

Genome-wide reconstruction of complex structural variants using read clouds

Noah Spies¹⁻³ , Ziming Weng³, Alex Bishara⁴, Jennifer McDaniel¹, David Catoe¹, Justin M Zook¹, Marc Salit^{1,2}, Robert B West³, Serafim Batzoglou⁴ & Arend Sidow^{2,3,5} 

In read cloud approaches, microfluidic partitioning of long genomic DNA fragments and barcoding of shorter fragments derived from these fragments retains long-range information in short sequencing reads. This combination of short reads with long-range information represents a powerful alternative to single-molecule long-read sequencing. We develop Genome-wide Reconstruction of Complex Structural Variants (GROC-SVs) for SV detection and assembly from read cloud data and apply this method to Illumina-sequenced 10x Genomics sarcoma and breast cancer data sets. Compared with short-fragment sequencing, GROC-SVs substantially improves the specificity of breakpoint detection at comparable sensitivity. This approach also performs sequence assembly across multiple breakpoints simultaneously, enabling the reconstruction of events exhibiting remarkable complexity. We show that chromothriptic rearrangements occurred before copy number amplifications, and that rates of single-nucleotide variants and SVs are not correlated. Our results support the use of read cloud approaches to advance the characterization of large and complex structural variation.

Despite their importance in evolution and disease, structural variants (SVs)—large-scale DNA deletions, duplications, inversions and translocations that contribute the majority of nucleotide variation between individuals¹ and many cancer genomes^{2,3}—remain difficult to characterize. DNA breakage and fusion can connect any two genomic loci, and thus the number of possible SVs is practically limitless. Repetitive loci, uneven or biased sequencing coverage, and the typically short length of sequenced fragments complicate accurate detection. In addition, SVs can be very complex⁴⁻⁸. For example, in chromothripsis³, a chromosome shatters into many pieces that are then apparently randomly reassembled, leading to massive rearrangements.

Complex events are particularly difficult to interpret. Short-fragment sequence analysis can only confidently relate breakpoints that are within the fragment size distribution, which is typically <500 bp. Longer distance reconstruction (e.g., ref. 2) is based on the assumption that downstream events occur in the

same haplotype, and that all breakpoints have been accurately identified. Single-molecule long-read approaches are better suited for SV detection, but they tend to be limited by throughput, cost and high per-base error rates.

Read clouds marry the high throughput and accuracy of standard Illumina sequencing with long-fragment information by incorporating a barcode tag during a molecular partitioning step⁹⁻¹². The 10x Genomics platform produces read cloud libraries with dramatically higher numbers of partitions than those of previous methods, and thus this platform enables new applications¹³. In 10x Genomics library preparation, long DNA fragments are diluted into 10⁵ to 10⁶ microfluidic droplets, each containing a unique barcode; and randomly primed amplification in each droplet produces identically tagged short fragments from the handful of long fragments. Barcoded short fragments are Illumina sequenced, and their alignments to the reference genome form clusters. We refer to the clusters of identically barcoded linked reads as clouds. Each cloud allows us to infer the extent of an originating large DNA molecule, even though the sparse sampling by short reads means we only directly observe a subset of positions within each long fragment.

We have developed GROC-SVs to identify, sequence assemble and reconstruct complex SVs from read cloud data (<https://github.com/grocsvs/grocsvs> and **Supplementary Software**). We show that 10x Genomics data substantially improves the detection of SVs compared with standard short-fragment sequencing and that it enables the reconstruction of large-scale complex SVs. In addition, we use read cloud information to produce high-quality assemblies of sequences spanning breakpoints, and this allows us to better interpret local complexity. We applied GROC-SVs to characterize chromothripsis and subsequent evolution of structural variation in a liposarcoma and to analyze SVs in a breast cancer cell line.

RESULTS

Sequence data generation and characteristics

We Illumina sequenced 10x GemCode libraries from each of seven spatially distinct sites within a well-differentiated liposarcoma,

¹Genome-scale Measurements Group, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. ²Joint Initiative for Metrology in Biology, Stanford, California, USA. ³Department of Pathology, Stanford University School of Medicine, Stanford, California, USA. ⁴Department of Computer Science, Stanford University, Stanford, California, USA. ⁵Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to A.S. (arend@stanford.edu) or N.S. (nspies@stanford.edu).

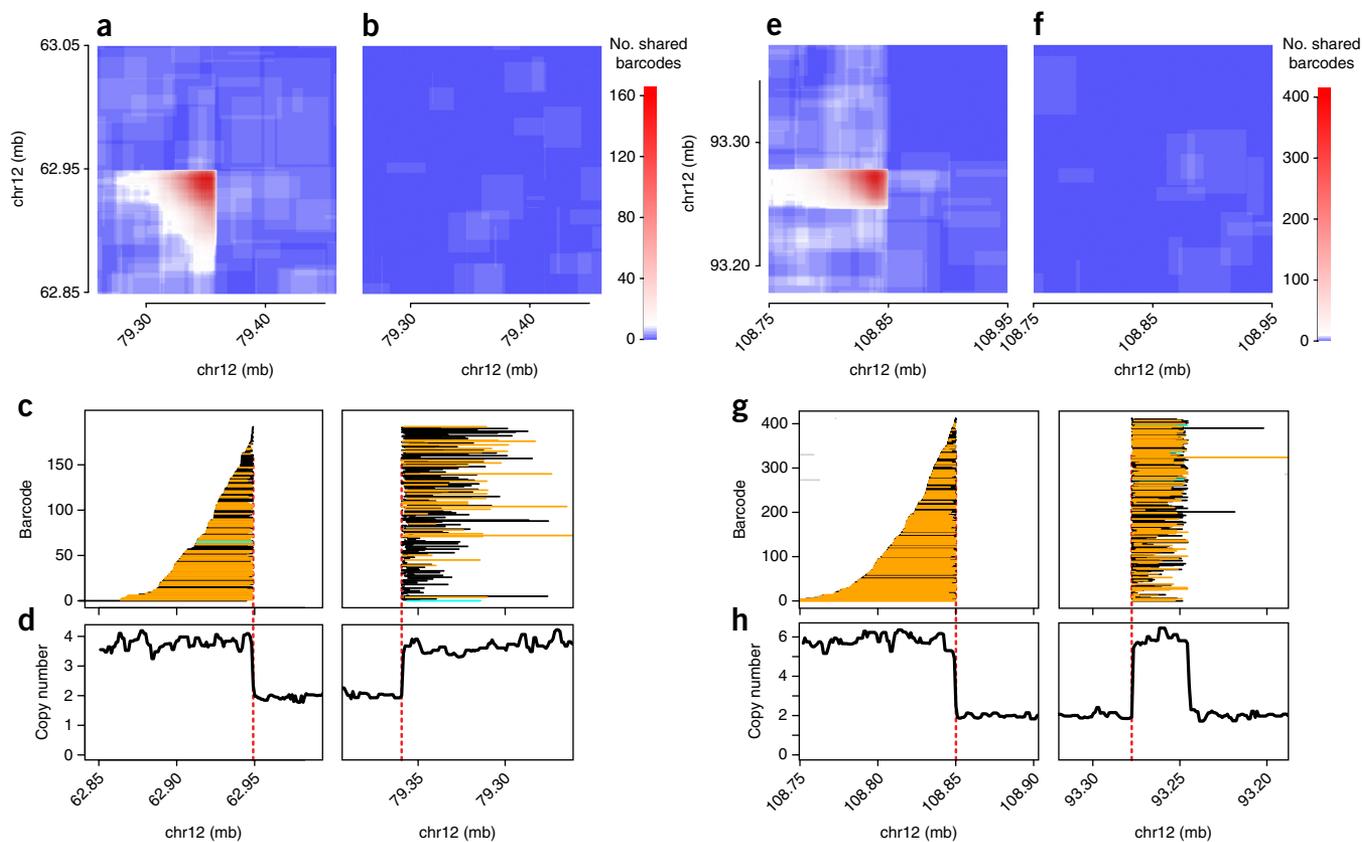


Figure 1 | Examples of breakpoint signals in 10x GemCode data. (**a–d**) A simple breakpoint in sarcoma sample 0. (**e–h**) Two breakpoints in close proximity in sarcoma sample 0. (**a,e**) Heat maps indicate barcode similarity for each pair of genomic locations. In **a**, the greatest signal forms a corner at the breakpoint coordinates. In **e**, two breakpoints can be discerned by a dropoff in signal at 108.85 mb (*x*-axis) in addition to 93.27 Mb and 93.25 Mb (*y*-axis). (**b,f**) The same region from **a** and **e** in the control sample. (**c,g**) Inferred extent of breakpoint-supporting read clouds (corresponding to input fragments). Each row is one cloud, colored according to its assignment to a haplotype; supporting haplotype, orange; unassigned, black; nonsupporting haplotype, cyan; and nonsupporting cloud in the same barcode as a supporting cloud, gray (derived from independent fragments in the same molecular partition). The long fragments tile across the breakpoint when ordered by their leftmost position in the left panel. (**d,h**) Copy number profiles based on the short-fragment data in the sarcoma. Decreasing coordinates depict the minus strand.

as well as a control sample from the same patient. For comparison and validation, we also Illumina sequenced PCR-free and long-insert (~7 kb) mate-pair libraries.

Size selection of sarcoma DNA resulted in a tight 10x-fragment-size distribution (mean > 30 kb; 95th percentile = ~80 kb; **Supplementary Fig. 1** and **Supplementary Table 1**). Each genomic position was covered, on average, by ~250 long fragments, but this was accomplished at an overall sequence coverage of ~25 \times through sparse sampling of each long fragment by short reads. We also analyzed second-generation 10x Chromium data from HCC1143 breast cancer and matched normal cell lines. Prepared without size selection, HCC1143 fragment sizes covered a wide distribution (mean = 41 kb, 95th percentile = 148 kb).

Overview of GROC-SVs

GROC-SVs first looks for statistical evidence for long fragments that span breakpoints by quantifying barcode similarity between all pairs of genomic locations (see **Supplementary Fig. 2a** for an overview). Barcode similarities are highest between any two nearby loci, since input long fragments tend to overlap both loci, whereas loci separated by distances larger than the input fragment

size share few or no barcodes. This is because each barcoded partition only contains a few randomly drawn genomic fragments, and thus the chance that multiple partitions contain long fragments from the same two distant loci is small. The presence of multiple shared barcodes between two distant locations (above background) is therefore indicative of a breakpoint where the two locations are joined. After breakpoint identification and refinement, GROC-SVs performs sequence assembly of the linked reads from the relevant breakpoints. This includes the reconstruction of complex events on the basis of breakpoints that are connected by long fragments (**Supplementary Fig. 2b**).

Breakpoint detection

Barcode similarity is highest near a breakpoint and drops off at distances proportional to the fragment size distribution (**Fig. 1**; see **Supplementary Fig. 3** and **Supplementary Note 1** for more detailed explanations). In the matched normal samples, the same region exhibits only low background similarity, indicative of a somatic SV (**Fig. 1b**). Using larger numbers of molecular partitions (barcodes) reduces background, and fragment lengths extend further from each breakpoint when using DNA with higher molecular weight (**Supplementary Fig. 4**).

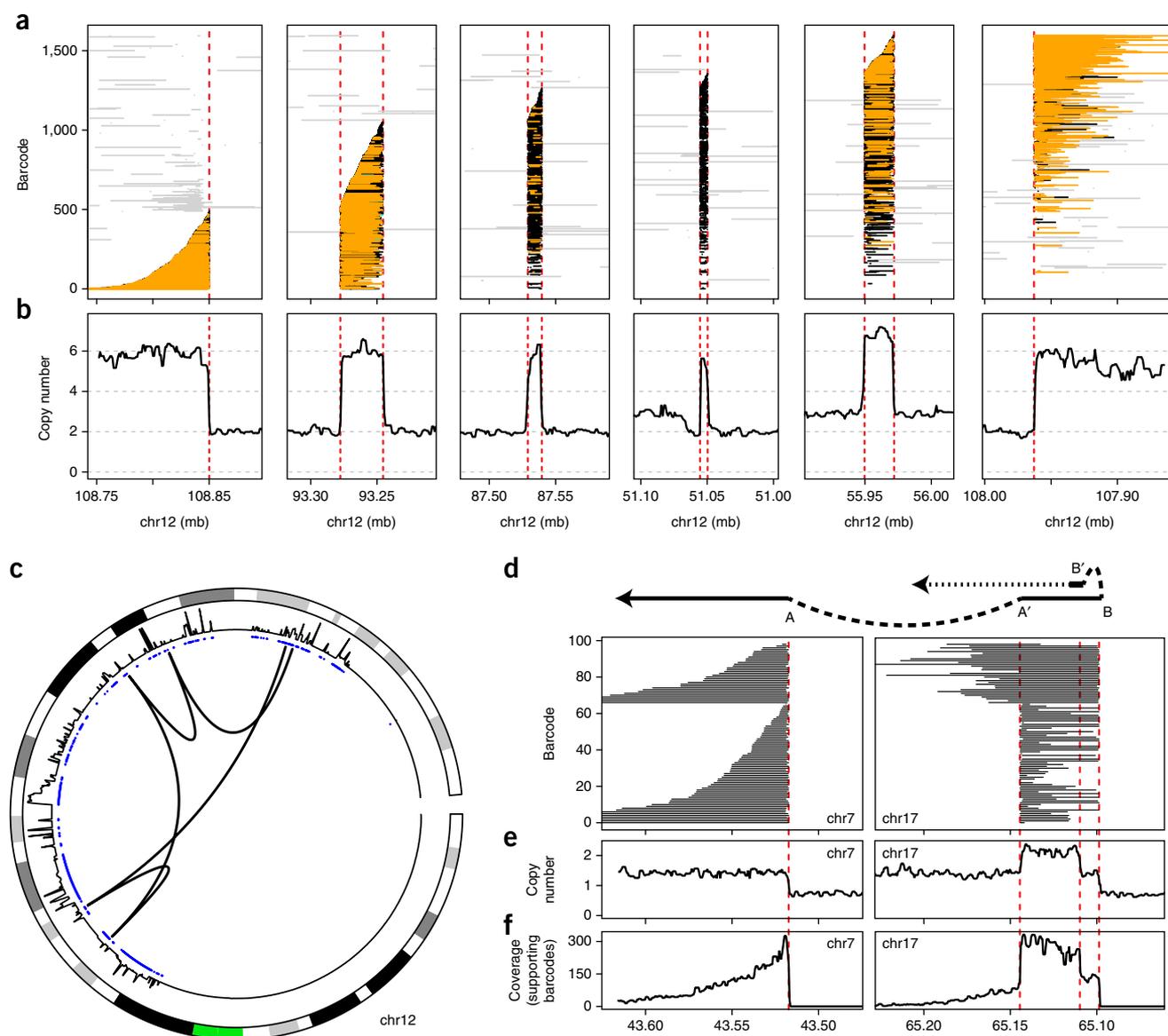


Figure 2 | Reconstruction of complex events. (a–c) Read clouds that support a complex event in the sarcoma. (a) Clouds tile across five consecutive breakpoints, (colors as in Fig. 1), with (b) consistent copy number profiles. (c) Circos plot with arcs depicting breakpoint connections illustrates that the event connects distant segments from the long arm of chromosome 12. From outside to inside, chromosome ideogram (green, centromere), copy number profiles and copy number aberration calls (blue, amplifications; red, deletions) are shown. (d) A complex event in cell line HCC1143 and its corresponding (e) copy number and (f) sequence read coverage.

All supporting read clouds end near the putative breakpoint location (Fig. 1c), a signal that is used during breakpoint refinement. In size-selected samples (as in the sarcoma) the clouds, ordered by their position relative to one side of the breakpoint, tile across the breakpoint such that those starting farthest from the breakpoint tend to extend the least into the other side of the breakpoint. Short-fragment sequencing coverage profiles support changes in copy number at many SV breakpoints (Fig. 1d).

A second example from the sarcoma illustrates the nature of barcode similarity when two breakpoints are in close proximity. In such cases, signal drops off suddenly at two nearby locations along one chromosome axis (Fig. 1e,f). When tiling the read clouds, it becomes apparent that there are two breakpoints at the 93-mb locus (Fig. 1g) which are consistent with copy number profiles

(Fig. 1h). A substantial number of fragments span more than one breakpoint, and this suggests that it is possible to use 10x long-fragment information to directly link breakpoints in proximity (see below).

Sequence assembly and complex event reconstruction

To better characterize breakpoints, GROCVS attempts to perform sequence assembly using the long-fragment information present in the barcoded reads. First, it identifies barcodes that are shared among multiple breakpoints, which suggests that some long fragments spanned across them; breakpoints that do not share barcodes are retained as singletons. For each collection of events, supporting barcodes are identified; and all reads, including unmappable or low-mapping-quality reads, are

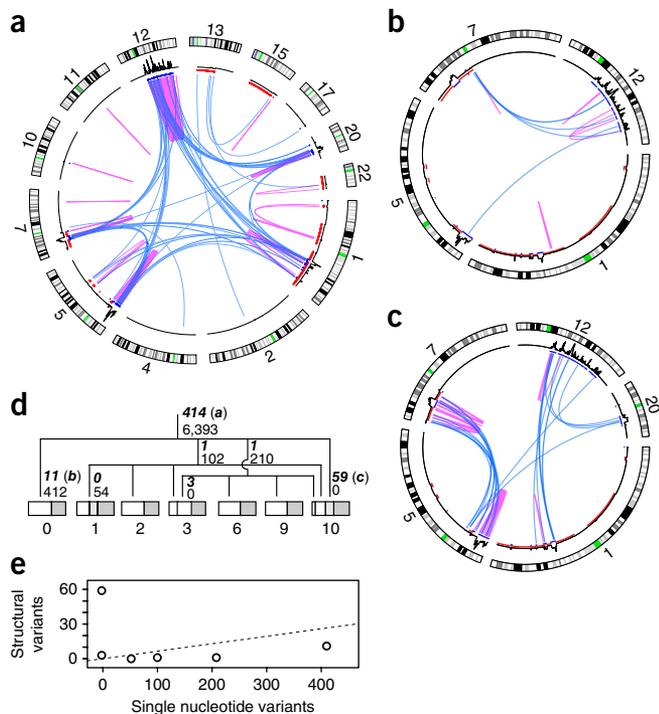


Figure 3 | Somatic genome evolution of the sarcoma. (a) Circos plots of the 414 ancestral (truncal) events. (b) Events private to sample 0. (c) Events private to sample 10. Colors as in **Figure 2c**. Blue, interchromosomal events; magenta, intrachromosomal events. (d) Lineage tree of the samples reconstructed from high-confidence somatic SNVs. Number of SNVs supporting each branch are in roman font, number of breakpoints are in bold italic with circos plot panel letters indicated for plots a–c. Samples are subdivided proportionally to somatic allele frequencies to indicate subclone size. Portion corresponding to normal contribution (e.g., infiltrating lymphocytes) is in dark gray. (e) Number of SVs vs SNVs for each branch in lineage tree. The ancestral branch is shown as a dashed line indicating the rate at which SVs would accumulate relative to SNVs under a constant rate model.

gathered and assembled (**Supplementary Fig. 2b**). The resulting contigs are aligned to the reference genome to identify precise breakpoint locations.

In many cases, we are able to use the sequence assemblies directly to reconstruct the order of multiple genomic segments within complex SVs. However, even when the sequence assemblies are incomplete, we can reconstruct complex events using the fact that adjacent genomic segments within a complex event will share more barcodes than distant segments within the same event.

Using this approach, we identified 12 events with four or more breakpoints and 60 events with two or three breakpoints in the sarcoma. As a fraction of all somatic breakpoints, 204/503 (41%) were assigned to complex events made up of at least two breakpoints. The ordering and assembly of five breakpoints in one example of a complex event that spans 75 kb (**Fig. 2a–c** and **Supplementary Note 1**) illustrates how the clouds tile and thereby connect neighboring breakpoints. Copy number profiles are consistent with the reconstruction (**Fig. 2b**). Strikingly, the variant connects sequence from all over the long arm of chromosome 12 (**Fig. 2c**).

In the nonchromothriptic breast cancer cell line, we reconstructed 11 complex somatic events with a total of 24 breakpoints,

including a large inverted repeat that illustrates both the potential complexity of structural variation and the power of read clouds to resolve it (**Fig. 2d–f**).

Structural variant discovery, comparison and validation

The sarcoma genome harbored substantial structural variation, which was represented by a total of 503 called somatic breakpoints (**Fig. 3a** and **Supplementary Figs. 3a** and **5**). The highest density of events, involving 174 breakpoints, occurred on the long arm of chromosome 12 (**Fig. 3a**).

One expectation regarding SV detection with 10x data is that its high physical coverage improves the signal-to-noise ratio compared to that of standard short-read SV detection approaches. The number of SV-supporting 10x fragments correlated highly with the number of supporting mate pairs ($\rho = 0.89$; **Supplementary Fig. 6**) and fairly well with short fragments ($\rho = 0.71$; **Supplementary Fig. 6b**). Strikingly, a median 3.2 times as many 10x barcodes as short fragments supported an event. The overall validation rate of our breakpoints by mate pairs was 94.6% (424/448), and this increased to 98.6% (351/356) when examining only successfully assembled SVs. Sensitivity and specificity were lower for events substantially smaller than the average fragment lengths (**Supplementary Note 2**). To compare the read cloud approach with previous methods, we applied commonly used tools to our standard Illumina libraries to identify large-scale SVs¹⁴. We found that only 65.1% (375/576) of the short-fragment-called somatic events were validated by mate-pair data (**Supplementary Fig. 7**).

Genome evolution within the sarcoma

The 414 breakpoints present in all sarcoma samples but not in the control are shared events that were present in a common ancestor. They include the chromothripsis on chromosome 12, with the vast majority of the other events involving chromosomes 1, 5, 7 and 20. In addition, we found an ancestral rearrangement followed by high-level amplification harboring the characteristic liposarcoma driver gene *MDM2* (ref. 15).

We also identified 89 SVs that were present in subsets of the samples (but not in the control). The majority of these involved chromosomes 5, 7, and 12; and they were private to one of the samples—59 in sample 10, 11 in sample 0, and 3 in sample 3. The remaining 19 SVs had an indeterminate subclone distribution.

The nonancestral SVs and the inferred presence of subclones suggest that our sampling captured some evolutionary history within the sarcoma. We therefore set out to determine the evolutionary relationships between samples and to analyze the dynamics of SV accumulation based on the inferred phylogenetic tree. We used somatic single-nucleotide variants (SNVs) to build the tree¹⁶ (**Fig. 3**) because they are more common than SVs. In agreement with the SVs and copy number profiles, the majority (6,393/7,171) of high-confidence somatic SNVs were ancestral to all cells in the samples. We identified an additional four subclones based on the presence of SNVs shared between subsets of samples.

When mapping SVs onto the SNV tree, SVs do not appear to accumulate proportionally to the number of cell divisions, but instead they tend to occur in bursts, clustering in evolutionary time. Gradual SV accumulation with cellular generations would be expected to give a constant ratio between the number of SVs

and SNVs present in a subclone. Instead, we see very low numbers of SVs compared with SNVs for most of the lower branches on the evolutionary tree, with the major exception of an outlier subclone that is private to sample 10 (Fig. 3e). This utter lack of agreement between SNV and SV rates suggests that SV accumulation is episodic, similar to what has been observed for copy number variation in breast cancer¹⁷.

Further evidence for the episodic nature of SV accumulation is found in the differential localization of the breakpoints, depending on exactly when they occurred during the evolution of the sarcoma. The 414 truncal events are highly enriched for involvement of chromosome 12, mostly intrachromosomally, with some involvement of chromosomes 1, 5, 7 and 20 (Fig. 3a). The private events in sample 1 mostly fell near regions of chromosomes 7 and 12 that harbor truncal structural variation (Fig. 3b). Strikingly, a large majority (43/59, 73%) of breakpoints present in the subclone private to sample 10 occurred within or between chromosomes 5 and 7 (Fig. 3c). In contrast, only 30% of ancestral mutations occurred within or between those chromosomes. This enrichment was highly significant ($P < 10^{-9}$, Fisher exact test), and this supported the occurrence of a sudden series of events affecting a small portion of the tumor genome. Thus these structural events likely occurred in a short enough time span that SNVs could not accumulate to sufficient levels for the subclone to be observed directly.

DISCUSSION

The high rate of breakpoint validation with mate-pair data demonstrates that 10x data combined with GROC-SVs provides a substantial improvement in large-scale SV detection compared with that of standard short-fragment sequencing and analysis. We note that, because mate-pair libraries span a limited range of fragment sizes, they are not well suited to reconstructing complex SVs. We expect that other methods leveraging read cloud data for breakpoint detection will also show improved accuracy on account of the substantially improved physical coverage and long-distance information.

Breakpoint detection is only the first step in SV characterization. We designed GROC-SVs to take full advantage of 10x data by simultaneously analyzing multiple breakpoints when it detects a complex SV and by performing SV sequence assembly. Sequence assembly serves as both validation (incorrect breakpoint calls would not lead to a consistent assembly), and base-pair-level reconstruction. GROC-SVs thus differs substantially from the current version of the LongRanger package from 10x Genomics, which only performs the SV detection step. In addition, we note that GROC-SVs also supports multisample analysis and so enables accurate calling of somatic SVs when paired tumor and normal samples are available.

We focused our efforts on large-scale rearrangements in tumor genomes because of their importance and the poor performance of short-fragment methods on these types of variants (Supplementary Note 2). To date, genome-scale reconstruction of complex SVs has been limited to cases where the breakpoints are spaced no longer than the fragment insert size (typically ~500 bp); or it has involved indirect inference that events are related, based on their proximity and orientation in the reference genome^{2,18}. Using 10x data, we were able to directly reconstruct the order of large-scale genomic rearrangements involving many breakpoints without the need for any

assumptions about pairs of breakpoints. In the sarcoma genome, where chromothripsis produced dramatic genomic change, we found that 40% of our breakpoints fell within complex SVs, with adjacent breakpoints frequently separated by tens of kilobases.

Most SVs in the sarcoma were shared across all seven spatially distinct locations and therefore must have occurred early in the evolution of the tumor. These ancestral events include the 174 chromosome-12 chromothripsis breakpoints and subsequent copy number amplifications as well as an additional 240 breakpoints. In contrast, while 778 subclonal SNVs were detected, corresponding to four distinct subclone lineages, very few SVs other than the ancestral ones were shared across subclones. Thus, the sarcoma must have undergone an initial period of substantial structural instability, accumulating hundreds of rearrangements and copy number changes before converging to a stable genomic configuration. Our results are concordant with a model recently proposed for liposarcomas in which chromothripsis is followed by breakage–fusion–bridge cycles and subsequent chromosome linearization¹⁹. In addition to the ancestral SVs, we found a subclone private to sample 10 with 59 breakpoints that likely occurred in an additional recent burst of genome instability.

Using GROC-SVs, we show that 10x data allow for direct, data-driven reconstruction of complex structural variation. This reconstruction is accomplished at high sensitivity and excellent specificity compared to short-fragment data and with less laboratory effort and fewer sample requirements than specialized library or mate-pair approaches. Two distinct substrates, a chromothriptic sarcoma and a less highly rearranged breast cancer cell line, demonstrate the wide applicability of the approach. Our evolutionary analysis of the sarcoma predicts substantial future advances in reconstructing the full cancer genome and in understanding each tumor's structural evolution.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank K. Giorda, S. Kyriazopoulou-Panagiotopoulou and M. Schnall-Levin for their assistance in preparing and analyzing the 10x data; and we thank D. Ramazzotti for analyzing mutation spectra. This work was supported by the Stanford Center for Computational, Evolutionary and Human Genomics (N.S.), R01CA183904 (NIH/NCI; R.B.W., S.B. and A.S.), and the BRCA Foundation (A.S.). Certain commercial equipment, instruments or materials are identified in this document. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

AUTHOR CONTRIBUTIONS

N.S., Z.W., A.B., R.B.W., J.M.Z., M.S., S.B. and A.S. designed the experiments and/or analyses. N.S., Z.W., J.M. and D.C. conducted the experiments. N.S. wrote analysis software. N.S. and A.S. analyzed the data. N.S. and A.S. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
2. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
3. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
4. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
5. Chiang, C. *et al.* Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* **44**, 390–397, S1 (2012).
6. Tupler, R. *et al.* A complex chromosome rearrangement with 10 breakpoints: tentative assignment of the locus for Williams syndrome to 4q33----q35.1. *J. Med. Genet.* **29**, 253–255 (1992).
7. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
8. Quinlan, A.R. & Hall, I.M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* **28**, 43–53 (2012).
9. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
10. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
11. Voskoboinik, A. *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2**, e00569 (2013).
12. Bishara, A. *et al.* Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).
13. Zheng, G.X.Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
14. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
15. Oliner, J.D., Kinzler, K.W., Meltzer, P.S., George, D.L. & Vogelstein, B. Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* **358**, 80–83 (1992).
16. Newburger, D.E. *et al.* Genome evolution during progression to breast cancer. *Genome Res.* **23**, 1097–1108 (2013).
17. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
18. Greenman, C.D. *et al.* Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* **22**, 346–361 (2012).
19. Garsed, D.W. *et al.* The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**, 653–667 (2014).

ONLINE METHODS

Sample preparation and library construction. Sections (0.5 cm thick, 14 cm diameter) of a well-differentiated liposarcoma tumor, obtained under informed consent (Stanford IRB) from the Stanford Tissue Bank, were cut into multiple pieces, snap frozen with liquid nitrogen, and stored at -80°C . Genomic DNA was extracted from seven spatially distinct sites of this sarcoma as well as from matched control kidney tissue of the same patient. We extracted genomic DNA from about 20 mg tissue using Genra Puregene Tissue Kit (Qiagen, cat. no. 158667). Tissue was ground in liquid nitrogen, lysed in cell lysis solution and proteinase K and digested with RNase A. Protein was pelleted and removed by adding protein precipitation solution followed by centrifugation. Genomic DNA was precipitated with isopropanol and resuspended in buffer EB. Purified genomic DNA was aliquoted and stored at -20°C .

Genomic DNA was separated by running about 1 μg DNA on a 1% low-melting-point agarose gel using pulsed field gel electrophoresis (PFGE). DNA of size 50–100 kb was then recovered by β -agarase I digestion and filter concentration (NEB, cat. no. M0392S). 1.2 ng of size-selected DNA was partitioned and barcoded using the 10x Genomics GemCode platform¹³. Libraries were then sequenced with a HiSeq2500 to ~ 25 -fold sequence coverage.

Standard short-fragment Illumina libraries were prepared for all seven sarcoma samples plus the matched normal control. Mate-pair libraries were prepared for sarcoma samples 0, 9 and 10 as well as for the matched control.

For short-fragment DNA libraries, 1 μg of total genomic DNA was sheared to 350 bp. PCR-free libraries were then constructed using Illumina's TruSeq DNA PCR-free library preparation kit and sequenced with the Illumina HiSeqX system to ~ 35 -fold sequence coverage.

For large-insert mate-pair libraries, 4 μg of total genomic DNA was fragmented with Tagment Enzyme and gel size-selected to build 7 kb-insert mate-pair libraries using Illumina's Nextera Mate Pair Sample Preparation Kit (FC-132-1001) (tagmentation, strand displacement, gel size selection, circularization, linear DNA digestion, circulated DNA shearing, streptavidin bead binding, end repairing, A-tailing, adaptor ligation, and PCR amplification). Libraries were sequenced with HiSeq2500 to ~ 20 -fold sequence coverage.

Breakpoint detection. GROC-SVs is implemented as a multisample analysis pipeline, allowing the simultaneous analysis of multiple tumor and matched normal samples or of multiple related individuals.

10x Genomics sequencing libraries are first demultiplexed, and droplet barcodes are called using the provided scripts, then reads are aligned to the reference genome using *bwa mem*²⁰ or *RFA*¹² (which has been implemented in the Long Ranger pipeline as the "Lariat" aligner). Barcodes are then ranked in decreasing order by the number of sequenced reads, and barcodes comprising 90% of all reads are retained while the remainder, which are enriched for experimental artifacts, are filtered. Next, read clouds are identified as previously¹³. Briefly, reads with the same barcode are combined into a single barcode if the largest distance between any adjacent reads is less than a certain distance threshold and the reads on either end are of high map quality. This distance threshold was fixed as per ref. 13 at 60 kb for the data produced in this paper; but an appropriate threshold, typically ~ 20 kb for Chromium data, can be learned directly from the data in order to increase sensitivity for smaller events.

GROC-SVs begins SV detection by identifying all barcodes overlapping each 10 kb genomic window and then performing an all-by-all comparison. Some independent fragments with the same barcode can cause a low level of background similarity, typically <1 (Chromium) or 0–5 (GemCode) barcodes at any given pair of positions. A pair of loci (x,y) is considered a structural variant candidate if the number of shared barcodes exceeds that expected based on the total number of barcodes (proportional to copy number) at each locus. For computational efficiency, this initial test is performed as a binomial test (a more rigorous test is applied later for each structural variant).

Next, candidate SV loci are clustered, and candidate breakpoints are extracted based on peaks in the distribution of read cloud ends. This takes advantage of the fact that read clouds are expected to end suddenly near each of the breakpoints; performing this operation only on those barcodes that are shared between the two loci both dramatically improves the signal and reduces the background. Candidate breakpoints are identified in each sample separately.

At this point, the breakpoints have been identified typically to within several kilobases of the correct location. The next step is to perform refinement on the breakpoint coordinates to obtain approximately nucleotide-level accuracy. This step takes all read clouds within 20 kb of the candidate site and selects only those clouds with barcodes shared on both sides of the breakpoint. Then, separately for each breakend (the two half-open intervals that make up each breakpoint), the maximum point of read cloud density is found and then walked toward the putative breakpoint location until the read cloud density drops off suddenly to background levels, indicating the presence of the breakpoint location. We found that this procedure typically identifies the correct breakpoint location to within several nucleotides if the breakpoint is uniquely mappable with short reads. In the case that the breakpoint region is not uniquely mappable, the inferred breakpoint location will be the last well-mappable ($\text{mapq} \geq 30$) position before the breakpoint. Breakpoint refinement occurs across samples together, so all fragments spanning a breakpoint are used for refinement, even if the event is only present in a small subclone within a sample.

Copy numbers were not used in the detection of SVs and were only calculated to gain a better understanding of the context for SVs. Because the coverage profiles for the first-generation 10x GemCode libraries showed substantial GC bias, we used standard PCR-free Illumina libraries to calculate copy number, normalized to the matched normal and normalized for DNA content within a sample. Coverage was typically higher for the tumor samples because of the many, large single-copy genomic regions.

Sequence assembly of breakpoints. Next, a permissive clustering step groups breakpoints together if they share a substantial proportion of their barcodes. Clusters are formed using a simple threshold based on the Jaccard Index, defined as the number of barcodes shared between the loci divided by the total number of barcodes. This Jaccard Index can be viewed as a sort of 'allele frequency,' where the numerator counts the number of fragments supporting the event, and the denominator counts the number of fragments in the reference and alternate alleles. However, this is an approximation, because it is difficult to confidently assign any individual fragment to one allele, since both reference- and

alternate-allele-supporting fragments can end near either breakpoint location. Theoretically, another confounder is the nonzero rate of ‘barcode collisions,’ where one fragment occurs near breakpoint x and an independent fragment occurs near breakpoint y , both in the same barcode. However, barcode collisions typically contribute a negligible amount to the numerator, since the average number of barcode collisions is very small for most genomic regions (<1 for GemCode and $\ll 1$ for Chromium in normal copy number regions, and only appreciably higher for extreme copy number outliers).

Within each cluster, the barcodes supporting each event are pooled together, and all reads originating from these supporting barcodes are collected. Sequence assembly is then performed on the collected reads using `idba_ud`²¹. As each barcode marks multiple fragments, many of the reads do not derive from a breakpoint-supporting genomic region. However, because fragments are randomly assigned to a barcode, these nonsupporting fragments should be distributed randomly throughout the genome. Thus, combined sequence coverage is highest near the breakpoints, which should be covered by every barcode, and low elsewhere. Therefore, most assembled contigs actually derive from the SV haplotype.

As with breakpoint refinement, sequence assembly is performed multisample, so spanning fragments can be used for assembly even if they occur in samples with very low allele frequency. `idba_ud` was selected because of its good performance across a wide range of sequence coverage, which is highest near the breakpoints and then low farther away. Contigs are then aligned against the reference genome, and breakpoint locations are called where appropriate. Note that this assembly process may discover additional breakpoints that were not significant in the genome-wide breakpoint detection step for various reasons.

Phasing germline haplotypes across structural variants. In addition to providing high physical coverage of structural variant breakpoints, the long-fragment information in the 10x data allows for phasing of small variants with respect to the germline haplotypes¹³. Read clouds overlapping a heterozygous short variant can be assigned to one of the haplotypes. The low sequence coverage C_R of each fragment means that some read clouds, especially shorter ones, will not cover a short variant informative for haplotype assignment. However, the high physical coverage C_F results in a high total number of phased fragments for most genomic regions.

Because the structural variant breakpoints are distant from one another in the genome, the haplotypes called by 10x’s Long Ranger software are identified independently for each side of the breakpoint, and so the standard phasing process does not uncover the phase arrangement for the tumor genome. However, nearly all informative fragments near each breakpoint support a single haplotype, and this indicates that each side of the breakpoint only contributes a single haplotype to the event (Fig. 1c,g). Thus we can use the predominant haplotype on either side of a breakpoint to locally phase the genomic regions that participate in the SV.

We identified 239 somatic breakpoints in the sarcoma with at least 20 phased read clouds supporting each side of the breakpoint. Of these events, the vast majority (229, or 96%) were supported by only a single haplotype combination, which is expected because the probability of the same exact SV occurring at the same

position on both haplotypes is vanishingly small. In contrast, systematic errors resulting from, for example, genome repetitiveness, should affect all haplotypes equally. Therefore, the high percentage of events supported by only a single haplotype combination not only supports the validity of our phasing across breakpoints but also provides evidence that the breakpoint calls themselves do not result from substantial systematic biases.

Genome-wide reconstruction of complex events. Following sequence assembly, a more rigorous complex event reconstruction is performed. First, breakends sharing a substantial proportion of barcodes are again clustered together. The resulting clusters are represented as graphs with breakends represented as nodes, and connections between nearby (contiguous genomic segments) and distant (noncontiguous structural variants) breakends represented as edges. Because fragments may span many breakpoints at once, there may be barcode similarity between breakends that are separated by one or more breakpoints. Thus, for each breakend, we first select the assembly-supported breakpoint if one exists. The remaining breakpoints are selected based on the highest barcode support (nearby breakends should share more barcodes than distant ones). This process uses the high-quality information present in the sequence assemblies but can still perform complex event reconstruction even for breakpoints that cannot be sequence assembled.

Postprocessing. During postprocessing, a P value is assigned to each breakpoint. This P value is calculated by randomly sampling the correct number of barcodes for each breakend from the background distribution of fragments per barcode and then calculating the number of shared barcodes. Resampling is performed 100 times; then the significance of the observed versus the resampled number of shared barcodes is calculated using a ranksum test. This resampling procedure takes into account the effect of differences in genome coverage as well as the nonuniform partitioning of fragments across barcodes.

Additional filters are applied, primarily for use when analyzing germline events to identify candidate confounding segmental duplications (segmental duplications should be present in both tumor and matched normal samples and are thus removed when analyzing somatic events). One filter of note compares the observed fragment lengths across breakpoints to those expected based on the background distribution. Structural variants should show long fragment support at tens of kilobases away from each breakend. In contrast, segmental duplications and other repetitive genomic sequences often result in short supporting read clouds.

A final postprocessing step assigns a present-absent call to each event for each sample. This genotype combines the resampling P value calculated above and requires a minimum allele frequency (again calculated using the Jaccard index). Note that heterozygous and homozygous calls are not calculated because these are difficult to accurately define for the different types of structural variant and especially when copy numbers are variable. SV calls were considered to be somatic if there was no more than one supporting barcode in the control sample; results were nearly identical when using cutoffs of zero or two barcodes instead.

Validation and comparison to short-fragment methods. Mate-pair validation was performed by counting the number of mate

pairs in the expected orientation and distance relative to the two breakends. We only used reads with a very conservative mapping quality filter of $\text{mapq} \geq 55$. The rationale for using reads with this high mapq filter was that true events should typically have mates mapping several kilobases away from the breakend, escaping any local repetitiveness around a breakend. We analyzed the background distribution of random genomic regions, and we found that the vast majority of regions shared zero mate pairs, and thus we used a conservative cutoff of 50 mate pairs to consider an event to be validated. We also tried a more lenient cutoff of ten mate pairs and obtained similar results.

Identification of large-scale SVs from the standard short-fragment Illumina sequencing libraries was performed using LUMPY¹⁴. We also performed these analyses using delly²²; but only the LUMPY SV calls are shown, as these validate more consistently using mate-pair data. A threshold of <1 reads in the control sample supporting the event was used to filter out germline events and artifacts, with the remainder of the events therefore inferred to be somatic. This conservative threshold produced the most specific results possible, and a slightly less conservative threshold of <2 control reads did not substantially affect sensitivity, while it did substantially decrease specificity.

Evolutionary analysis. Evolutionary trees relating samples within the sarcoma were built as in refs. 16 and 23. The alternate allele frequencies of the SNVs of the two phylogenetically informative classes are highly consistent with the allele frequencies of the

ancestral SNVs. The frequencies of SNVs present in the mixed-lineage samples (3 and 10) are consistent with one another, with their sums matching the ancestral frequencies. The mutation spectrum of the somatic SNVs (data not shown) closely matches that of germline events; this suggests that they were caused by replication errors without special mutational mechanisms, and that they accumulated at a rate proportional to the number of cell divisions. Finally, as expected, the most phylogenetically similar samples were in close spatial proximity to one another within the tumor (**Supplementary Fig. 5**). These lines of evidence support the idea that we were able to construct a robust evolutionary tree of our samples that could form the basis for interpreting the accumulation of SVs in this tumor (**Fig. 3d**).

Data availability statement. GROC-SVs is open source and available at <https://github.com/grocsvs/grocsvs>. Raw sequencing data are available from dbGaP with accession code [phs001255.v1.p1](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101255).

20. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
21. Peng, Y., Leung, H.C.M., Yiu, S.M. & Chin, F.Y.L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
22. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
23. Weng, Z. *et al.* Cell-lineage heterogeneity and driver mutation recurrence in pre-invasive breast neoplasia. *Genome Med.* **7**, 28 (2015).