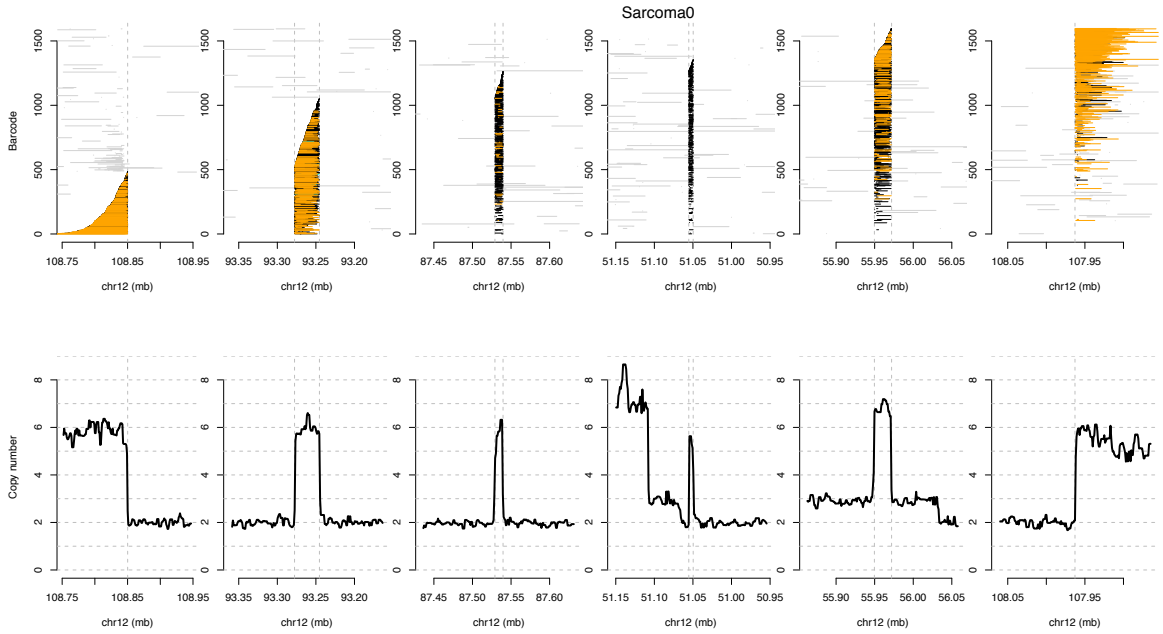
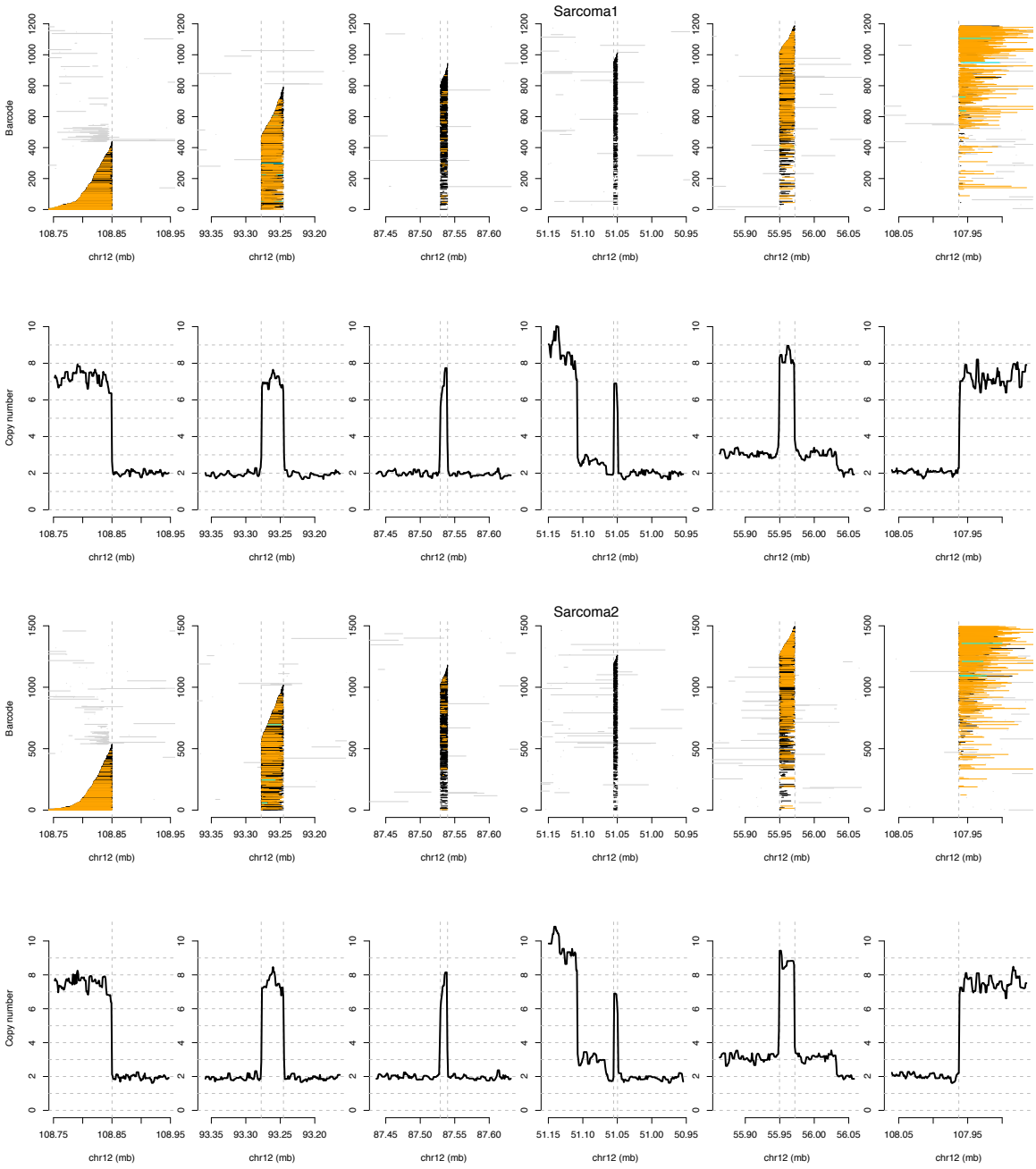
Supplementary Note 1 – Interpretation of complex structural variant examples

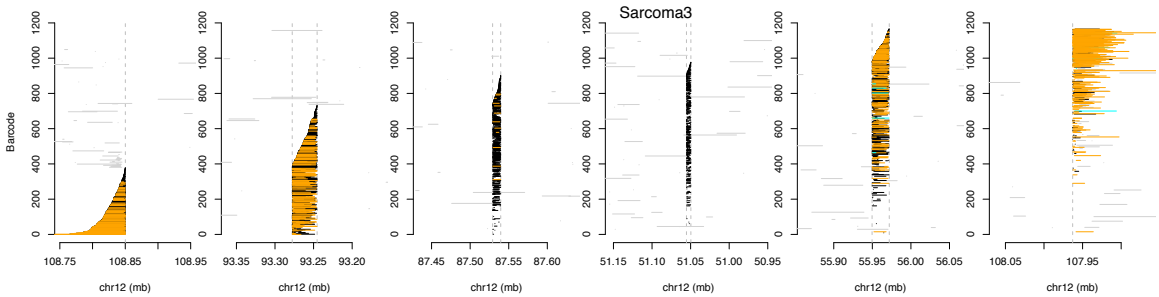
Shown below is the example event from Figure 2a for all 7 sarcoma samples. As in Figure 2, orange lines represent read clouds supporting the SV from the major haplotype, cyan lines represent supporting read clouds from the minor haplotype (there are very few of these) and black lines represent supporting read clouds that were not assigned to a haplotype (these are more frequent for the shorter read clouds as they have a lower chance of overlapping an informative single nucleotide polymorphism). Gray lines represent non-supporting read clouds, which result either from independent fragments in the same barcode (so-called “barcode collisions”) or from truly spanning fragments whose closest read is too distant from the breakpoint to be considered as supporting read clouds.

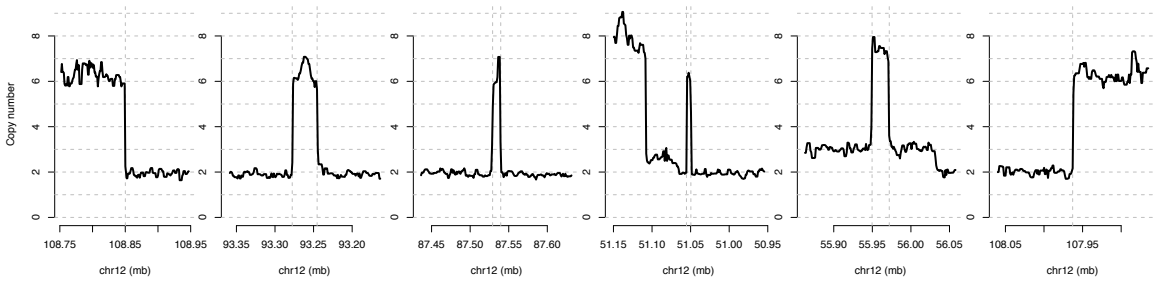
Note that the average fragment lengths are longer in some samples, for example Sarcoma 10, resulting in more fragments that span the entire event from start to end.

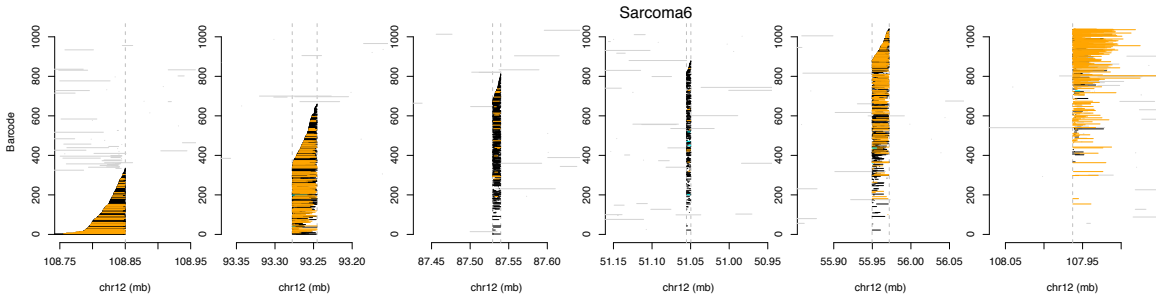
The copy number profiles (bottom of each pair) are normalized to the matched control sample, so 2 represents the normal, diploid state. This SV appears to be at copy number 4, since the normal chromosomal DNA is at copy number 2, and the regions included in the SV are at copy number 6. The exception to this pattern is the 5th panel from the left, the 56 mb locus, which has background DNA copy number 3 and the SV region appears to be at copy number 7, again supporting copy number of 7–3=4 for the SV.

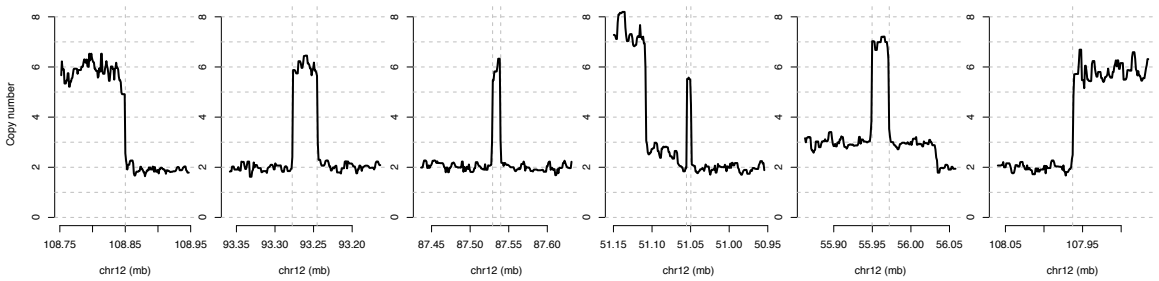


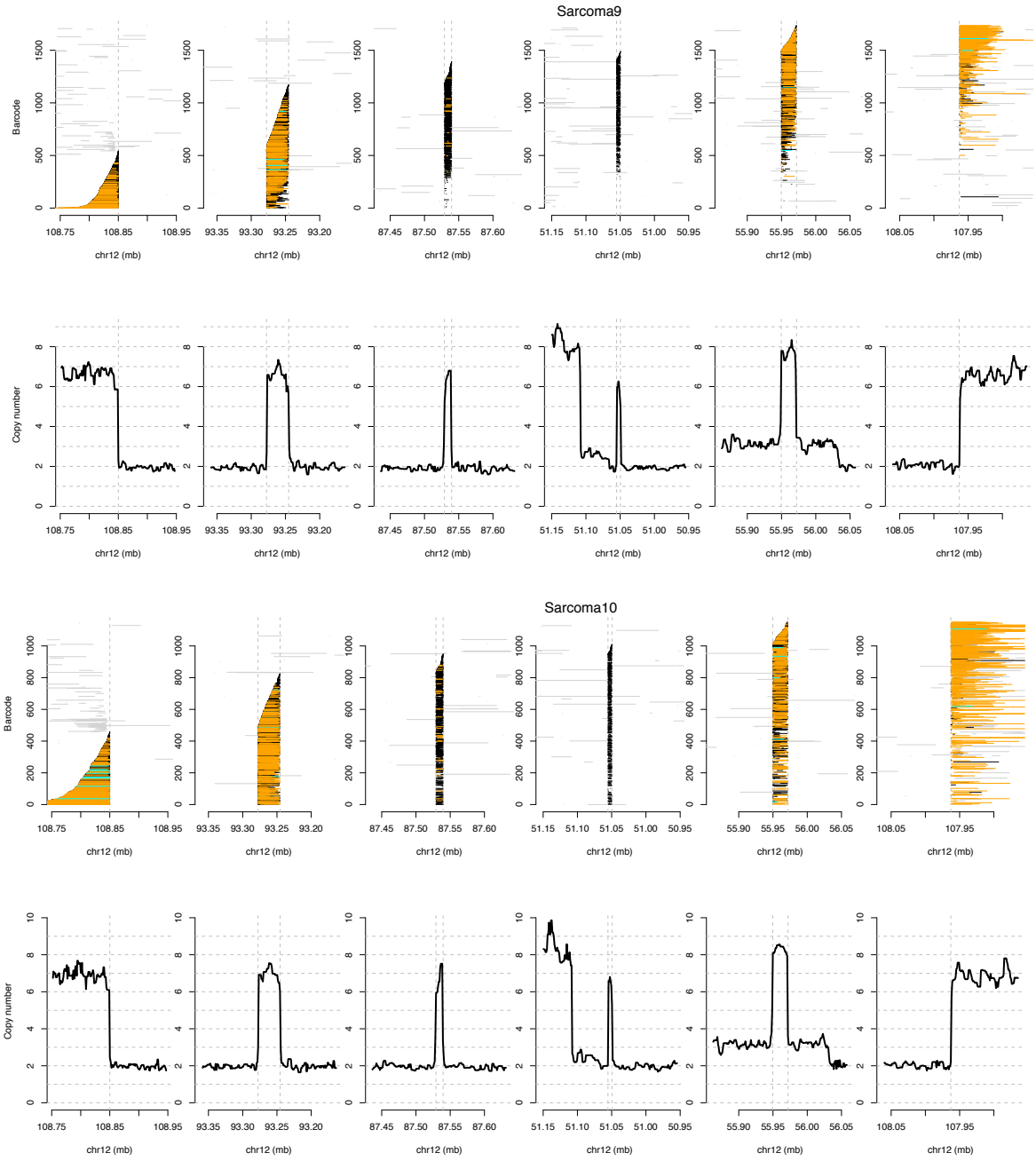

Sarcoma1
Barcode
Copy number
chr12 (mb)
Sarcoma2
Barcode
Copy number
chr12 (mb)


Sarcoma3
Barcode
chr12 (mb)


Copy number
chr12 (mb)


Sarcoma6
Barcode
chr12 (mb)


Copy number
chr12 (mb)


Sarcoma9
Barcode
Copy number
chr12 (mb)
Sarcoma10
Barcode
Copy number
chr12 (mb)

Supplementary Note 2 – Detection of smaller structural variants

As described in the main text, the barcode patterns used by GROC-SVs to detect SVs provide a high signal to background ratio for detecting large-scale structural variation, such as frequently found in tumors. However, barcode patterns are similar between genomic loci that are close to one another and thus the SV detection algorithm used by GROC-SVs is expected to exhibit lower sensitivity and specificity for SVs that are smaller than the average long-fragment length. Practically, GROC-SVs is unable to find SVs smaller than 10kb in data from standard 10x libraries.

To assess the ability of GROC-SVs to detect smaller SVs, we applied GROC-SVs to the NA12878 genome, for which comprehensive SV callsets and sequencing data are available. GROC-SVs identified 36 high-confidence candidate SVs in NA12878, of which 17 (47%) were corroborated using mate-pair data (Eberle 2017). When considering a more inclusive set of SVs, including additional lower confidence SVs, 56/227 (25%) of GROC-SVs calls were corroborated by mate-pair data.

Out of this inclusive set of SVs called by GROC-SVs, 43 perfectly or partially overlapped SVs described in Pendleton et al (2015), which were called using Pacific Biosciences long reads. For comparison, when considering only the Pendleton deletions of size ≥10kb (this was the only SV type identified in that size range), we found that 60/228 (26%) were corroborated by mate-pair data. We note that we used mate-pair data for validation because that sequencing platform is orthogonal to both the PacBio-based calls made by Pendleton and the 10x-based calls made by GROC-SVs.

References

Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang HY, Humphray SJ, Halpern AL, Kruglyak S, Margulies EH, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 2017 Jan;27(1):157-164. doi: 10.1101/gr.210500.116.

Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015 Aug;12(8):780-6. doi: 10.1038/nmeth.3454

		Sarcoma0	Sarcoma1	Sarcoma2	Sarcoma3	Sarcoma6	Sarcoma9	Sarcoma10	SarcomaC	HCC1143	HCC1143_BL
Fragment Lengths (bp)	10%	8,625	3,956	4,974	1,964	4,783	6,754	5,122	2,601	1,498	1,271
	25%	25,824	17,573	20,327	9,025	14,884	22,835	21,980	11,043	5,304	3,909
	50%	43,070	44,840	44,476	26,709	32,406	38,706	51,708	28,742	21,369	14,843
	75%	56,244	62,231	61,171	47,739	51,483	48,691	70,101	43,055	57,011	43,208
	90%	70,559	79,806	75,906	59,464	62,901	56,907	100,961	52,677	106,764	85,675
	95%	81,163	95,388	87,497	66,166	70,829	62,395	124,431	58,197	147,857	122,273
	mean	42,366	43,730	43,183	29,590	34,258	35,903	52,443	28,392	41,331	32,608
	std	24,497	30,186	27,888	22,465	23,227	18,887	37,903	19,285	53,562	45,907
Physical depth (C_F)		350.6 ± 189.1	195.0 ± 107.9	236.0 ± 142.8	189.5 ± 95.9	203.1 ± 90.2	272.9 ± 143.5	215.3 ± 125.8	209.0 ± 34.0	144.3 ± 68.5	126.2 ± 24.6
Read coverage per fragment position (C_R)		0.07 ± 0.10	0.13 ± 0.15	0.11 ± 0.13	0.14 ± 0.16	0.11 ± 0.12	0.08 ± 0.10	0.10 ± 0.10	0.12 ± 0.15	0.34 ± 0.20	0.27 ± 0.21
Barcodes (after filtering)		166,892	157,257	168,818	164,865	172,893	183,770	167,595	176,451	700,176	651,125

Supplementary Table 1: 10x sequencing library statistics.

Fragment lengths (top rows, in bp) for size-selected sarcoma libraries and HCC1143 breast cancer cell line. Physical depth C_F indicates the average fold-coverage of each genomic position by inferred long fragments. Read coverage C_R indicates the average coverage of each long fragment by short reads. Note that since C_R is less than 1.0, many positions in each long fragment are not covered by short-reads. Barcodes indicates the number of different molecular partitions analyzed by GROC-SVs (note that some low-quality molecular partitions are filtered as described in the methods).