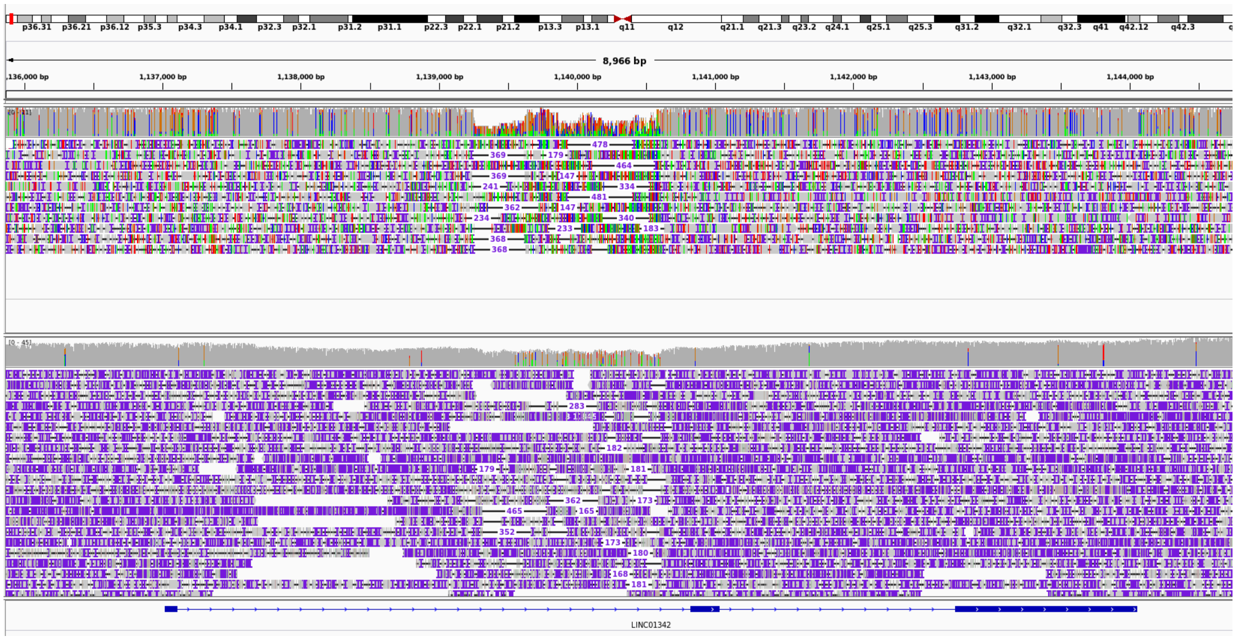

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3
8,966 bp
1,136,000 bp
1,137,000 bp
1,138,000 bp
1,139,000 bp
1,140,000 bp
1,141,000 bp
1,142,000 bp
1,143,000 bp
1,144,000 bp
[0 - 45]
LINC01342

Figure S8 | CSUB chromosome 1, position 1140181, length 61 bp

Figure 9 & 10: This SV is a confirmed deletion, as can be seen in the overall alignment and the individual PacBio alignments.

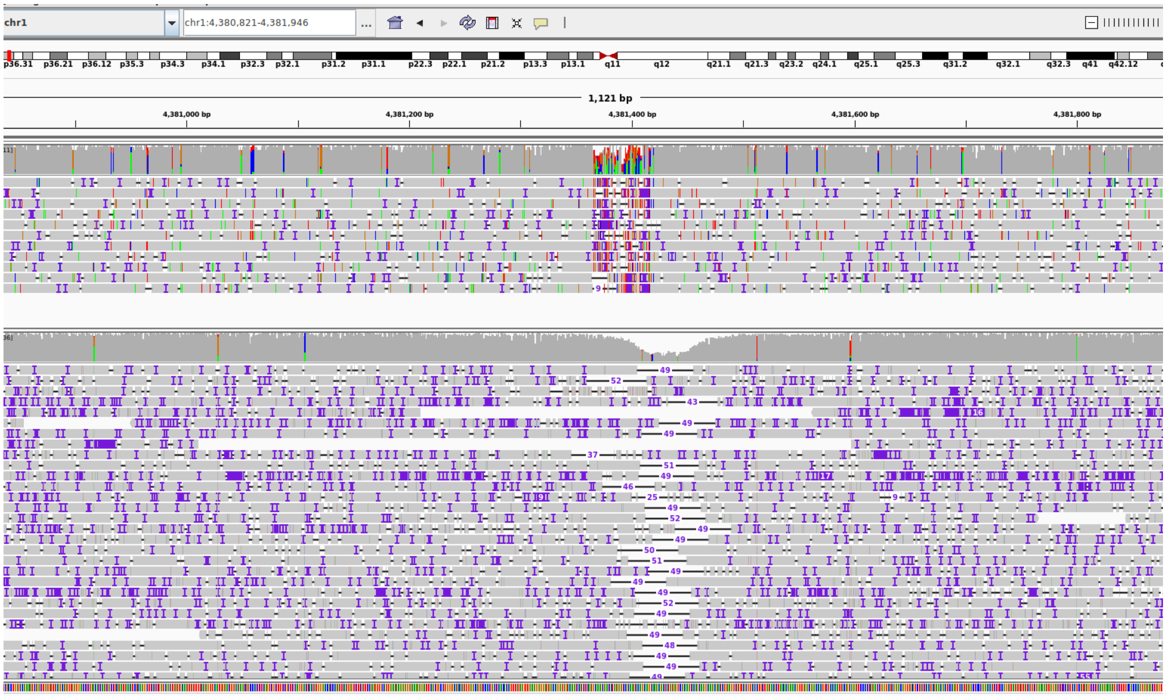
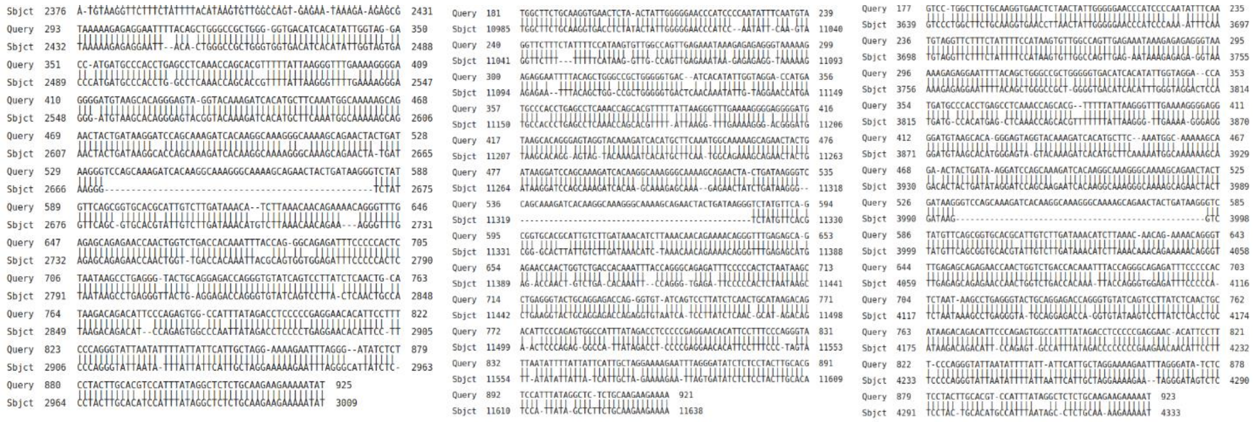


Figure S9 | CSUB chromosome 1, position 4381366, length 51 bp.



```
Sbjct  2376   A-TGTAAGGTTCTTTCTATTTTACATAAGTGTTGGCCAGT-GAGAA-TAAAGA-AGAGCG  2431
Query  293    TAAAAAGAGAGGAATTTTACAGCTGGGCCGCTGGG-GGTGACATCACATATTGGTAG-GA  350
Sbjct  2432   TAAAAAGAGAGGAATT--ACA-CTGGGCCGCTGGGTGGTGACATCACATATTGGTAGTGA  2488
Query  351    CC-ATGATGCCCACCTGAGCCTCAAACCAGCACGTTTTTATTAAGGGTTTGAAAAGGGGA  409
Sbjct  2489   CCCATGATGCCCACCTG-GCCTCAAACCAGCACCGTTTTATTAAGGGTTTTGAAAAGGGA  2547
Query  410    GGGGATGTAAGCACAGGGAGTA-GGTACAAAGATCACATGCTTCAAATGGCAAAAAGCAG  468
Sbjct  2548   GGG-ATGTAAGCACAGGGAGTACGGTACAAAGATCACATGCTTCAAATGGCAAAAAGCAG  2606
Query  469    AACTACTGATAAGGATCCAGCAAAGATCACAAGGCAAAGGGCAAAAGCAGAACTACTGAT  528
Sbjct  2607   AACTACTGATAAGGCACCAGCAAAGATCACAAGGCAAAAGGGCAAAGCAGAACTA-TGAT  2665
Query  529    AAGGGTCCAGCAAAGATCACAAGGCAAAGGGCAAAAGCAGAACTACTGATAAGGGTCTAT  588
Sbjct  2666   AAGGG-----------------------------------------------------TCTAT  2675
Query  589    GTTCAGCGGTGCACGCATTGTCTTGATAAACA--TCTTAAACAACAGAAAACAGGGTTTG  646
Sbjct  2676   GTTCAGC-GTGCACGTATTGTCTTGATAAACATGTCTTAAACAACAGAA---AGGGTTTG  2731
Query  647    AGAGCAGAGAACCAACTGGTCTGACCACAAATTTACCAG-GGCAGAGATTTCCCCCACTC  705
Sbjct  2732   AGAGCAGAGAACCAACTGGT-TGACCACAAATTACGCAGTGGTGGAGATTTCCCCCACTC  2790
Query  706    TAATAAGCCTGAGGG-TACTGCAGGAGACCAGGGTGTATCAGTCCTTATCTCAACTG-CA  763
Sbjct  2791   TAATAAGCCTGAGGGTTACTG-AGGAGACCAGGGTGTATCAGTCCTTA-CTCAACTGCCA  2848
Query  764    TAAGACAGACATTCCCAGAGTGG-CCATTTATAGACCTCCCCCGAGGAACACATTCCTTT  822
Sbjct  2849   TAAGACAGACAT--CCAGAGTGGCCCAATTATAGACCTCCCCTGAGGAACACATTCC-TT  2905
Query  823    CCCAGGGTATTAATATTTTATTATTCATTGCTAGG-AAAAGAATTTAGGG--ATATCTCT  879
Sbjct  2906   CCCAGGGTATTAATA-TTTATTATTCATTGCTAGGAAAAAGAATTTAGGGCATTATCTC-  2963
Query  880    CCTACTTGCACGTCCATTTATAGGCTCTCTGCAAGAAGAAAAATAT  925
Sbjct  2964   CCTACTTGCACATCCATTTATAGGCTCTCTGCAAGAAGAAAAATAT  3009
```

```
Query  181    TGGCTTCTGCAAGGTGAACTCTA-ACTATTGGGGGAACCCATCCCCAATATTTCAATGTA  239
Sbjct  10985  TGGCTTCTGCAAGGTGACCTCTATACTATTGGGGGAACCCATCC--AATATT-CAA-GTA  11040
Query  240    GGTTCTTTCTATTTTCCATAAGTGTTGGCCAGTTGAGAAATAAAGAGAGAGGGTAAAAAG  299
Sbjct  11041  GGTTCTTT---TTTTCATAAG-GTTG-CCAGTTGAGAAATAA-GAGAGAGG-TAAAAAG  11093
Query  300    AGAGGAATTTTACAGCTGGGCCGCTGGGGGTGAC--ATCACATATTGGTAGGA-CCATGA  356
Sbjct  11094  AGAGAA--TTTACAGCTGG-CCGCTGGGGGTGACTCAACAAATATTG-TAGGAACCATGA  11149
Query  357    TGCCCACCTGAGCCTCAAACCAGCACGTTTTTATTAAGGGTTTGAAAAGGGGAGGGGATG  416
Sbjct  11150  TGCCACCCTGAGCCTCAAACCAGCACGTTTT-ATTAAGG-TTTGAAAAGGG-ACGGGATG  11206
Query  417    TAAGCACAGGGAGTAGGTACAAAGATCACATGCTTCAAATGGCAAAAAGCAGAACTACTG  476
Sbjct  11207  TAAGCACAGG-AGTAG-TACAAAGATCACATGCTTCAA-TGGCAGAAAGCAGAACTACTG  11263
Query  477    ATAAGGATCCAGCAAAGATCACAAGGCAAAGGGCAAAAGCAGAACTA-CTGATAAGGGTC  535
Sbjct  11264  ATAAGGATCCAGCAAAGATCACAA-GCAAAGAGCAAA--GAGAACTATCTGATAAGGG--  11318
Query  536    CAGCAAAGATCACAAGGCAAAGGGCAAAAGCAGAACTACTGATAAGGGTCTATGTTCA-G  594
Sbjct  11319  -------------------------------------------------TCTATGTTCACG  11330
Query  595    CGGTGCACGCATTGTCTTGATAAACATCTTAAACAACAGAAAACAGGGTTTGAGAGCA-G  653
Sbjct  11331  CGG-GCACTTATTGTCTTGATAAACATC-TAAACAACAGAAAACAGGGTTTGAGAGCATG  11388
Query  654    AGAACCAACTGGTCTGACCACAAATTTACCAGGGCAGAGATTTCCCCCACTCTAATAAGC  713
Sbjct  11389  AG-ACCAACT-GTCTGA-CACAAATT--CCAGGG-TGAGA-TTCCCCCACTCTAATAAGC  11441
Query  714    CTGAGGGTACTGCAGGAGACCAG-GGTGT-ATCAGTCCTTATCTCAACTGCATAAGACAG  771
Sbjct  11442  CTGAAGGTACTGCAGGAGACCAGAGGTGTAATCA-TCCTTATCTCAAC-GCAT-AGACAG  11498
Query  772    ACATTCCCAGAGTGGCCATTTATAGACCTCCCCCGAGGAACACATTCCTTTCCCAGGGTA  831
Sbjct  11499  A-ACTCCCAGAG-GGCCA-TTATAGACCT-CCCCGAGGAACACATTCCTTTCCC-TAGTA  11553
Query  832    TTAATATTTTATTATTCATTGCTAGGAAAAGAATTTAGGGATATCTCTCCTACTTGCACG  891
Sbjct  11554  TT-ATATATTATTA-TCATTGCTA-GAAAAGAA-TTAGTGATATCTCTCCTACTTGCACA  11609
Query  892    TCCATTTATAGGCTC-TCTGCAAGAAGAAAA  921
Sbjct  11610  TCCA-TTATA-GCTCTTCTGCAAGAAGAAAA  11638
```

```
Query  177    GTCC-TGGCTTCTGCAAGGTGAACTCTAACTATTGGGGGAACCCATCCCCAATATTTCAA  235
Sbjct  3639   GTCCCTGGCTTCTGCAAGGTGAACCTTAACTATTGGGGGAACCCATCCCAAA-ATTTCAA  3697
Query  236    TGTAGGTTCTTTCTATTTTCCATAAGTGTTGGCCAGTTGAGAAATAAAGAGAGAGGGTAA  295
Sbjct  3698   TGTAGGTTCTTTCTATTTTCCATAAGTGTTGGCCAGTTGAG-AATAAAGAGAGA-GGTAA  3755
Query  296    AAAGAGAGGAATTTTACAGCTGGGCCGCTGGGGGTGACATCACATATTGGTAGGA--CCA  353
Sbjct  3756   AAAGAGAGGAATTTTACAGCTGGGCCGCT-GGGGTGACATCACATTTGGGTAGGACTCCA  3814
Query  354    TGATGCCCACCTGAGCCTCAAACCAGCACG--TTTTTATTAAGGGTTTGAAAAGGGGAGG  411
Sbjct  3815   TGATG-CCACATGAG-CTCAAACCAGCACGTTTTTTTATTAAGGG-TTGAAAA-GGGAGG  3870
Query  412    GGATGTAAGCACA-GGGAGTAGGTACAAAGATCACATGCTTC--AAATGGC-AAAAAGCA  467
Sbjct  3871   GGATGTAAGCACATGGGAGTA-GTACAAAGATCACATGCTTCAAAAATGGCAAAAAAGCA  3929
Query  468    GA-ACTACTGATA-AGGATCCAGCAAAGATCACAAGGCAAAGGGCAAAAGCAGAACTACT  525
Sbjct  3930   GACACTACTGATATAGGATCCAGCAAGAATCACAAGGCAAAGGGCAAAAGCAGAACTACT  3989
Query  526    GATAAGGGTCCAGCAAAGATCACAAGGCAAAGGGCAAAAGCAGAACTACTGATAAGGGTC  585
Sbjct  3990   GATAAG---------------------------------------------------GTC  3998
Query  586    TATGTTCAGCGGTGCACGCATTGTCTTGATAAACATCTTAAAC-AACAG-AAAACAGGGT  643
Sbjct  3999   TATGTTCAGCGGTGCACGTATTGTCTTGATAAACATCTTAAACAAACAGAAAAACAGGGT  4058
Query  644    TTGAGAGCAGAGAACCAACTGGTCTGACCACAAATTTACCAGGGCAGAGATTTCCCCCAC  703
Sbjct  4059   TTGAGAGCAGAGAACCAACTGGTCTGACCACAAA-TTACCAGGGTGGAGATTTCCCCCA-  4116
Query  704    TCTAAT-AAGCCTGAGGGTACTGCAGGAGACCAGGGTGTATCAGTCCTTATCTCAACTGC  762
Sbjct  4117   TCTAATAAAGCCTGAGGGTA-TGCAGGAGACCA-GGTGTATAAGTCCTTATCTCACCTGC  4174
Query  763    ATAAGACAGACATTCCCAGAGTGGCCATTTATAGACCTCCCCCGAGGAAC-ACATTCCTT  821
Sbjct  4175   ATAAGACAGACATT-CCAGAGT-GCCATTTATAGACCCCCCCGAAGAACAACATTCCTT  4232
Query  822    T-CCCAGGGTATTAATATTTTATT-ATTCATTGCTAGGAAAAGAATTTAGGGATA-TCTC  878
Sbjct  4233   TCCCCAGGGTATTAATATTTTATTAATTCATTGCTAGGAAAAGAA--TAGGGATAGTCTC  4290
Query  879    TCCTACTTGCACGT-CCATTTATAGGCTCTCTGCAAGAAGAAAAAT  923
Sbjct  4291   TCCTAC-TGCACATGCCATTTAATAGC-CTCTGCAA-AAGAAAAAT  4333
```

Figure S10 | Individual alignment of three PacBio reads for ‘CSUB chromosome 1, position 4381366, length 51 bp’.

Figure 11: Several reads show insertions around this position.

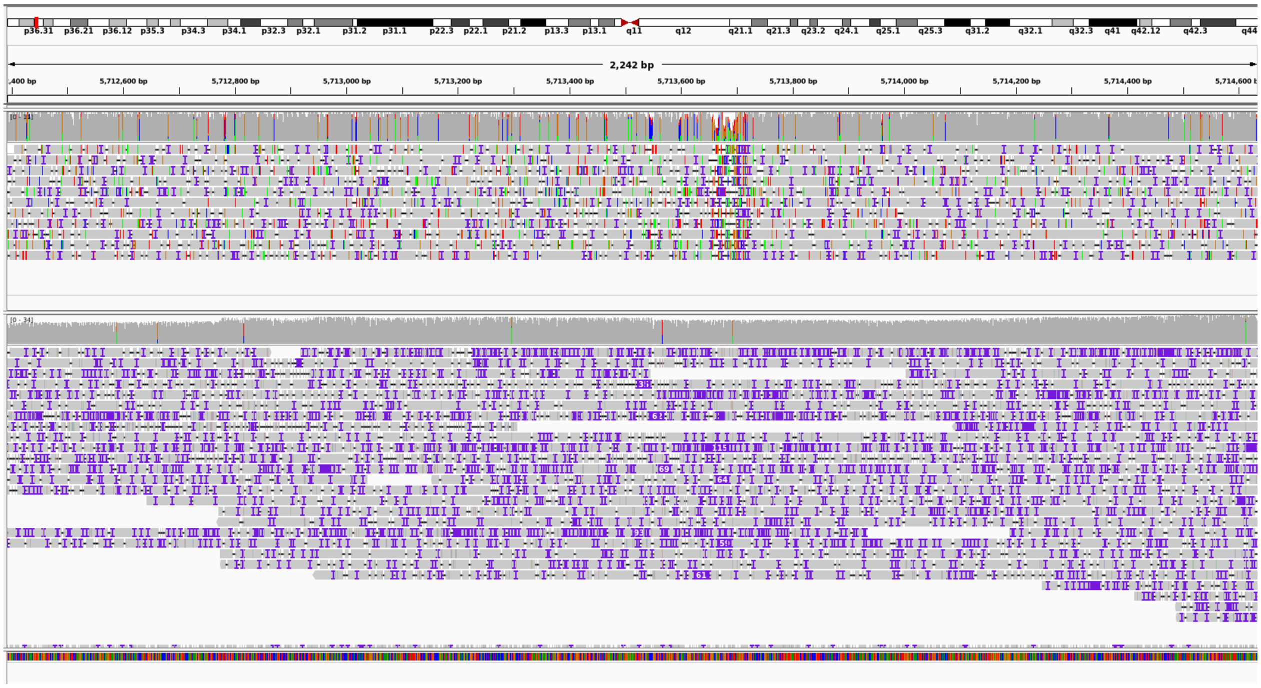

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3 q44
2,242 bp
,400 bp
5,712,600 bp
5,712,800 bp
5,713,000 bp
5,713,200 bp
5,713,400 bp
5,713,600 bp
5,713,800 bp
5,714,000 bp
5,714,200 bp
5,714,400 bp
5,714,600 b
[0 - 33]
[0 - 34]

Figure S11 | CSUB chromosome 1, position 5713658, length 62 bp

Figure 12 & 13: We did not observe a SV at either position by inspecting the alignments.

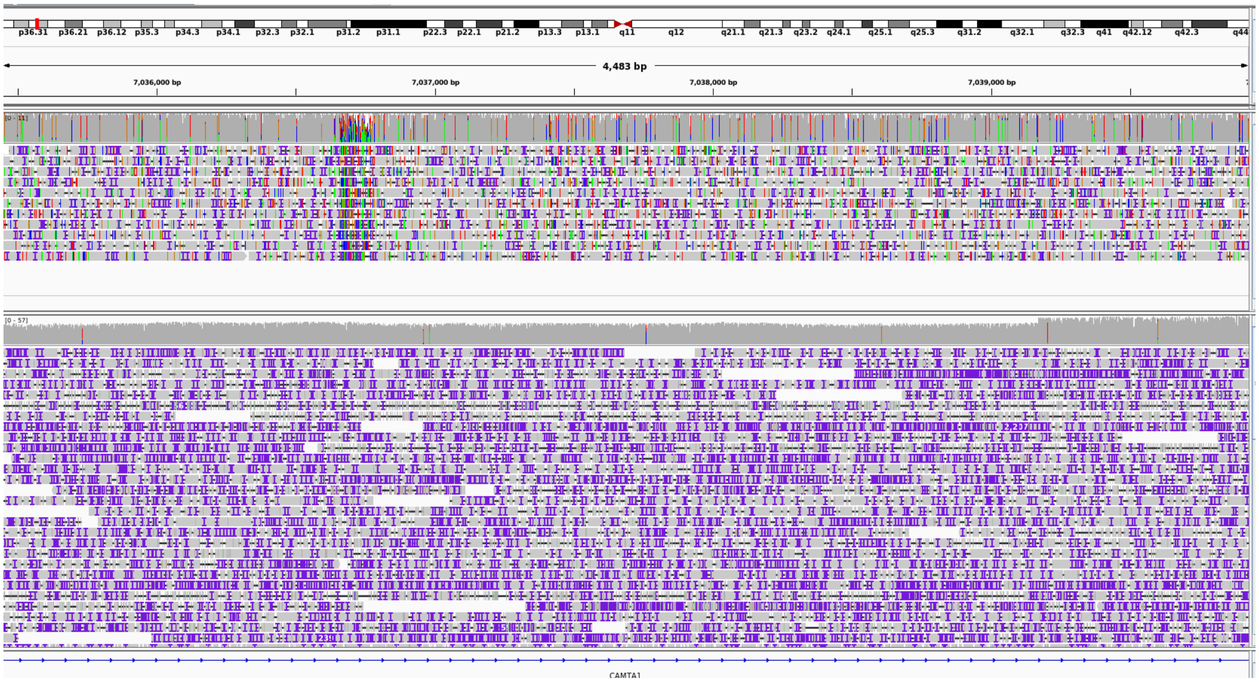

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3 q44
4,483 bp
7,036,000 bp
7,037,000 bp
7,038,000 bp
7,039,000 bp
[0 - 31]
[0 - 57]
CAMTA1

Figure S12 | CSUB chromosome 1, position 7036659, length 119 bp

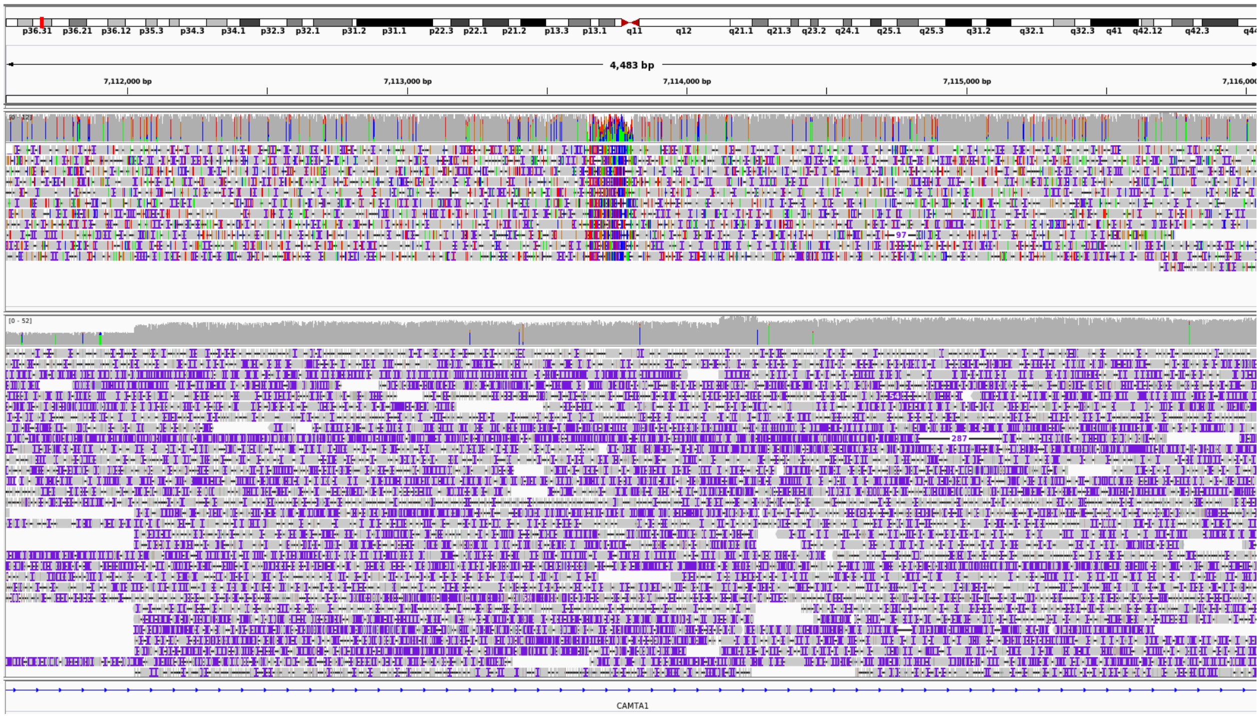

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3 q44
4,483 bp
7,112,000 bp
7,113,000 bp
7,114,000 bp
7,115,000 bp
7,116,000
[0 - 52]
287
97
CAMTA1

Figure S13 | CSUB chromosome 1, position 7113644, length 165 bp

Figure 14 & 15: This SV is a confirmed insertion, as can be seen in the overall alignment and the individual PacBio alignments.

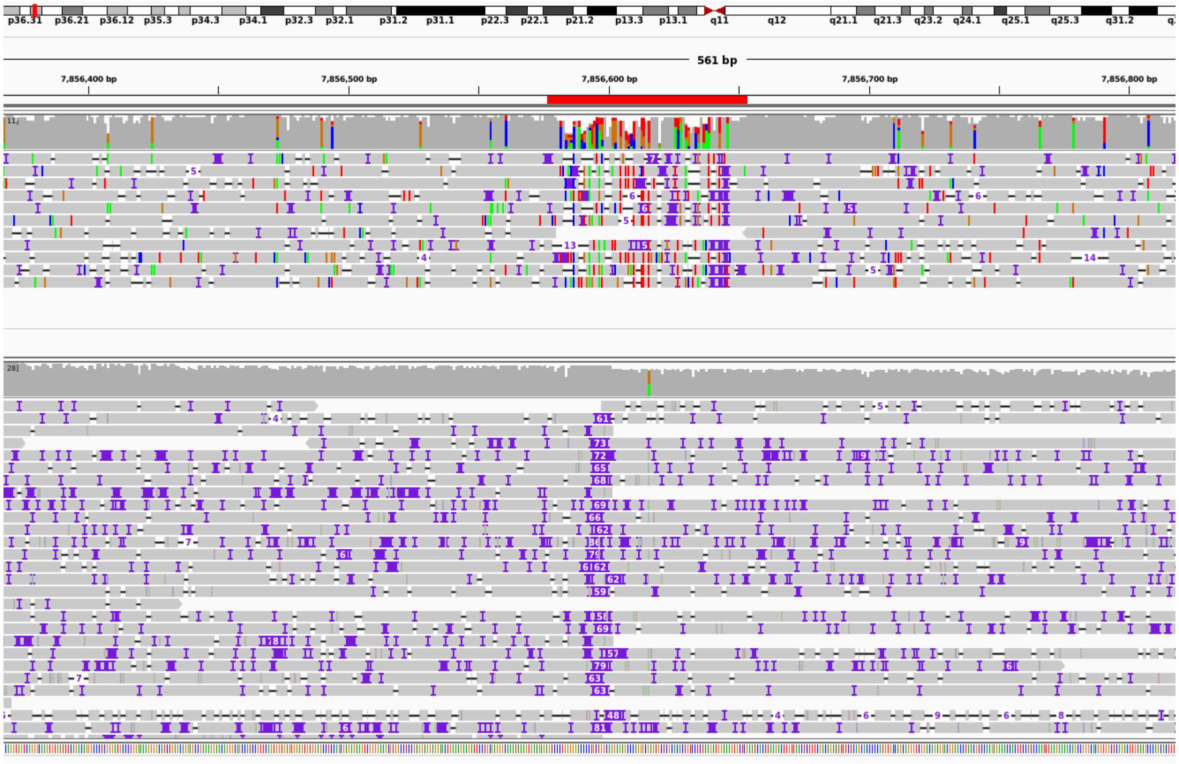
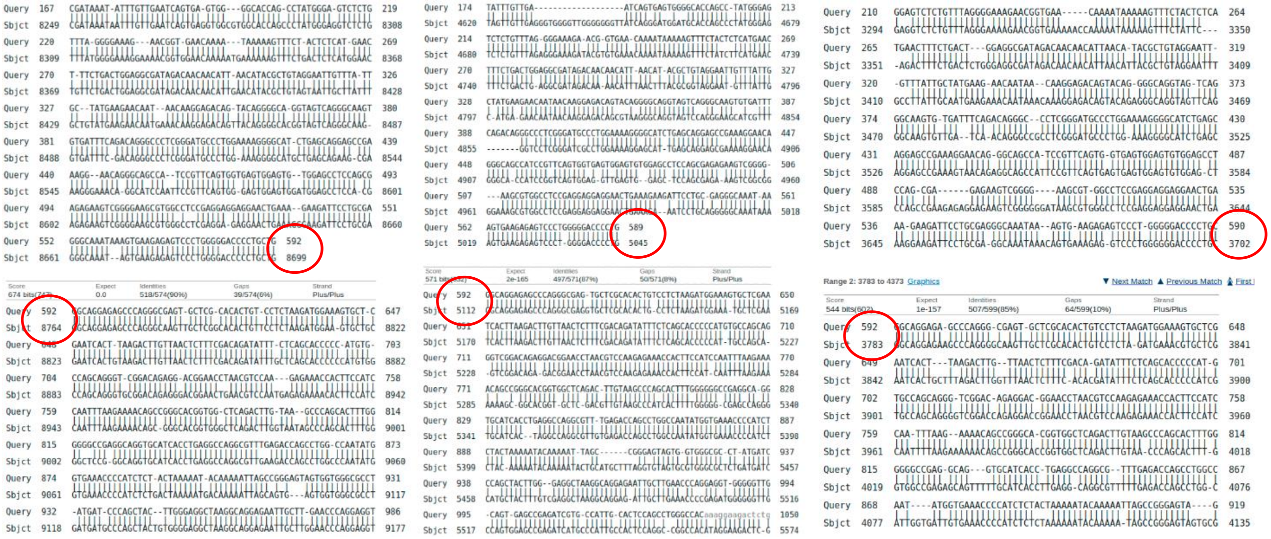

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2
561 bp
7,856,400 bp
7,856,500 bp
7,856,600 bp
7,856,700 bp
7,856,800 bp
28]

Figure S14 | CSUB chromosome 1, position 7856583, length 62 bp



```
Query  167   CGATAAAT-ATTTGTTGAATCAGTGA-GTGG---GGCACCAG-CCTATGGGA-GTCTCTG  219
Sbjct  8249  CGATAAATAATTTGTTGAATCAGTGAGGTGGCGTGGCACCAGCCCTATGGGAGGTCTCTG  8308
Query  220   TTTA-GGGGAAAG---AACGGT-GAACAAAA---TAAAAAGTTTCT-ACTCTCAT-GAAC  269
Sbjct  8309  TTTATGGGGAAAGGAAAACGGTGGAACAAAAATGAAAAAAGTTTCTGACTCTCATGGAAC  8368
Query  270   T-TTCTGACTGGAGGCGATAGACAACAACATT-AACATACGCTGTAGGAATTGTTTA-TT  326
Sbjct  8369  TGTTCTGACTGGAGGCGATAGACAACAACATTGAACATACGCTGTAGTAATTGCTTATTT  8428
Query  327   GC--TATGAAGAACAAT--AACAAGGAGACAG-TACAGGGGCA-GGTAGTCAGGGCAAGT  380
Sbjct  8429  GCTGTATGAAGAACAATGAAACAAGGAGACAGTTACAGGGGCACGGTAGTCAGGGCAAG-  8487
Query  381   GTGATTTCAGACAGGGCCCTCGGGATGCCCTGGAAAAGGGGCAT-CTGAGCAGGAGCCGA  439
Sbjct  8488  GTGATTTC-GACAGGGCCCTCGGGATGCCCTGG-AAAGGGGCATGCTGAGCAGAAG-CGA  8544
Query  440   AAGG--AACAGGGCAGCCA--TCCGTTCAGTGGTGAGTGGAGTG--TGGAGCCTCCAGCG  493
Sbjct  8545  AAGGGAAACA-GGCATCCAATTCCGTTCAGTGG-GAGTGGAGTGGATGGAGCCTCCA-CG  8601
Query  494   AGAGAAGTCGGGGAAGCGTGGCCTCCGAGGAGGAGGAACTGAAA--GAAGATTCCTGCGA  551
Sbjct  8602  AGAGAAGTCGGGGAAGCGTGGCCTCGAGGA-GAGGAACTGAAAGGCAAGATTCCTGCGA  8660
Query  552   GGGCAAATAAAGTGAAGAGAGTCCCTGGGGGACCCCTGCTG  592
Sbjct  8661  GGGCAAAT--AGTGAAGAGAGTCCCTGGGGACCCCCTGCTG  8699

Score            Expect   Identities     Gaps         Strand
674 bits(747)    0.0      518/574(90%)   39/574(6%)   Plus/Plus

Query  592   GGCAGGAGAGCCCAGGGCGAGT-GCTCG-CACACTGT-CCTCTAAGATGGAAAGTGCT-C  647
Sbjct  8764  GGCAGGAGAGCCCAGGGCAAGTTGCTCGGCACACTGTTCCTCTAAGATGGAA-GTGCTGC  8822
Query  648   GAATCACT-TAAGACTTGTTAACTCTTTCGACAGATATTT-CTCAGCACCCCC-ATGTG-  703
Sbjct  8823  GAATCACTGTAAGACTTGTTAACTCTTTCGACAGATATTTGCTCAGCACCCCCCATGTGG  8882
Query  704   CCAGCAGGGT-CGGACAGAGG-ACGGAACCTAACGTCCAA---GAGAAACCACTTCCATC  758
Sbjct  8883  CCAGCAGGGTGCGGACAGAGGGACGGAACTGAACGTCCAATGAGAGAAAACACTTCCATC  8942
Query  759   CAATTTAAGAAAACAGCCGGGCACGGTGG-CTCAGACTTG-TAA--GCCCAGCACTTTGG  814
Sbjct  8943  CAATTTAAGAAAACAGC-GGGCACGGTGGGCTCAGACTTGGTAATAGCCCAGCACTTTGG  9001
Query  815   GGGGCCGAGGCAGGTGCATCACCTGAGGCCAGGCGTTTGAGACCAGCCTGG-CCAATATG  873
Sbjct  9002  GGCTCCG-GGCAGGTGCATCACCTGAGGCCAGGCGTTGAAGACCAGCCTGGCCCAATATG  9060
Query  874   GTGAAACCCCATCTCT-ACTAAAAAT-ACAAAAATTAGCCGGGAGTAGTGGTGGGCGCCT  931
Sbjct  9061  GTGAAACCCCATCTCTGACTAAAAATGACAAAAATTAGCAGTG---AGTGGTGGGCGCCT  9117
Query  932   -ATGAT-CCCAGCTAC--TTGGGAGGCTAAGGCAGGAGAATTGCTT-GAACCCAGGAGGT  986
Sbjct  9118  GATGATGCCCAGCTACTGTGGGGAGGCTAAGGCAGGAGAATTGCTTGGAACCCAGGAGGT  9177
```

```
Query  174   TATTTGTTGA------------------ATCAGTGAGTGGGGCACCAGCC-TATGGGAG  213
Sbjct  4620  TAGTTGTTGAGGGTGGGGTTGGGGGGGTTATCAGGGATGGATGCACCAGCCCTATGGGAG  4679
Query  214   TCTCTGTTTAG-GGGAAAGA-ACG-GTGAA-CAAAATAAAAAGTTTCTACTCTCATGAAC  269
Sbjct  4680  TCTCTGTTTAGAGGGAAAGATACGTGTGAAACAAAATAAAAAGTTTCTATCTTCATGAAC  4739
Query  270   TTTCTGACTGGAGGCGATAGACAACAACATT-AACAT-ACGCTGTAGGAATTGTTTATTG  327
Sbjct  4740  TTTCTGACTG-AGGCGATAGACAA-AACATTTAACTTTACGCGGTAGGAAT-GTTTATTG  4796
Query  328   CTATGAAGAACAATAACAAGGAGACAGTACAGGGGCAGGTAGTCAGGGCAAGTGTGATTT  387
Sbjct  4797  C-ATGA-GAACAATAACAAGGAGACAGCGTAAGGGCAGGTAGTCCAGGGAAGCATCGTTT  4854
Query  388   CAGACAGGGCCCTCGGGATGCCCTGGAAAAGGGGCATCTGAGCAGGAGCCGAAAGGAACA  447
Sbjct  4855  -------GGTCCTCGGGATCGCCTGGAAAAGGAGCAT-TGAGCAGGAGCGAAAAGGAACA  4906
Query  448   GGGCAGCCATCCGTTCAGTGGTGAGTGGAGTGTGGAGCCTCCAGCGAGAGAAGTCGGGG-  506
Sbjct  4907  GGGCA-CCATCCGGTCAGTGGAG-GTTGAGTG--GAGC-TCCAGCGAGA-AAGTCGGCGG  4960
Query  507   ---AAGCGTGGCCTCCGAGGAGGAGGAACTGAAAGAAGATTCCTGC-GAGGGCAAAT-AA  561
Sbjct  4961  GGAAAGCGTGGCCTCCGAGGAGGAGGAACTGAAAGA--AATCCTGCAGGGGGCAAATAAA  5018
Query  562   AGTGAAGAGAGTCCCTGGGGGACCCCTG  589
Sbjct  5019  AGTGAAGAGAGTCCCT-GGGGACCCCTG  5045

Score            Expect   Identities     Gaps         Strand
571 bits(632)    2e-165   497/571(87%)   50/571(8%)   Plus/Plus

Query  592   GGCAGGAGAGCCCAGGGCGAG-TGCTCGCACACTGTCCTCTAAGATGGAAAGTGCTCGAA  650
Sbjct  5112  GGCAGGAGAGCCCAGGGCGAGGTGCTCGCACACTG-CCTCTAAGATGGAAA-TGCTCGAA  5169
Query  651   TCACTTAAGACTTGTTAACTCTTTCGACAGATATTTCTCAGCACCCCCATGTGCCAGCAG  710
Sbjct  5170  TCACTTAAGACTTGTTAACTCTTTCGACAGATATTTCTCAGCACCCCCAT-TGCCAGCA-  5227
Query  711   GGTCGGACAGAGGACGGAACCTAACGTCCAAGAGAAACCACTTCCATCCAATTTAAGAAA  770
Sbjct  5228  -GTCGGACAGA-GACGGAACCTAACGTCCAAGAGAAACCACTTCCAT-CAATTTAAGAAA  5284
Query  771   ACAGCCGGGCACGGTGGCTCAGAC-TTGTAAGCCCAGCACTTTGGGGGGCCGAGGCA-GG  828
Sbjct  5285  AAAAGC-GGCACGGT-GCTC-GACGTTGTAAGCCCATCACTTTTGGGGG-CGAGCCAGGG  5340
Query  829   TGCATCACCTGAGGCCAGGCGTT-TGAGACCAGCCTGGCCAATATGGTGAAACCCCATCT  887
Sbjct  5341  TGCATCAC--TAGGCCAGGCGTTGTGAGACCAGCCTGGCCAATATGGTGAAACCCCATCT  5398
Query  888   CTACTAAAAATACAAAAAT-TAGC------CGGGAGTAGTG-GTGGGCGC-CT-ATGATC  937
Sbjct  5399  CTAC-AAAAATACAAAAATACTGCATGCTTTAGGTGTAGTGCGTGGGCGCTCTGATGATC  5457
Query  938   CCAGCTACTTGG--GAGGCTAAGGCAGGAGAATTGCTTGAACCCAGGAGGT-GGGGGTTG  994
Sbjct  5458  CATGCTACTTTGTCGAGGCTAAGGCAGGAG-ATTGCTTGAAACCCCGAGATGGGGGGTTG  5516
Query  995   -CAGT-GAGCCGAGATCGTG-CCATTG-CACTCCAGCCTGGGCCACaaaggaagactctg  1050
Sbjct  5517  CCAGTGGAGCCGAGATCATGCCCATTGCCACTCCAGGC-CGGCCACATAGGAAGACTC-G  5574
```

```
Query  210   GGAGTCTCTGTTTAGGGGAAAGAACGGTGAA-----CAAAATAAAAAGTTTCTACTCTCA  264
Sbjct  3294  GAGGTCTCTGTTTAGGGAAAAGAACGGTGAAAAACCAAAAATAAAAAGTTTCTATTC---  3350
Query  265   TGAACTTTCTGACT---GGAGGCGATAGACAACAACATTAACA-TACGCTGTAGGAATT-  319
Sbjct  3351  -AGACTTTCTGACTCTGGGAGGCGATAGACAACAACATTAACATTACGCTGTAGGAATTT  3409
Query  320   -GTTTATTGCTATGAAG-AACAATAA--CAAGGAGACAGTACAG-GGGCAGGTAG-TCAG  373
Sbjct  3410  GCCTTATTGCAATGAAGAAACAATAAACAAAGGAGACAGTACAGAGGGCAGGTAGTTCAG  3469
Query  374   GGCAAGTG-TGATTTCAGACAGGGC--CCTCGGGATGCCCTGGAAAAGGGGCATCTGAGC  430
Sbjct  3470  GGCAAGTGTTGA--TCA-ACAGGGCCGCCTCGGGATGCCCTGG-AAAGGGGCATCTGAGC  3525
Query  431   AGGAGCCGAAAGGAACAG-GGCAGCCA-TCCGTTCAGTG-GTGAGTGGAGTGTGGAGCCT  487
Sbjct  3526  AGGAGCCGAAAGTAACAGAGGCAGCCATTCCGTTCAGTGAGTGAGTGGAGTGTGGAG-CT  3584
Query  488   CCAG-CGA------GAGAAGTCGGGG----AAGCGT-GGCCTCCGAGGAGGAGGAACTGA  535
Sbjct  3585  CCAGCCGAAGAGAGGAGAAGTCGGGGGGATAAGCGTGGGCCTCCGAGGAGGAGGAACTGA  3644
Query  536   AA-GAAGATTCCTGCGAGGGCAAATAA--AGTG-AAGAGAGTCCCT-GGGGGACCCCTGC  590
Sbjct  3645  AAGGAAGATTCCTGCGA-GGCAAATAAACAGTGAAAGAG-GTCCCTGGGGGGACCCCTGC  3702

Range 2: 3783 to 4373  Graphics        ▼ Next Match  ▲ Previous Match  First

Score            Expect   Identities     Gaps          Strand
544 bits(602)    1e-157   507/599(85%)   64/599(10%)   Plus/Plus

Query  592   GGCAGGAGA-GCCCAGGG-CGAGT-GCTCGCACACTGTCCTCTAAGATGGAAAGTGCTCG  648
Sbjct  3783  GGCAGGAGAAGCCCAGGGGCAAGTTGCTCGCACACTGTCCTCTA-GATGAAACGTGCTCG  3841
Query  649   AATCACT---TAAGACTTG--TTAACTCTTTCGACA-GATATTTCTCAGCACCCCCAT-G  701
Sbjct  3842  AATCACTGCTTTAGACTTGGTTTAACTCTTTC-ACACGATATTTCTCAGCACCCCCATCG  3900
Query  702   TGCCAGCAGGG-TCGGAC-AGAGGAC-GGAACCTAACGTCCAAGAGAAACCACTTCCATC  758
Sbjct  3901  TGCCAGCAGGGGTCGGACCAGAGGACCGGAACCTAACGTCCAAGAGAAACCACTTCCATC  3960
Query  759   CAA-TTTAAG--AAAACAGCCGGGCA-CGGTGGCTCAGACTTGTAAGCCCAGCACTTTGG  814
Sbjct  3961  CAATTTTAAGAAAAAAACAGCCGGGCACCGGTGGCTCAGACTTGTAA-CCCAGCACTTT-G  4018
Query  815   GGGGCCGAG-GCAG---GTGCATCACC-TGAGGCCAGGCG--TTTGAGACCAGCCTGGCC  867
Sbjct  4019  GTGGCCGAGAGCAGTTTTTGCATCACCTTGAGG-CAGGCGTTTTTGAGACCAGCCTGG-C  4076
Query  868   AAT----ATGGTGAAACCCCATCTCTACTAAAAATACAAAAATTAGCCGGGAGTA----G  919
Sbjct  4077  ATTGGTGATTGTGAAACCCCATCTCTCTAAAAAATACAAAAA-TAGCCGGGAGTAGTGCG  4135
```

Figure S15 | Individual alignment of three PacBio reads for ‘CSUB chromosome 1, position 7856583, length 62 bp’.

Figure 16 & 17: This SV is a confirmed deletion, as can be seen in the overall alignment and the individual PacBio alignments. By comparing individual PacBio reads with the reference we concluded that the first duplicated sequence was deleted in the NA19240 (first yellow sequence in Figure 13).

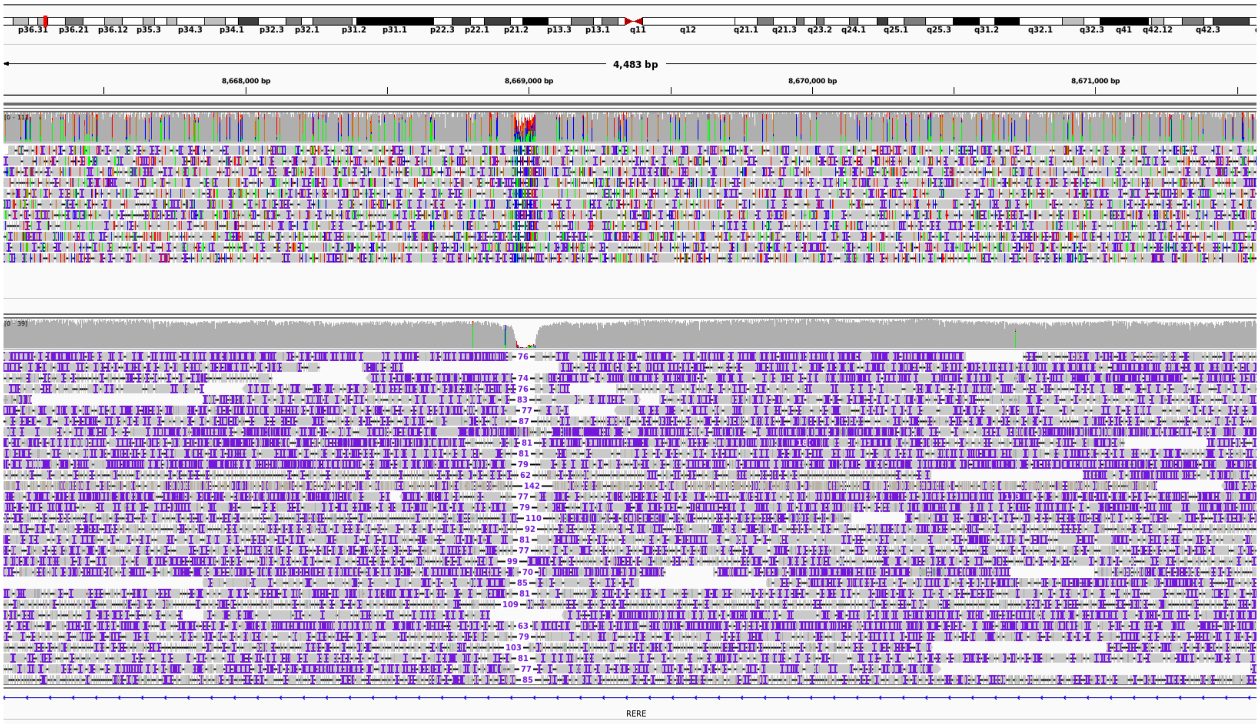
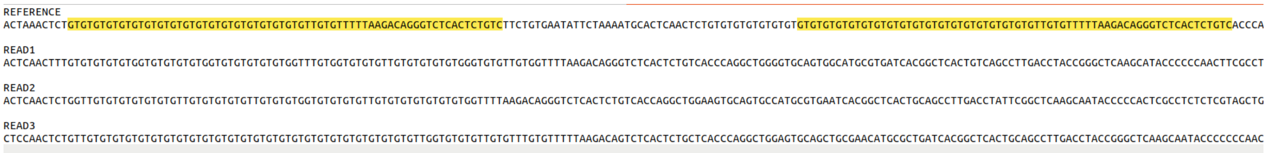


Figure S16 | CSUB chromosome 1, position 8668948, length 76 bp



```
REFERENCE
ACTAAACTCTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGTGTTTTTAAGACAGGGTCTCACTCTGTCTTCTGTGAATATTCTAAAATGCACTCAACTCTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGTGTTTTTAAGACAGGGTCTCACTCTGTCACCCA

READ1
ACTCAACTTTGTGTGTGTGTGGTGTGTGTGTGGTGTGTGTGTGGTTTGTGGTGTGTGTTGTGTGTGTGTGGGTGTGTTGTGGTTTTAAGACAGGGTCTCACTCTGTCACCCAGGCTGGGGTGCAGTGGCATGCGTGATCACGGCTCACTGTCAGCCTTGACCTACCGGGCTCAAGCATACCCCCAACTTCGCCT

READ2
ACTCAACTCTGGTTGTGTGTGTGTGTGTTGTGTGTGTGTTGTGTGTGGTGTGTGTGTTGTGTGTGTGTGTGGTTTTAAGACAGGGTCTCACTCTGTCACCAGGCTGGAAGTGCAGTGCCATGCGTGAATCACGGCTCACTGCAGCCTTGACCTATTCGGCTCAAGCAATACCCCCACTCGCCTCTCTCGTAGCTG

READ3
CTCCAACTCTGTTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGGTGTGTGTTGTGTTTGTGTTTTTAAGACAGTCTCACTCTGCTCACCCAGGCTGGAGTGCAGCTGCGAACATGCGCTGATCACGGCTCACTGCAGCCTTGACCTACCGGGCTCAAGCAATACCCCCCCAAC
```

Figure S17 | Individual alignment of three PacBio reads for ‘CSUB chromosome 1, position 8668948, length 76 bp’.

Figure 18 & 19: This SV is a confirmed insertion, as can be seen in the overall alignment and the individual PacBio alignments. Although it seems this insertion was heterozygous, as half of the reads do not show any inserted sequence.

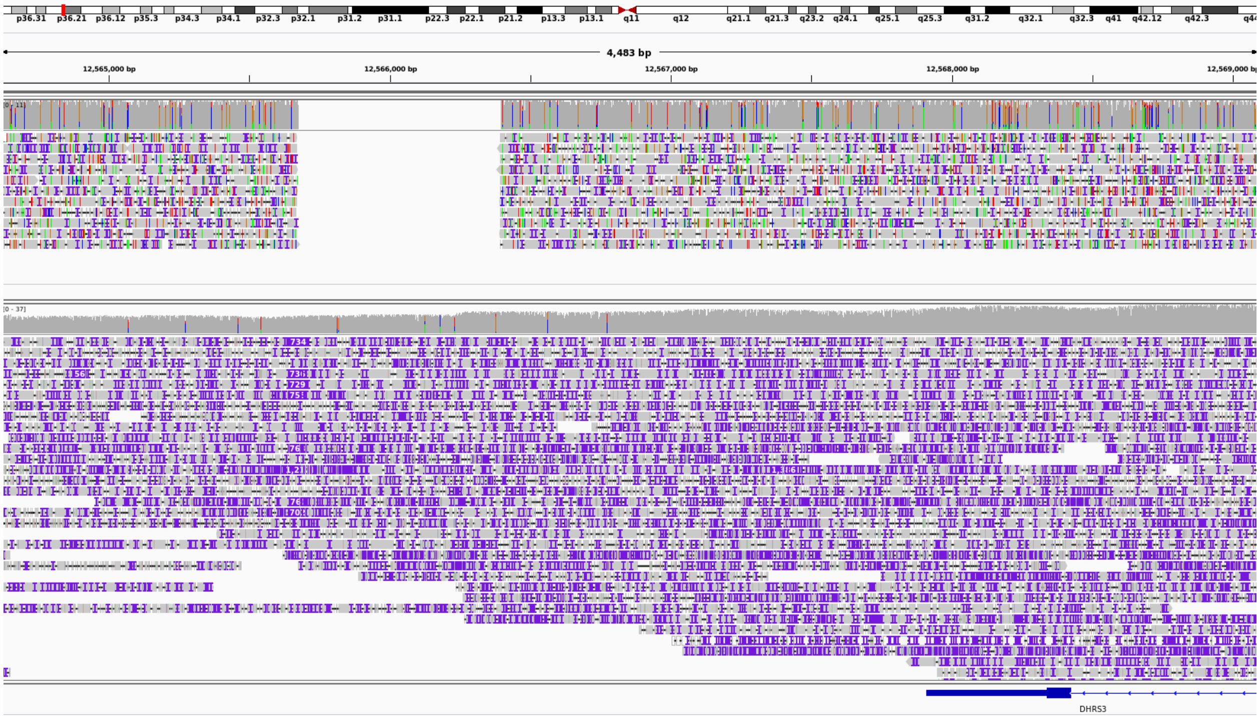
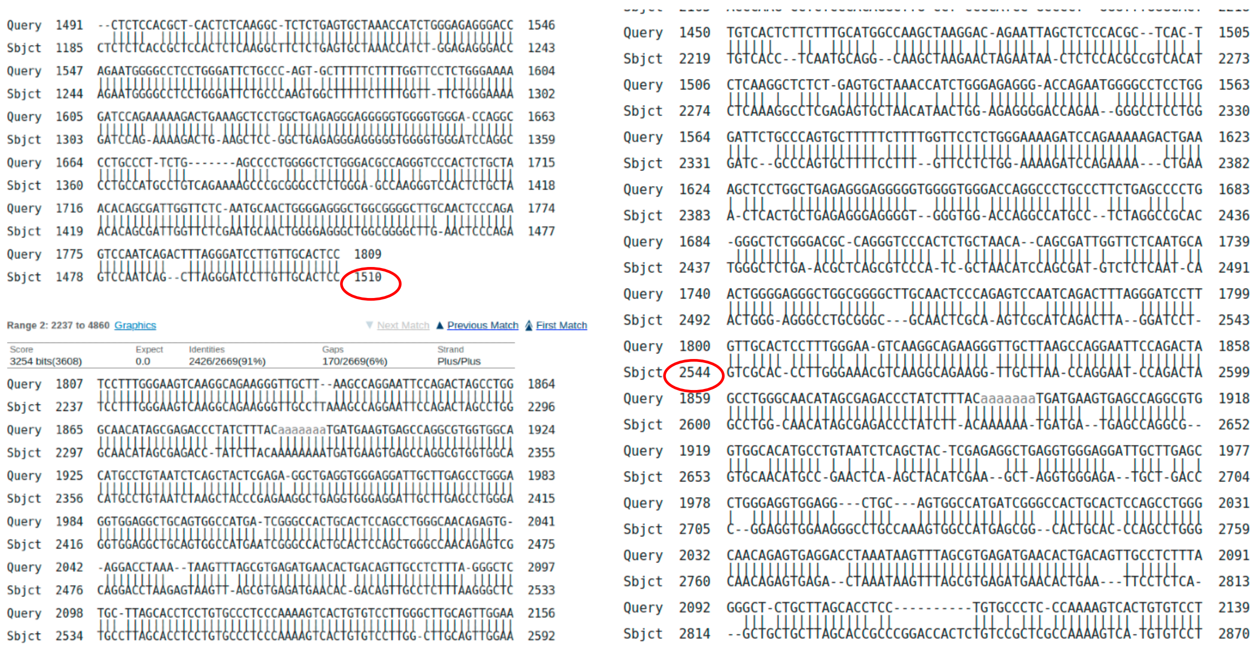


Figure S18 | CSUB chromosome 1, position 12565675, length 719 bp



```
Query  1491  --CTCTCCACGCT-CACTCTCAAGGC-TCTCTGAGTGCTAAACCATCTGGGAGAGGGACC  1546
Sbjct  1185  CTCTCTCACCGCTCCACTCTCAAGGCTTCTCTGAGTGCTAAACCATCT-GGAGAGGGACC  1243
Query  1547  AGAATGGGGCCTCCTGGGATTCTGCCC-AGT-GCTTTTTCTTTTGGTTCCTCTGGGAAAA  1604
Sbjct  1244  AGAATGGGGCCTCCTGGGATTCTGCCCAAGTGGCTTTTTCTTTTGGTT-TTCTGGGAAAA  1302
Query  1605  GATCCAGAAAAAGACTGAAAGCTCCTGGCTGAGAGGGAGGGGGTGGGGTGGGA-CCAGGC  1663
Sbjct  1303  GATCCAG-AAAAGACTG-AAGCTCC-GGCTGAGAGGGAGGGGGTGGGGTGGGATCCAGGC  1359
Query  1664  CCTGCCCT-TCTG-------AGCCCCTGGGGCTCTGGGACGCCAGGGTCCCACTCTGCTA  1715
Sbjct  1360  CCTGCCATGCCTGTCAGAAAAGCCCGCGGGCCTCTGGGA-GCCAAGGGTCCACTCTGCTA  1418
Query  1716  ACACAGCGATTGGTTCTC-AATGCAACTGGGGAGGGCTGGCGGGGCTTGCAACTCCCAGA  1774
Sbjct  1419  ACACAGCGATTGGTTCTCGAATGCAACTGGGGAGGGCTGGCGGGGCTTG-AACTCCCAGA  1477
Query  1775  GTCCAATCAGACTTTAGGGATCCTTGTTGCACTCC  1809
Sbjct  1478  GTCCAATCAG--CTTAGGGATCCTTGTTGCACTCC  1510
```

Range 2: 2237 to 4860 Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Identities	Gaps	Strand
3254 bits(3608)	0.0	2426/2669(91%)	170/2669(6%)	Plus/Plus

```
Query  1807  TCCTTTGGGAAGTCAAGGCAGAAGGGTTGCTT--AAGCCAGGAATTCCAGACTAGCCTGG  1864
Sbjct  2237  TCCTTTGGGAAGTCAAGGCAGAAGGGTTGCCTTAAAGCCAGGAATTCCAGACTAGCCTGG  2296
Query  1865  GCAACATAGCGAGACCCTATCTTTACaaaaaaaTGATGAAGTGAGCCAGGCGTGGTGGCA  1924
Sbjct  2297  GCAACATAGCGAGACC-TATCTTTACAAAAAAAATGATGAAGTGAGCCAGGCGTGGTGGCA  2355
Query  1925  CATGCCTGTAATCTCAGCTACTCGAGA-GGCTGAGGTGGGAGGATTGCTTGAGCCTGGGA  1983
Sbjct  2356  CATGCCTGTAATCTAAGCTACCCGAGAAGGCTGAGGTGGGAGGATTGCTTGAGCCTGGGA  2415
Query  1984  GGTGGAGGCTGCAGTGGCCATGA-TCGGGCCACTGCACTCCAGCCTGGGCAACAGAGTG-  2041
Sbjct  2416  GGTGGAGGCTGCAGTGGCCATGAATCGGGCCACTGCACTCCAGCTGGGCCAACAGAGTCG  2475
Query  2042  -AGGACCTAAA--TAAGTTTAGCGTGAGATGAACACTGACAGTTGCCTCTTTA-GGGCTC  2097
Sbjct  2476  CAGGACCTAAGAGTAAGTT-AGCGTGAGATGAACAC-GACAGTTGCCTCTTTAAGGGCTC  2533
Query  2098  TGC-TTAGCACCTCCTGTGCCCTCCCAAAAGTCACTGTGTCCTTGGGCTTGCAGTTGGAA  2156
Sbjct  2534  TGCCTTAGCACCTCCTGTGCCCTCCCAAAAGTCACTGTGTCCTTGG-CTTGCAGTTGGAA  2592
```

```
Query  1450  TGTCACTCTTCTTTGCATGGCCAAGCTAAGGAC-AGAATTAGCTCTCCACGC--TCAC-T  1505
Sbjct  2219  TGTCACC--TCAATGCAGG--CAAGCTAAGAACTAGAATAA-CTCTCCACGCCGTCACAT  2273
Query  1506  CTCAAGGCTCTCT-GAGTGCTAAACCATCTGGGAGAGGG-ACCAGAATGGGGCCTCCTGG  1563
Sbjct  2274  CTCAAAGGCCTCGAGAGTGCTAACATAACTGG-AGAGGGGACCAGAA--GGGCCTCCTGG  2330
Query  1564  GATTCTGCCCAGTGCTTTTTCTTTTGGTTCCTCTGGGAAAAGATCCAGAAAAAGACTGAA  1623
Sbjct  2331  GATC--GCCCAGTGCTTTTCCTTT--GTTCCTCTGG-AAAAGATCCAGAAAA---CTGAA  2382
Query  1624  AGCTCCTGGCTGAGAGGGAGGGGGTGGGGTGGGACCAGGCCCTGCCCTTCTGAGCCCCTG  1683
Sbjct  2383  A-CTCACTGCTGAGAGGGAGGGGT--GGGTGG-ACCAGGCCATGCC--TCTAGGCCGCAC  2436
Query  1684  -GGGCTCTGGGACGC-CAGGGTCCCACTCTGCTAACA--CAGCGATTGGTTCTCAATGCA  1739
Sbjct  2437  TGGGCTCTGA-ACGCTCAGCGTCCCA-TC-GCTAACATCCAGCGAT-GTCTCTCAAT-CA  2491
Query  1740  ACTGGGGAGGGCTGGCGGGGCTTGCAACTCCCAGAGTCCAATCAGACTTTAGGGATCCTT  1799
Sbjct  2492  ACTGGG-AGGGCCTGCGGGC---GCAACTCGCA-AGTCGCATCAGACTTA--GGATCCT-  2543
Query  1800  GTTGCACTCCTTTGGGAA-GTCAAGGCAGAAGGGTTGCTTAAGCCAGGAATTCCAGACTA  1858
Sbjct  2544  GTCGCAC-CCTTGGGAAACGTCAAGGCAGAAGG-TTGCTTAA-CCAGGAAT-CCAGACTA  2599
Query  1859  GCCTGGGCAACATAGCGAGACCCTATCTTTACaaaaaaaTGATGAAGTGAGCCAGGCGTG  1918
Sbjct  2600  GCCTGG-CAACATAGCGAGACCCTATCTT-ACAAAAAA-TGATGA--TGAGCCAGGCG--  2652
Query  1919  GTGGCACATGCCTGTAATCTCAGCTAC-TCGAGAGGCTGAGGTGGGAGGATTGCTTGAGC  1977
Sbjct  2653  GTGCAACATGCC-GAACTCA-AGCTACATCGAA--GCT-AGGTGGGAGA--TGCT-GACC  2704
Query  1978  CTGGGAGGTGGAGG---CTGC---AGTGGCCATGATCGGGCCACTGCACTCCAGCCTGGG  2031
Sbjct  2705  C--GGAGGTGGAAGGGCCTGCCAAAGTGGCCATGAGCGG--CACTGCAC-CCAGCCTGGG  2759
Query  2032  CAACAGAGTGAGGACCTAAATAAGTTTAGCGTGAGATGAACACTGACAGTTGCCTCTTTA  2091
Sbjct  2760  CAACAGAGTGAGA--CTAAATAAGTTTAGCGTGAGATGAACACTGAA---TTCCTCTCA-  2813
Query  2092  GGGCT-CTGCTTAGCACCTCC----------TGTGCCCTC-CCAAAAGTCACTGTGTCCT  2139
Sbjct  2814  --GCTGCTGCTTAGCACCGCCCGGACCACTCTGTCCGCTCGCCAAAAGTCA-TGTGTCCT  2870
```

Figure S19 | Individual alignment of two PacBio reads for ‘CSUB chromosome 1, position 12565675, length 719 bp’.

Figure 20: Unconfirmed SV.

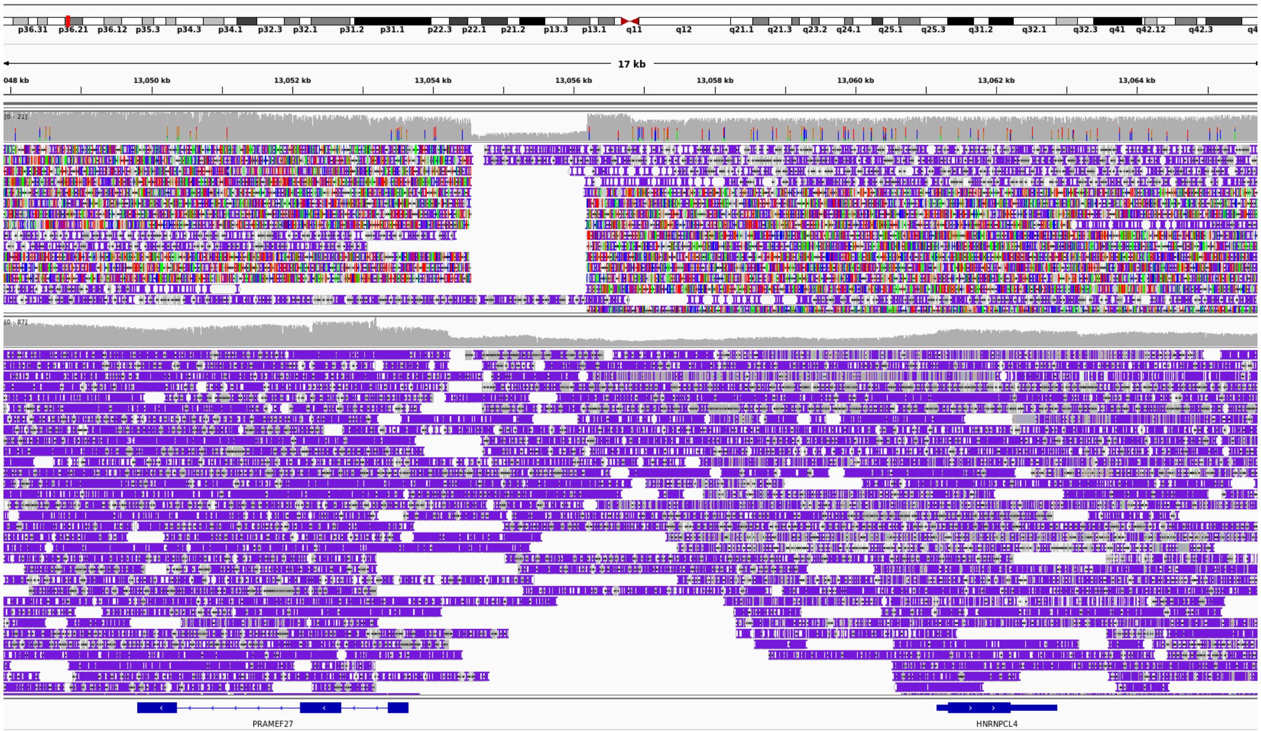


Figure S20 | CSUB chromosome 1, position 13054543, length 1646 bp

Figure 21: We observed an insertion, although this is a duplicated region which makes alignments less reliable.

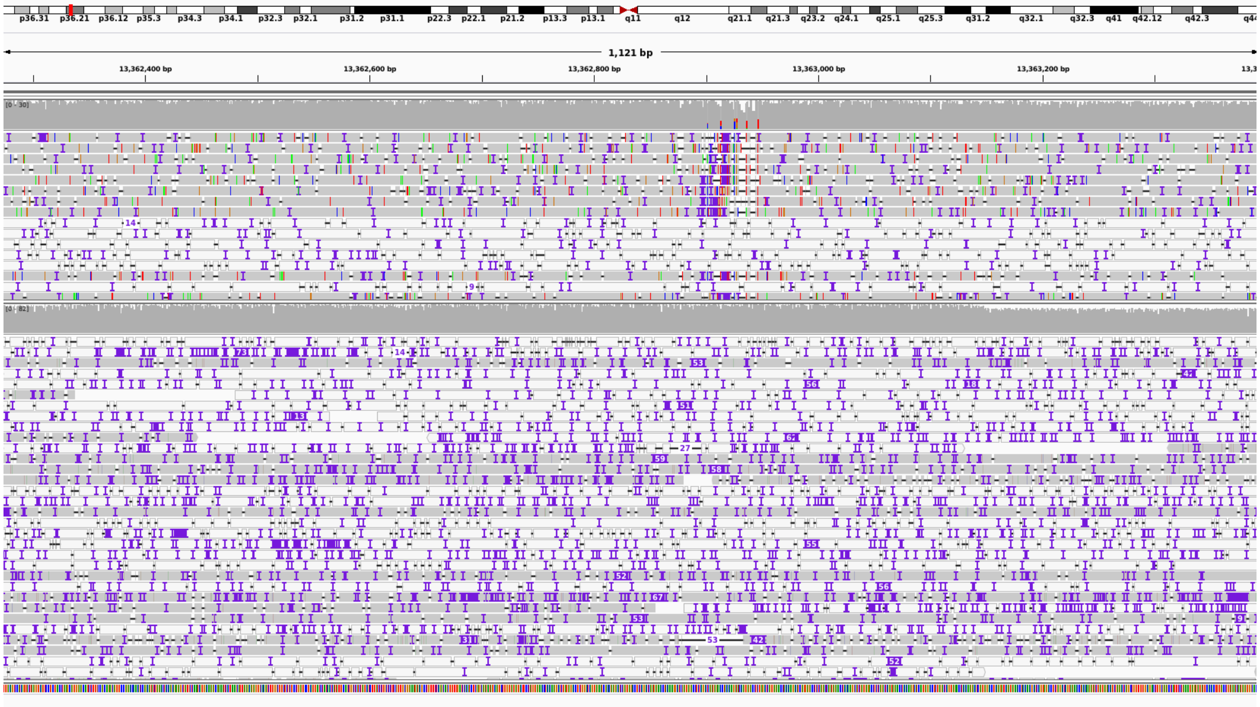

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3
1,121 bp
13,362,400 bp
13,362,600 bp
13,362,800 bp
13,363,000 bp
13,363,200 bp
[0 - 30]
[0 - 82]

Figure S21 | CSUB chromosome 1, position 13362897, length 53 bp

Figure 22: We observed an heterozygous insertion.

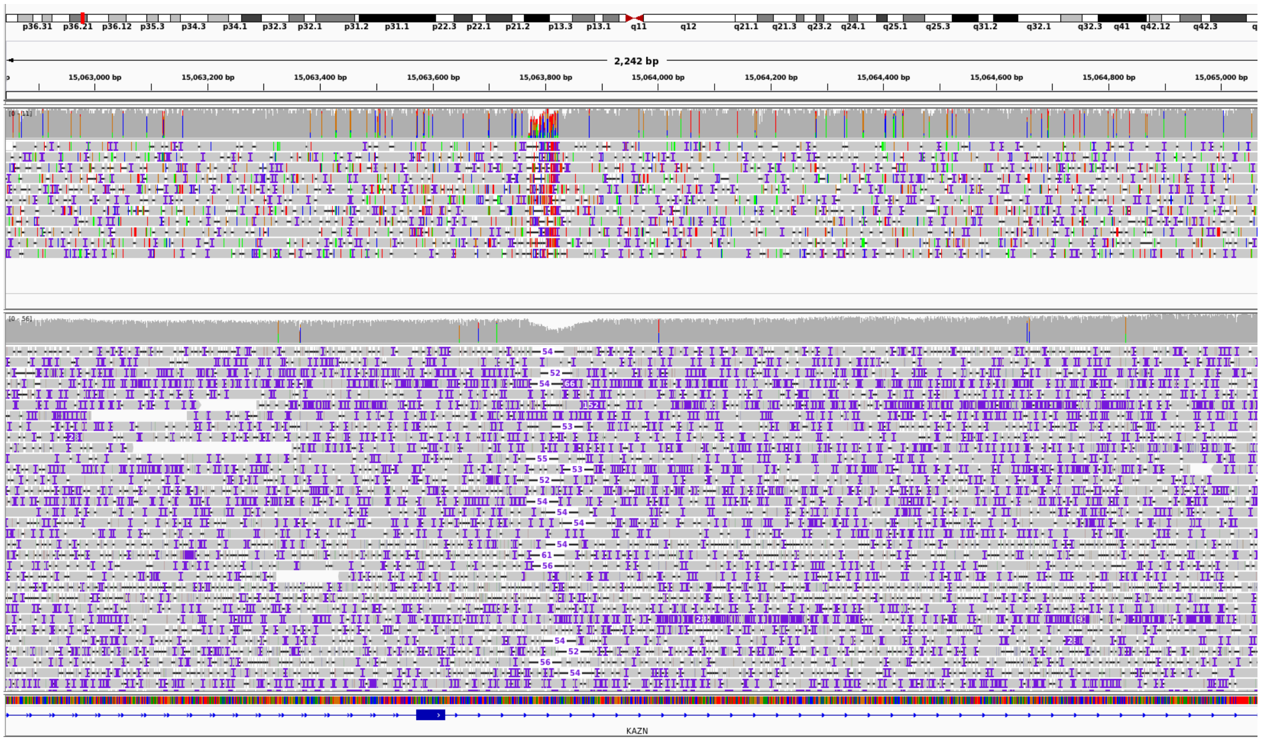

p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3
2,242 bp
15,063,000 bp
15,063,200 bp
15,063,400 bp
15,063,600 bp
15,063,800 bp
15,064,000 bp
15,064,200 bp
15,064,400 bp
15,064,600 bp
15,064,800 bp
15,065,000 bp
[0 - 11]
[0 - 56]
KAZN

Figure S22 | CSUB chromosome 1, position 6935636, length 64 bp

Figure 23: We observed a deletion.

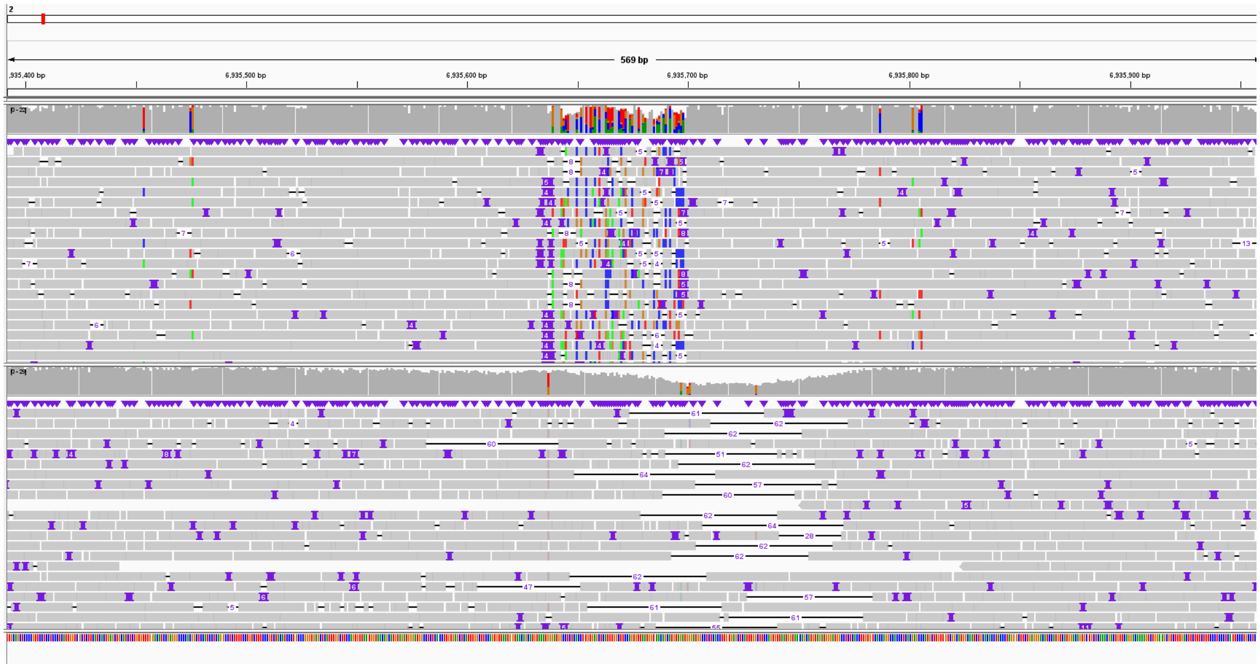


Figure S23 | CSUB chromosome 2, position 6935636, length 64 bp

Figure 24 - 26: We did not observe a SV at any of the positions by inspecting the alignments.

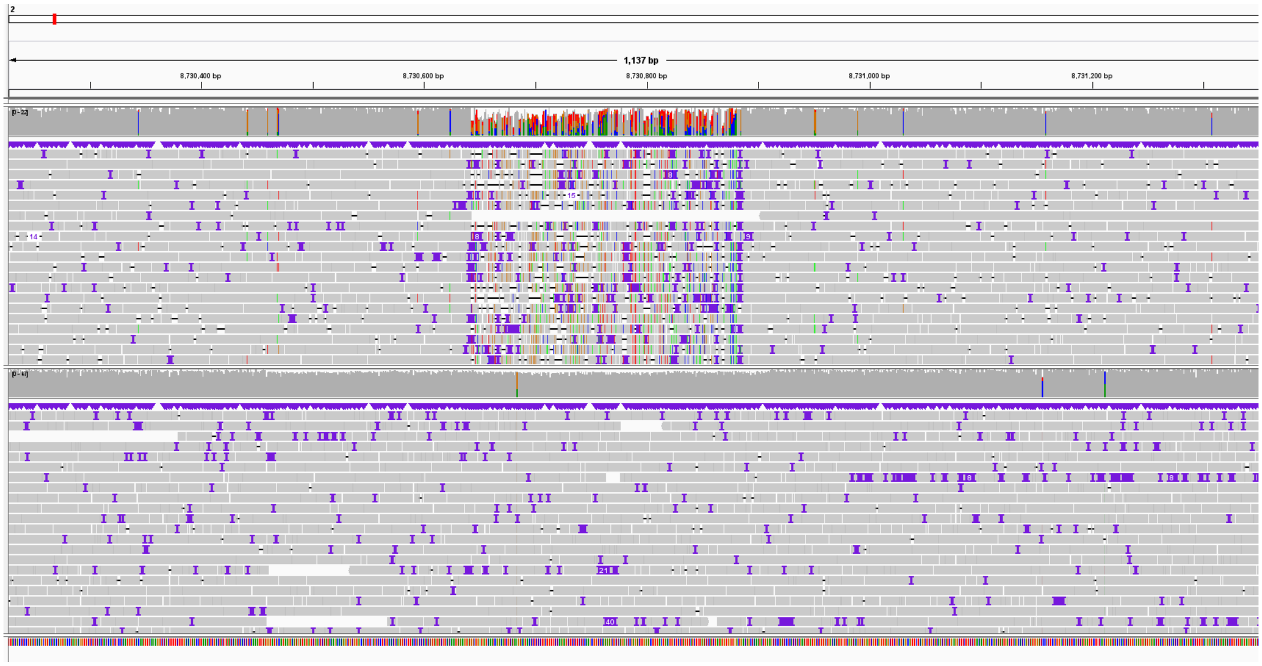


Figure S24 | CSUB chromosome 2, position 8730642, length 242 bp

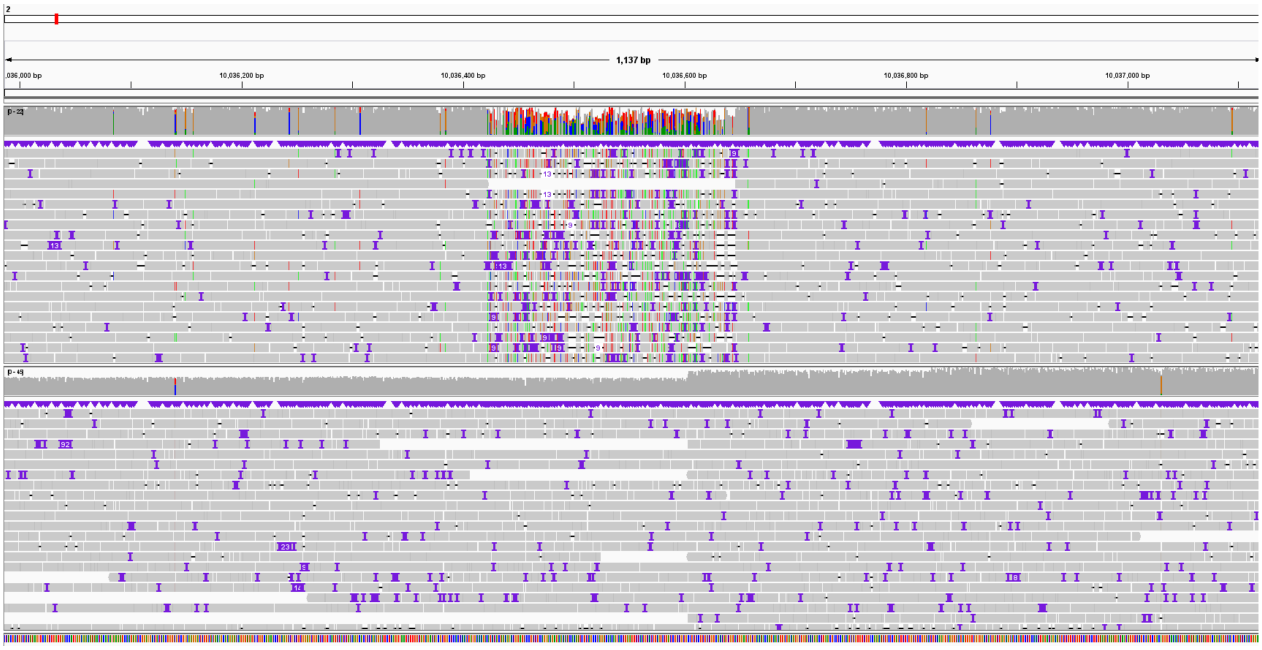


Figure S25 | CSUB chromosome 2, position 10036422, length 225 bp

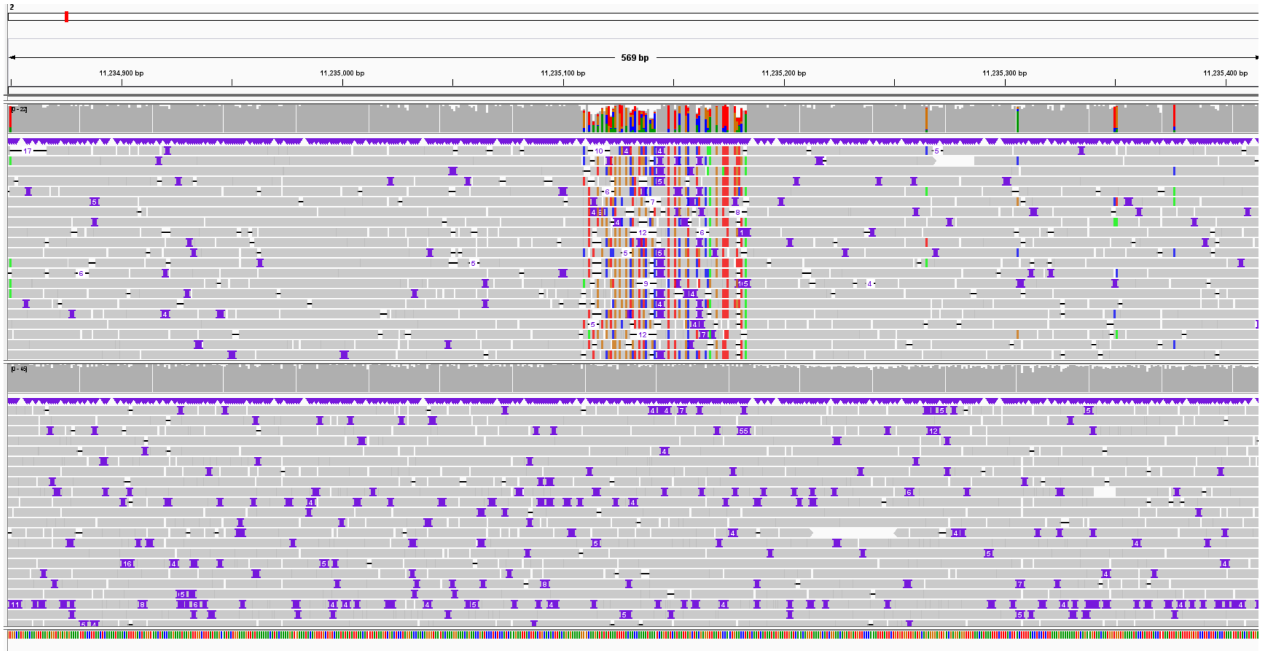


Figure S26 | CSUB chromosome 2, position 11235109, length 74 bp

Figure 27 & 28: This SV is a confirmed insertion, as can be seen in the overall alignment and the individual PacBio alignments.

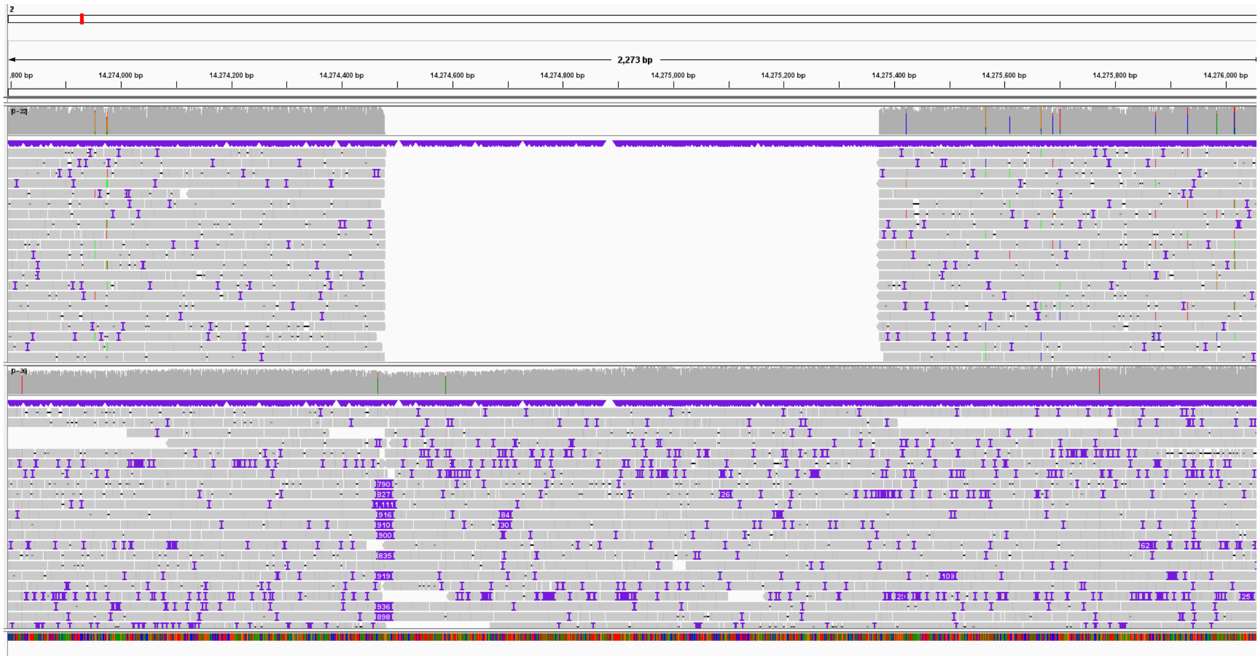
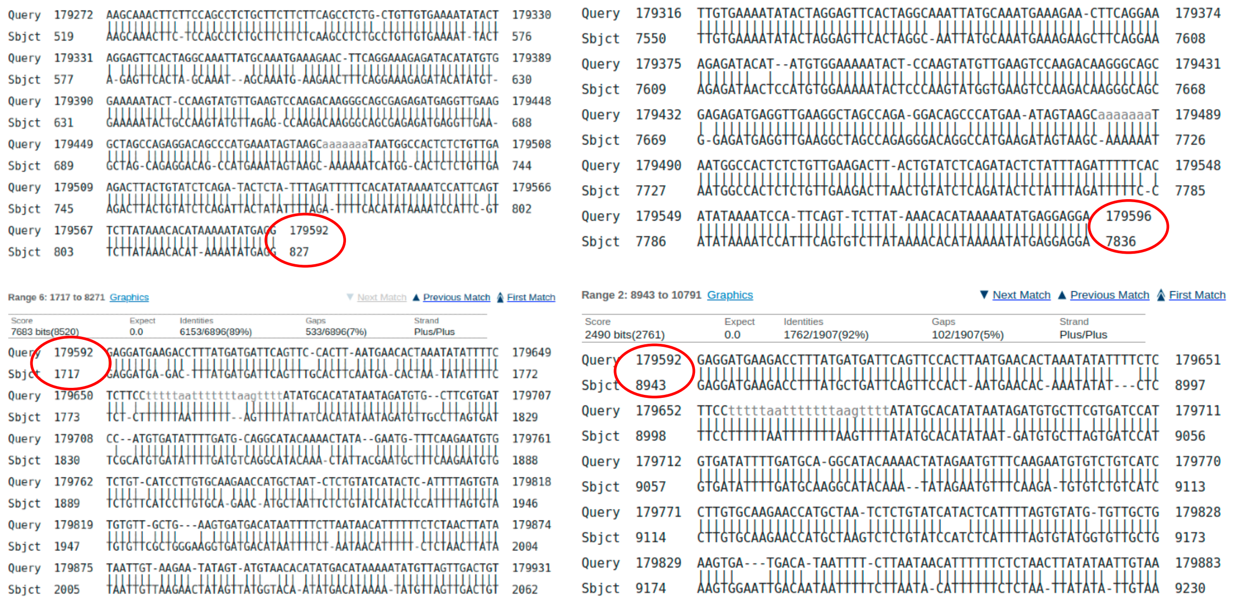


Figure S27 | CSUB chromosome 2, position 14274478, length 895 bp.



```
Query  179272  AAGCAAACTTCTTCCAGCCTCTGCTTCTTCTTCAGCCTCTG-CTGTTGTGAAAATATACT  179330
Sbjct  519     AAGCAAACTTC-TCCAGCCTCTGCTTCTTCTCAAGCCTCTGCCTGTTGTGAAAAT-TACT  576
Query  179331  AGGAGTTCACTAGGCAAATTATGCAAATGAAAGAAC-TTCAGGAAAGAGATACATATGTG  179389
Sbjct  577     A-GAGTTCACTA-GCAAAT--AGCAAATG-AAGAACTTTCAGGAAAGAGATACATATGT-  630
Query  179390  GAAAAATACT-CCAAGTATGTTGAAGTCCAAGACAAGGGCAGCGAGAGATGAGGTTGAAG  179448
Sbjct  631     GAAAAATACTGCCAAGTATGTTAGAG-CCAAGACAAGGGCAGCGAGAGATGAGGTTGAA-  688
Query  179449  GCTAGCCAGAGGACAGCCCATGAAATAGTAAGCaaaaaaaTAATGGCCACTCTCTGTTGA  179508
Sbjct  689     GCTAG-CAGAGGACAG-CCATGAAATAGTAAGC-AAAAAATCATGG-CACTCTCTGTTGA  744
Query  179509  AGACTTACTGTATCTCAGA-TACTCTA-TTTAGATTTTTCACATATAAAATCCATTCAGT  179566
Sbjct  745     AGACTTACTGTATCTCAGATTACTATATTTTAGA-TTTTCACATATAAAATCCATTC-GT  802
Query  179567  TCTTATAAACACATAAAAATATGAGG  179592
Sbjct  803     TCTTATAAACACAT-AAAATATGAGG  827
```

Range 6: 1717 to 8271 Graphics ▼ Next Match ▲ Previous Match ⍋ First Match

Score	Expect	Identities	Gaps	Strand
7683 bits(8520)	0.0	6153/6896(89%)	533/6896(7%)	Plus/Plus

```
Query  179592  GAGGATGAAGACCTTTATGATGATTCAGTTC-CACTT-AATGAACACTAAATATATTTTC  179649
Sbjct  1717    GAGGATGA-GAC-TTTATGATGATTCAGTTTGCACTTCAATGA-CACTAA-TATATTTTC  1772
Query  179650  TCTTCCttttaatttttttaagttttATATGCACATATAATAGATGTG--CTTCGTGAT  179707
Sbjct  1773    TCT-CTTTTTTAATTTTTTT--AGTTTTATTATCACATATAATAGATGTTGCCTTAGTGAT  1829
Query  179708  CC--ATGTGATATTTTGATG-CAGGCATACAAAACTATA--GAATG-TTTCAAGAATGTG  179761
Sbjct  1830    TCGCATGTGATATTTTGATGTCAGGCATACAAA-CTATTACGAATGCTTTCAAGAATGTG  1888
Query  179762  TCTGT-CATCCTTGTGCAAGAACCATGCTAAT-CTCTGTATCATACTC-ATTTTAGTGTA  179818
Sbjct  1889    TCTGTTCATCCTTGTGCA-GAAC-ATGCTAATTCTCTGTATCATACTCCATTTTAGTGTA  1946
Query  179819  TGTGTT-GCTG---AAGTGATGACATAATTTTCTTAATAACATTTTTTCTCTAACTTATA  179874
Sbjct  1947    TGTGTTCGCTGGGAAGGTGATGACATAATTTTCT-AATAACATTTTTT-CTCTAACTTATA  2004
Query  179875  TAATTGT-AAGAA-TATAGT-ATGTAACACATATGACATAAAAATATGTTAGTTGACTGT  179931
Sbjct  2005    TAATTGTTAAGAACTATAGTTATGGTACA-ATATGACATAAAA-TATGTTAGTTGACTGT  2062
```

```
Query  179316  TTGTGAAAATATACTAGGAGTTCACTAGGCAAATTATGCAAATGAAAGAA-CTTCAGGAA  179374
Sbjct  7550    TTGTGAAAATATACTAGGAGTTCACTAGGC-AATTATGCAAATGAAAGAAGCTTCAGGAA  7608
Query  179375  AGAGATACAT--ATGTGGAAAAATACT-CCAAGTATGTTGAAGTCCAAGACAAGGGCAGC  179431
Sbjct  7609    AGAGATAACTCCATGTGGAAAAATACTCCCAAGTATGGTGAAGTCCAAGACAAGGGCAGC  7668
Query  179432  GAGAGATGAGGTTGAAGGCTAGCCAGA-GGACAGCCCATGAA-ATAGTAAGCaaaaaaaT  179489
Sbjct  7669    G-GAGATGAGGTTGAAGGCTAGCCAGAGGGACAGGCCATGAAGATAGTAAGC-AAAAAAT  7726
Query  179490  AATGGCCACTCTCTGTTGAAGACTT-ACTGTATCTCAGATACTCTATTTAGATTTTTCAC  179548
Sbjct  7727    AATGGCCACTCTCTGTTGAAGACTTAACTGTATCTCAGATACTCTATTTAGATTTTTC-C  7785
Query  179549  ATATAAAATCCA-TTCAGT-TCTTAT-AAACACATAAAAATATGAGGAGGA  179596
Sbjct  7786    ATATAAAATCCATTTCAGTGTCTTATAAAACACATAAAAATATGAGGAGGA  7836
```

Range 2: 8943 to 10791 Graphics ▼ Next Match ▲ Previous Match ⍋ First Match

Score	Expect	Identities	Gaps	Strand
2490 bits(2761)	0.0	1762/1907(92%)	102/1907(5%)	Plus/Plus

```
Query  179592  GAGGATGAAGACCTTTATGATGATTCAGTTCCACTTAATGAACACTAAATATATTTTCTC  179651
Sbjct  8943    GAGGATGAAGACCTTTATGCTGATTCAGTTCCACT-AATGAACAC-AAATATAT---CTC  8997
Query  179652  TTCCttttaatttttttaagttttATATGCACATATAATAGATGTGCTTCGTGATCCAT  179711
Sbjct  8998    TTCCTTTTTAATTTTTTTAAGTTTTATATGCACATATAAT-GATGTGCTTAGTGATCCAT  9056
Query  179712  GTGATATTTTGATGCA-GGCATACAAAACTATAGAATGTTTCAAGAATGTGTCTGTCATC  179770
Sbjct  9057    GTGATATTTTGATGCAAGGCATACAAA--TATAGAATGTTTCAAGA-TGTGTCTGTCATC  9113
Query  179771  CTTGTGCAAGAACCATGCTAA-TCTCTGTATCATACTCATTTTAGTGTATG-TGTTGCTG  179828
Sbjct  9114    CTTGTGCAAGAACCATGCTAAGTCTCTGTATCCATCTCATTTTAGTGTATGGTGTTGCTG  9173
Query  179829  AAGTGA---TGACA-TAATTTT-CTTAATAACATTTTTTCTCTAACTTATATAATTGTAA  179883
Sbjct  9174    AAGTGGAATTGACAATAATTTTTCTTAATA-CATTTTTTCTCTAA-TTATATA-TTGTAA  9230
```

Figure S28 | Individual alignment of two PacBio reads for ‘CSUB chromosome 2, position 14274478, length 895 bp’.

Figure 29 : We did not observe a SV at this position by inspecting the alignments.

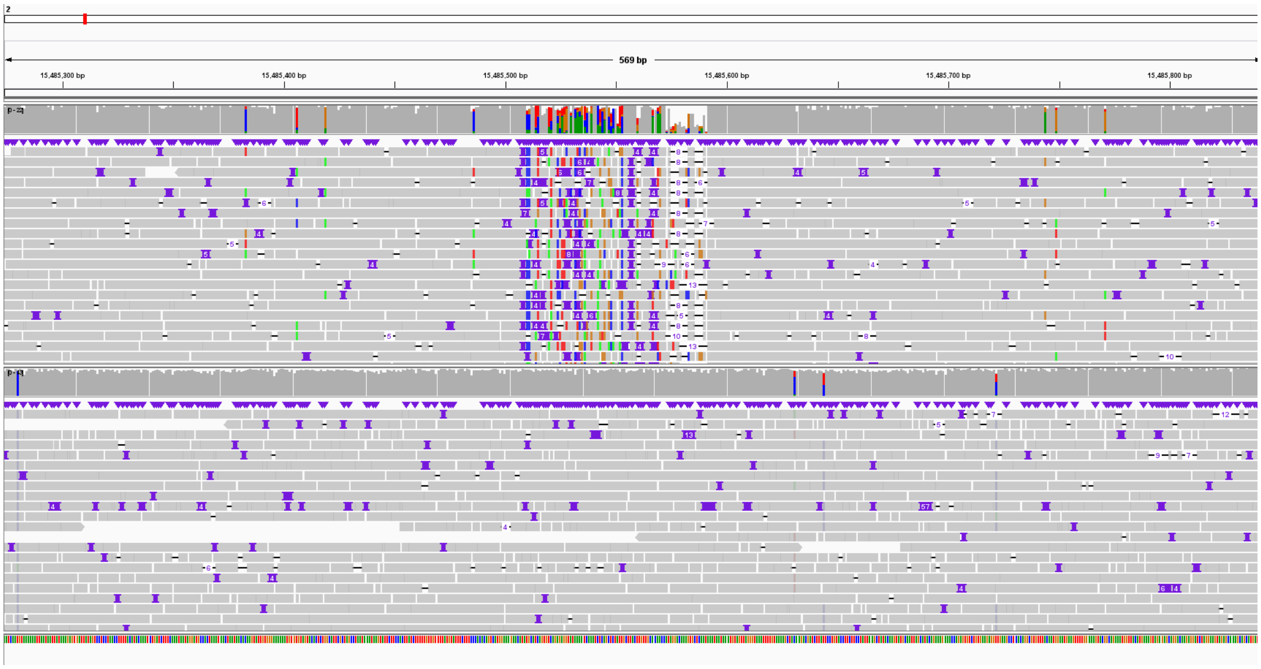


Figure S29 | CSUB chromosome 2, position 15485508, length 895 bp.

Figure 30 & 31: Insertions were observed across a tandem repeat region.

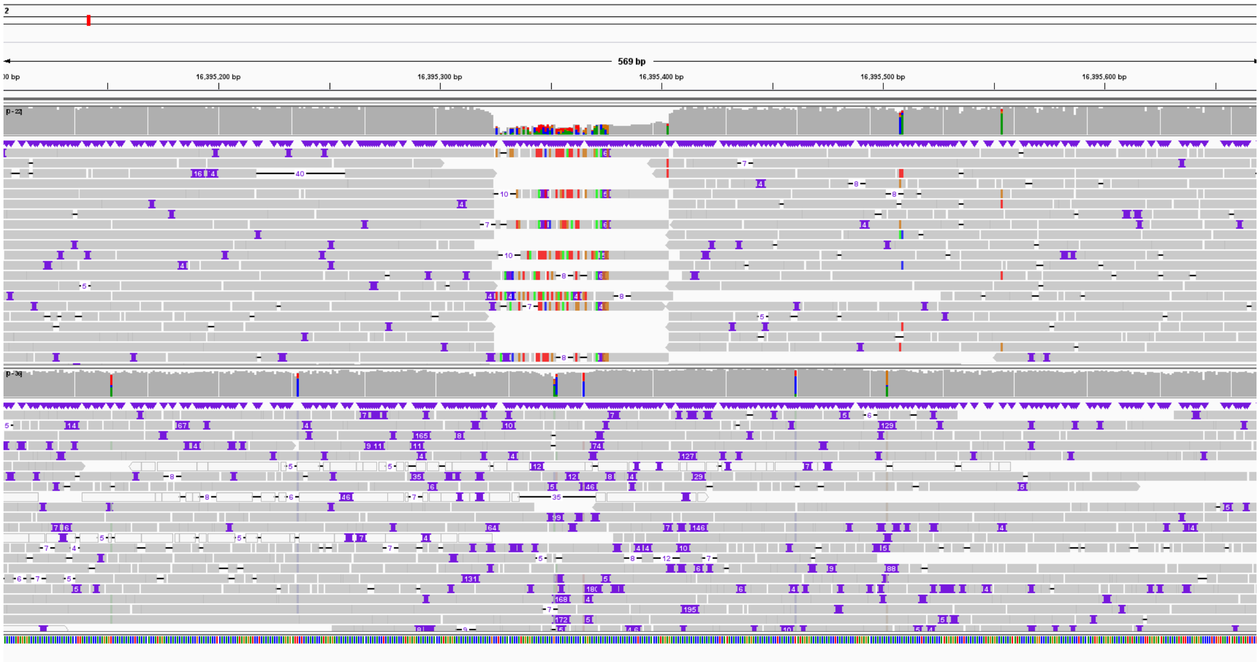


Figure S30 | CSUB chromosome 2, position 16395324, length 54 bp.

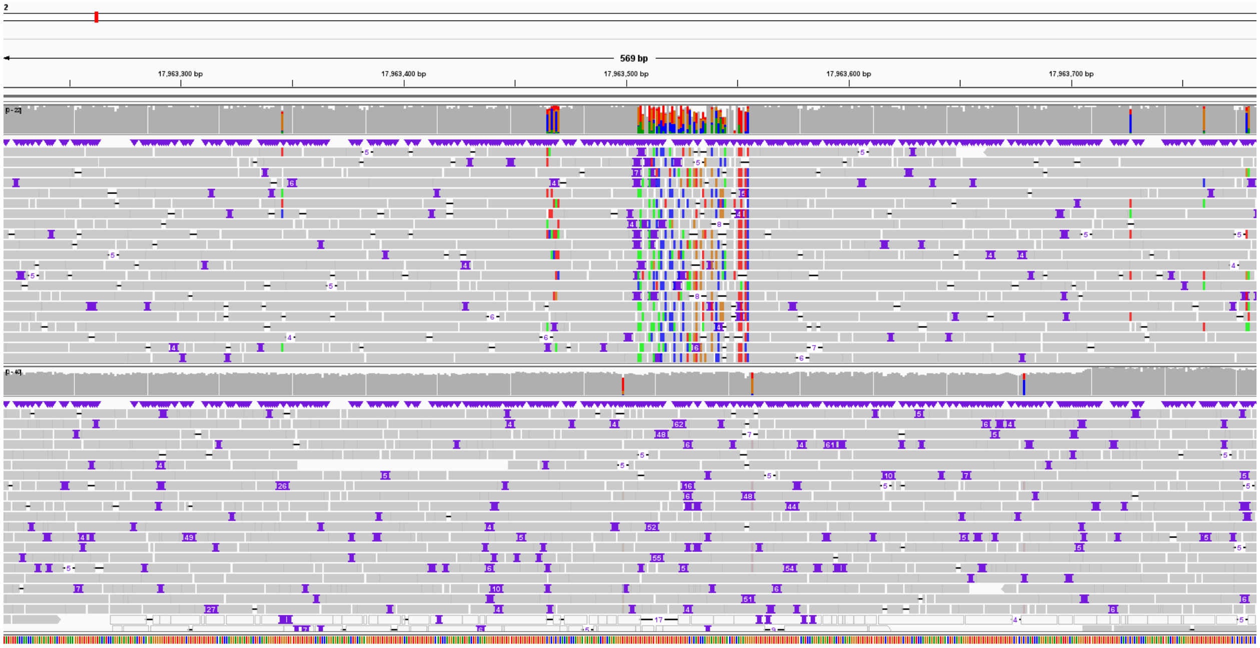


Figure S31 | CSUB chromosome 2, position 17963505, length 50 bp.

Figure 32: Insertions were observed across a repetitive region.

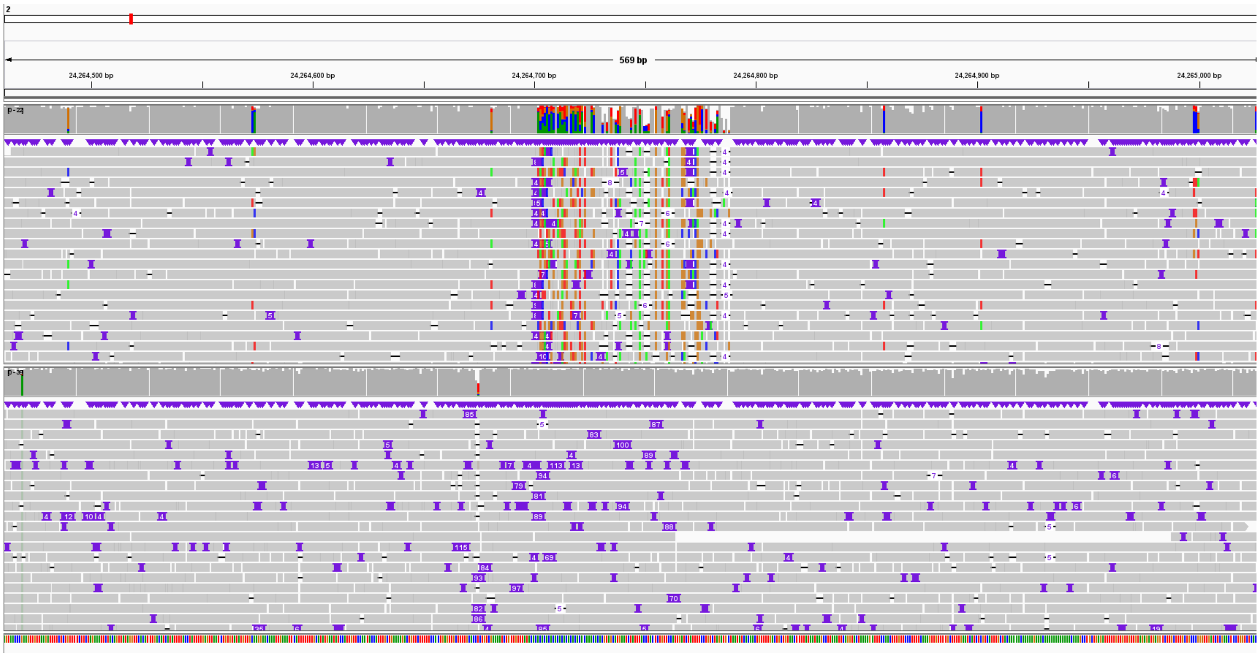


Figure S32 | CSUB chromosome 2, position 24264701, length 87 bp.

Figure 33 & 34 : We did not observe a SV at either position by inspecting the alignments.

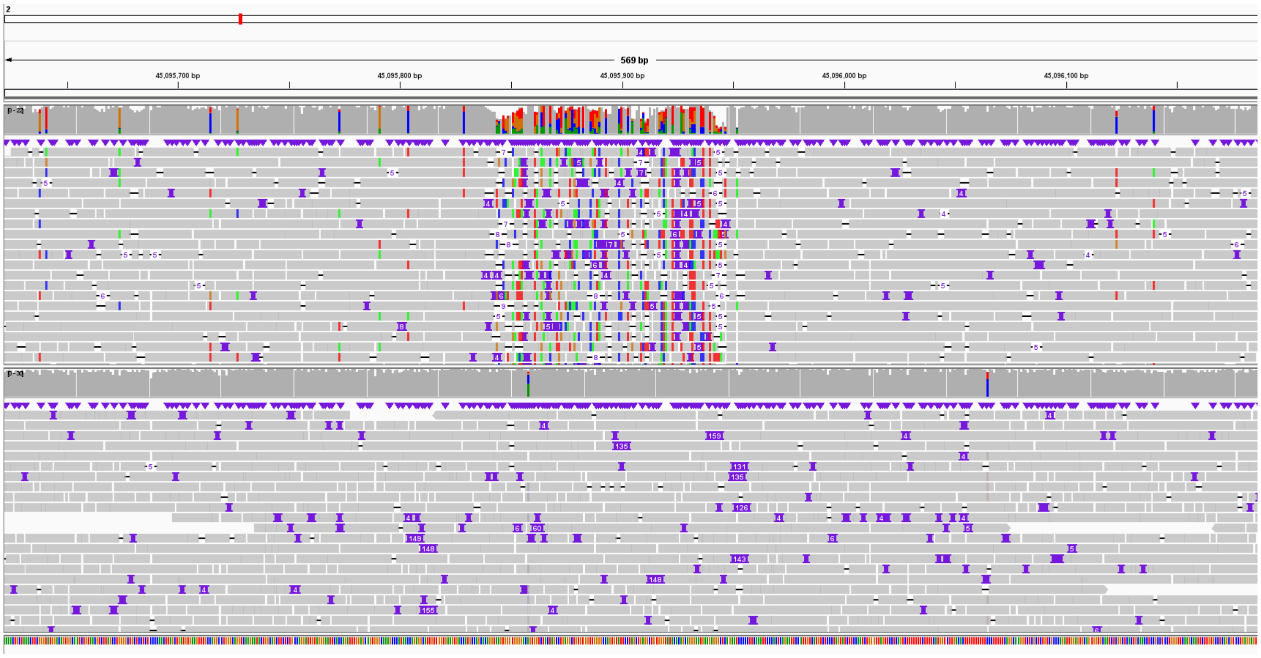


Figure S33 | CSUB chromosome 2, position 45095843, length 110 bp.

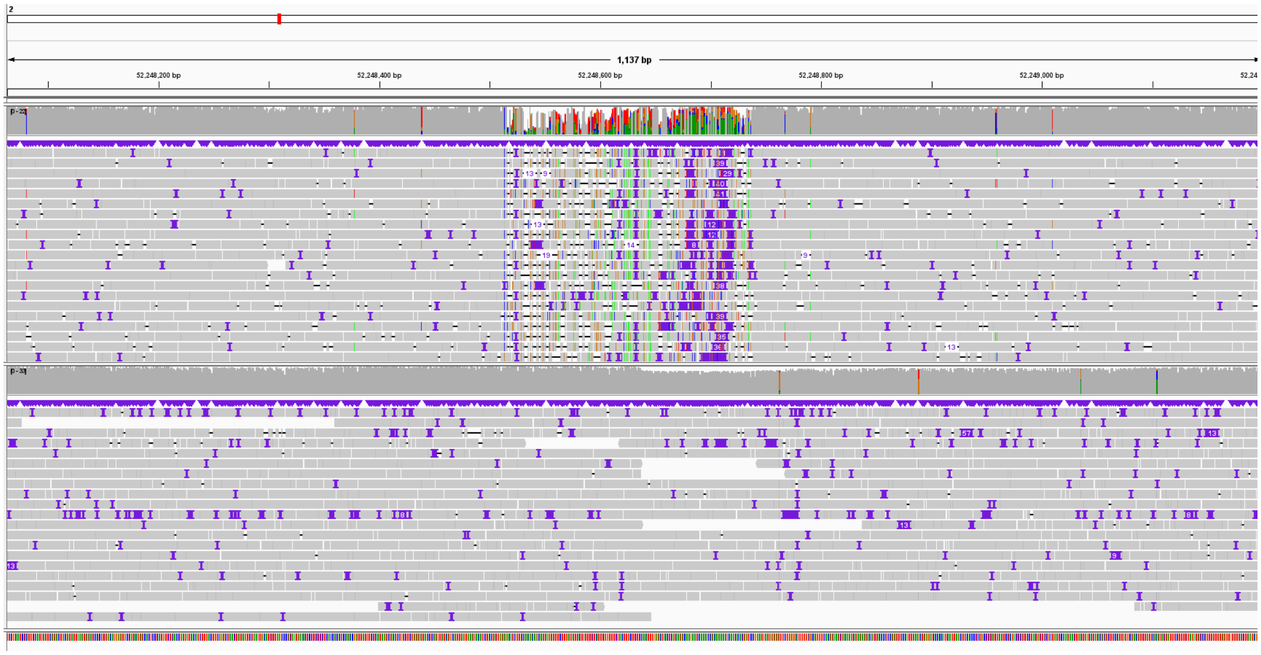


Figure S34 | CSUB chromosome 2, position 52248513, length 225 bp.

Figure 35 & 36: Insertions were observed across a repetitive region.

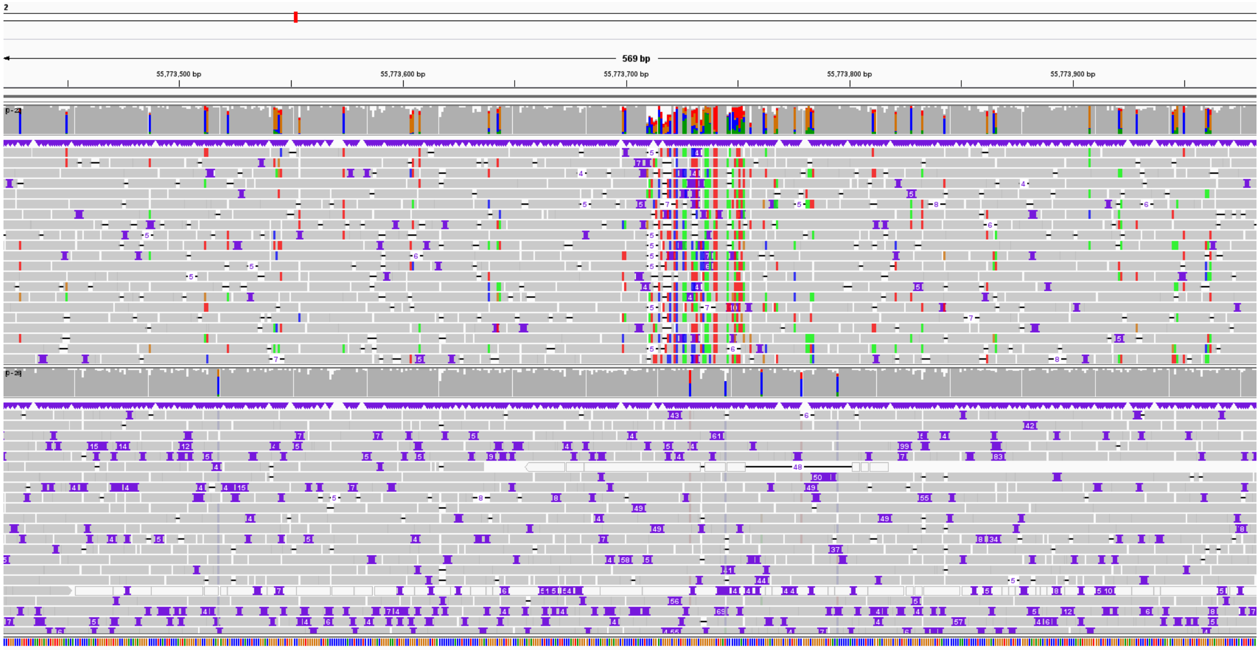


Figure S35 | CSUB chromosome 2, position 55773706, length 50 bp.

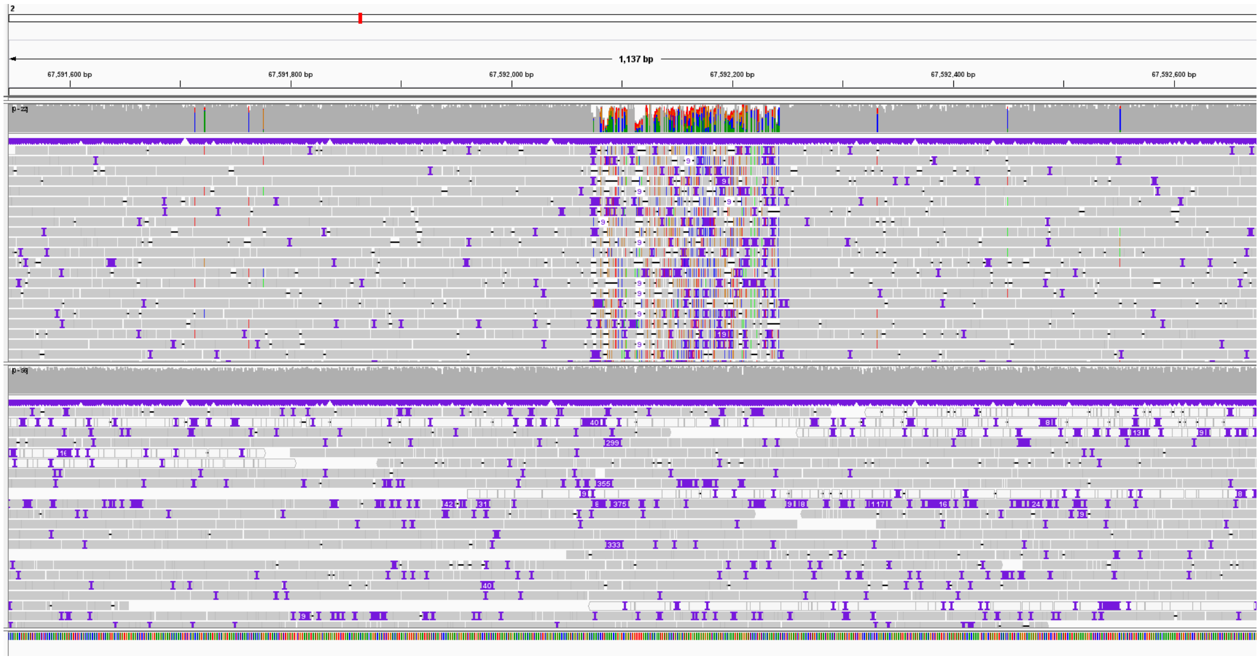


Figure S36 | CSUB chromosome 2, position 67592074, length 169 bp.

Figure 37: We observed an heterozygous deletion.

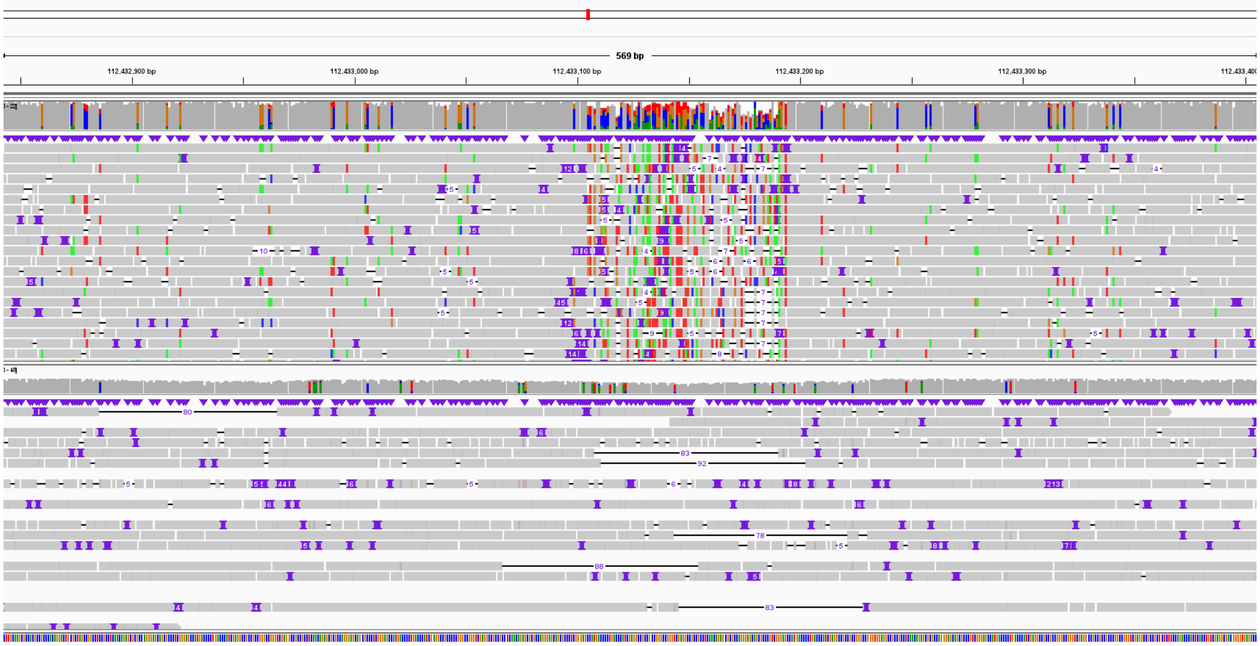


Figure S37 | CSUB chromosome 2, position 112433104, length 90 bp.

Figure 38: Insertions were observed across a tandem repeat region.

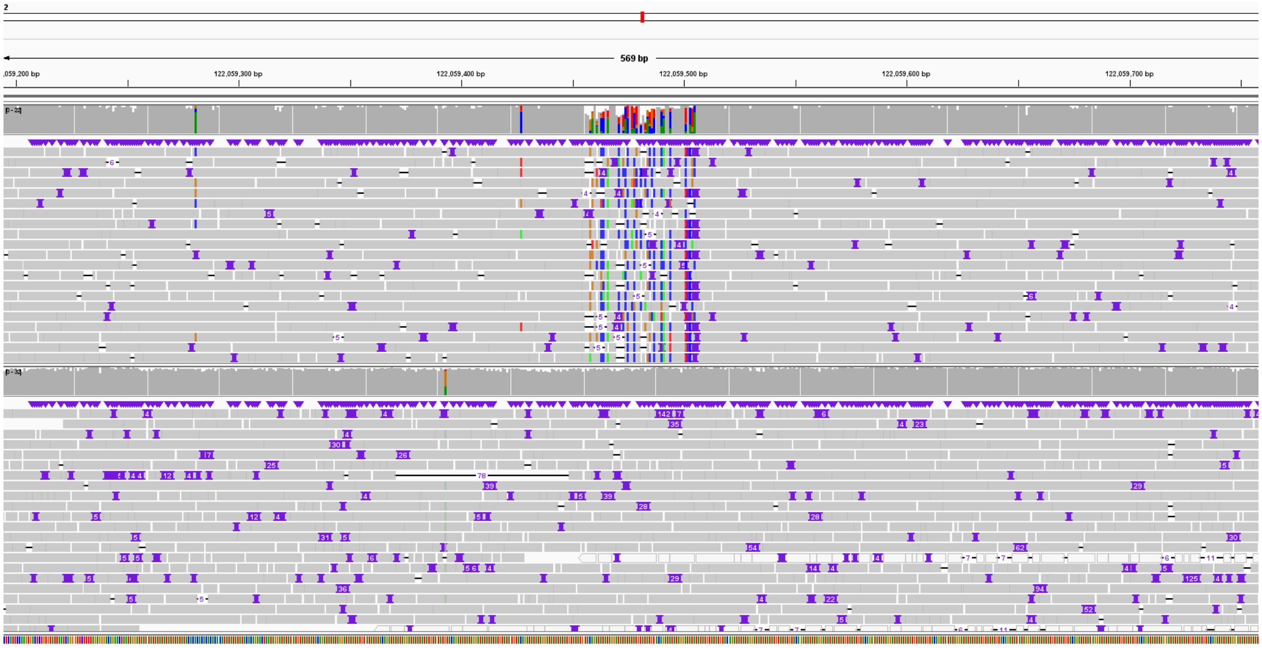


Figure S38 | CSUB chromosome 2, position 122059455, length 50 bp.

References

1 . Chaisson, M.J.P., et al. (2019). Multi-platform discovery of haplotype resolved structural variation in human genomes. *Nature Communications*, **10**, 1784. doi.org/10.1038/s41467-018-08148-z