nature genetics

Article

https://doi.org/10.1038/s41588-023-01540-6

Most large structural variants in cancer genomes can be detected without long reads

Received: 1 June 2021

Accepted: 19 September 2023

Published online: 9 November 2023

Check for updates

Zi-Ning Choo[1,2,3,4], Julie M. Behr[1,2,5], Aditya Deshpande[1,2,5], Kevin Hadi[1,2,4], Xiaotong Yao[1,2,5], Huasong Tian[1,2,6], Kaori Takai[7], George Zakusilo[7], Joel Rosiene[1,2], Arnaud Da Cruz Paula[8], Britta Weigelt[8], Jeremy Setton[8], Nadeem Riaz[8], Simon N. Powell[8], Klaus Busam[8], Alexander N. Shoushtari[8], Charlotte Ariyan[8], Jorge Reis-Filho[8], Titia de Lange[7] & Marcin Imieliński[1,2,6,9] ✉

Short-read sequencing is the workhorse of cancer genomics yet is thought to miss many structural variants (SVs), particularly large chromosomal alterations. To characterize missing SVs in short-read whole genomes, we analyzed 'loose ends'—local violations of mass balance between adjacent DNA segments. In the landscape of loose ends across 1,330 high-purity cancer whole genomes, most large (>10-kb) clonal SVs were fully resolved by short reads in the 87% of the human genome where copy number could be reliably measured. Some loose ends represent neotelomeres, which we propose as a hallmark of the alternative lengthening of telomeres phenotype. These pan-cancer findings were confirmed by long-molecule profiles of 38 breast cancer and melanoma cases. Our results indicate that aberrant homologous recombination is unlikely to drive the majority of large cancer SVs. Furthermore, analysis of mass balance in short-read whole genome data provides a surprisingly complete picture of cancer chromosomal structure.

It is widely thought that short-read sequencing (SRS), which usually generates ≤150-bp reads, has limited sensitivity for mapping cancer structural variants (SVs; copy number (CN) alterations and rearrangements) owing to the many homologous sequences in the human genome[1]. Indeed, more than two-thirds of the human genome consists of repetitive sequences[2], including transposable elements, satellites and telomeres. SVs that rearrange long homologous repeats are likely to be missed by SRS.

Cancer whole-genome profiling efforts have been carried out almost exclusively with SRS[3–5]. Hence, little is known about the nature and burden of cancer SVs missed by SRS. While most cancer rearrangements detected with SRS have negligible breakend homology[3,6–8], it is also unknown whether additional homologous recombination-driven mutational processes govern the evolution of rearrangements that are undetectable by SRS[1,9].

Owing to mass balance, every copy of every segment in a genome must either have both a left and right neighbor or reside at a chromosome end. Because rearrangements appose previously distant segment ends to create new junctions, CN alterations and rearrangements are physically coupled in the cancer genome; most CN alterations involve a rearrangement, and many rearrangements are associated with a CN alteration[4,10–13].

This coupling can be formalized as 'junction balance constraints' on a graph of genomic segments and their junctions[4] (Fig. 1a).

[1]New York Genome Center, New York, NY, USA. [2]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [3]Tri-institutional MD PhD Program, Weill Cornell Medicine, New York, NY, USA. [4]Physiology and Biophysics PhD Program, Weill Cornell Medicine, New York, NY, USA. [5]Tri-institutional PhD Program in Computational Biology and Medicine, New York, NY, USA. [6]Perlmutter Cancer Center, NYU Grossman School of Medicine, New York, NY, USA. [7]Laboratory of Cell Biology and Genetics, Rockefeller University, New York, NY, USA. [8]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [9]Department of Pathology, NYU Grossman School of Medicine, New York, NY, USA. ✉e-mail: mski@mskilab.org

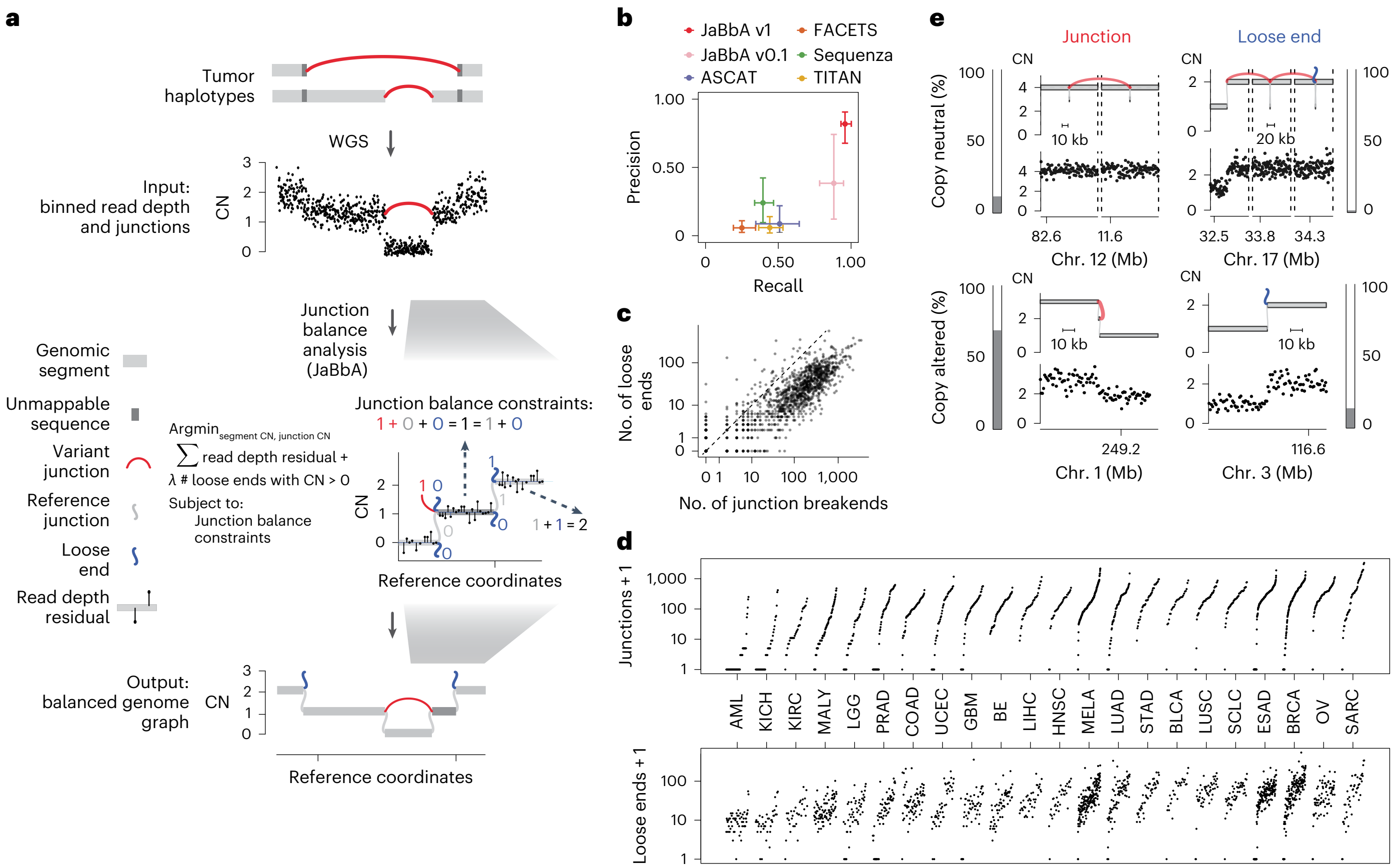

Fig. 1 | Mass balance violations in cancer genomes. a, Schematic for integrated SV detection in JaBbA. Rearranged tumor haplotypes (top) comprise genomic segments connected by variant DNA junctions. These haplotypes produce read depth changes (scatterplot) and variant adjacencies (red edges) in SRS whole-genome profiles (second track from top). JaBbA solves a mixed-integer program to identify the balanced genome graph that optimally explains the input (read depth and adjacencies; third track, right). Graph edges comprise reference or variant junctions and loose ends. Loose ends are placeholder edges that represent local violations of mass balance, which can occur at the breakends of junctions that are missing from the data (Methods). **b**, Precision and recall of SV breakend detection by JaBbA v1 in comparison to other state-of-the-art CN inference algorithms (JaBbA v0.1, ASCAT v2.5.2 (ref. 14), FACETS v0.6.2 (ref. 17), Sequenza v3.0 (ref. 16) and TITAN v1.28 (ref. 15)) in a simulated dataset of 500 samples. Points show medians across all samples, and error bars show the IQR. **c**, Somatic loose end count (*y* axis) versus somatic junction breakend count (*x* axis) identified by JaBbA across a pan-cancer cohort of 1,330 high-purity matched tumor–normal tissue samples. The line shows $x = y$, and points correspond to breakend counts + 1 on a $\log_{10}$ scale. **d**, Number of junction breakends (top) and number of loose ends (bottom) by tumor type. Counts are plotted on a log scale after adding 1. AML, acute myeloid leukemia; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; MALY, malignant lymphoma; LGG, low-grade glioma; PRAD, prostate adenocarcinoma; COAD, colon adenocarcinoma; UCEC, uterine corpus endometrial carcinoma; GBM, glioblastoma multiforme; BE, Barrett's esophagus; LIHC, liver hepatocellular carcinoma; HNSC, head and neck squamous cell carcinoma; MELA, melanoma; LUAD, lung adenocarcinoma; STAD, stomach adenocarcinoma; BLCA, bladder carcinoma; LUSC, lung squamous cell carcinoma; SCLC, small cell lung cancer; ESAD, esophageal adenocarcinoma; BRCA, breast carcinoma; OV, ovarian adenocarcinoma; SARC, sarcoma. **e**, Fractions and examples of copy-neutral and copy-altered breakends associated with junctions and/or loose ends. Outer bar plots show the fractional contribution of each of the four breakend classes (e.g. copy-altered loose ends in the bottom right) to the total number of SV breakends detected by JaBbA. In each subpanel, the top track shows the balanced genome graph with plot elements as in **a** and the bottom track shows binned purity- and ploidy-transformed tumor read depth.

These constraints state that the CN of each genomic segment is equal to the CN of the junctions connecting to its left and right sides. Enforcing these and other constraints within a statistical model enables the inference of balanced genome graphs and high-fidelity CN profiles from whole-genome SRS data, as shown with our previously published JaBbA (v0.1) algorithm[4].

JaBbA's statistical model allows for 'loose ends', which are 'placeholder' adjacencies that allow the graph to satisfy junction balance while violating mass balance (Fig. 1a). Loose ends allow JaBbA to be robust to missing data but also represent hypotheses about unmapped junctions. We reasoned that analysis of loose ends in JaBbA could be used to test the completeness of cancer genome reconstructions from SRS and assess the nature of missing SVs in SRS profiles. In particular, we focused on large (>10-kb) SVs that give rise to clonal chromosomal alterations in cancers (referred to as SVs below for brevity, unless otherwise qualified). Our goal was to understand the impact of mutational processes that specifically rearrange repetitive sequences, including aberrant homologous recombination, on cancer chromosomal structure.

Results

JaBbA v1 outperforms previous CN algorithms

We enhanced our previous JaBbA (v0.1; ref. 4) model with several methodological innovations to increase robustness to read depth waviness, improve algorithm convergence and enforce junction balance for allele-specific as well as total CN (Extended Data Fig. 1a–d and Methods). We also rigorously defined 'CN-unmappable' regions in the genome as positions surrounded by >90% repetitive bases in their 1-kb vicinity. CN-unmappable regions accounted for 13% of the genome (across read lengths and genome builds), primarily comprised regions in or around telomeres and centromeres, and showed high variance in read depth across a panel of diploid normal samples (Methods and

Extended Data Fig. 2). We then limited analysis with the updated model (JaBbA v1) to the 87% of the human genome that was CN-mappable.

To assess the accuracy of JaBbA v1 for SV breakend detection in CN-mappable regions, we simulated 500 SRS whole-genome profiles comprising binned (1-kb) read depth, single nucleotide polymorphism (SNP) read counts and SV junctions (Extended Data Fig. 3a–d and Methods). In these simulations, JaBbA v1 loose ends showed substantially higher precision (median of 43% versus 5%) and recall (median of 70% versus 54%) than JaBbA v0.1 loose ends for missing CN-mappable SVs in high-purity (>0.5) cancer genomes (Extended Data Fig. 3e). JaBbA v1 also showed markedly improved accuracy for overall CN-mappable SV breakend inference relative to both JaBbA v0.1 and four state-of-the-art cancer CN inference algorithms (ASCAT[14], TITAN[15], Sequenza[16] and FACETS[17]) (Extended Data Fig. 3f), particularly for high-purity samples (median precision of 82% (68–91%) and median recall of 96% (93–100%), with the interquartile range (IQR) in parentheses) (Fig. 1b). JaBbA v1 also accurately estimated both total and allelic CN (Extended Data Fig. 3g), suggesting that JaBbA v1 is a state-of-the-art algorithm for the inference of CN and missing SVs in cancer genomes.

Pan-cancer landscape of loose ends

We next applied JaBbA v1 to 1,330 high-purity tumor and matched normal SRS profiles previously analyzed in Hadi et al.[4] (see Methods for details), identifying 154,322 (clonal and somatic) junctions (median of 63 per tumor sample) and 48,835 somatic loose ends (median of 21 per tumor sample). The somatic loose end burden per sample varied across a 200-fold range and was correlated (Spearman $R^2 = 0.68$) with the junction burden (Fig. 1c,d).

Junction breakends may be reciprocal, meaning that they are near (within 10 kb) of another breakend with opposite orientation. Reciprocal breakends are usually copy-neutral (Fig. 1e, top left) which makes them difficult to detect through classic CN analyses. JaBbA's bookkeeping of mass balance across segments and junctions enables sensitive detection of reciprocal and nonreciprocal SVs at both copy-neutral and copy-altered genomic regions (Extended Data Fig. 4a–e). Across cancer, we found that most (85%) cancer junctions were both nonreciprocal and copy-altered (Fig. 1e, bottom left). Such junctions can arise from inherently nonreciprocal SVs, such as simple deletions, or begin as reciprocal translocations that undergo subsequent loss or gain of one of the derivative alleles (Extended Data Fig. 4f). Like somatic junction breakends, somatic loose ends were predominantly (92%) copy-altered (Fig. 1e, bottom right), although copy-neutral loose ends were also identified (Fig. 1e, top right). Taken together, these results suggest that loose ends arise by breakage and repair mutational processes similar to those generating junction breakends.

Loose ends harbor repetitive and foreign sequences

To study the sequence context around loose ends, we defined a canonical axis originating at the loose end with coordinates increasing along the DNA strand whose 3′ terminus matches the side of a segment on which a loose end is found, which we refer to as the loose end's 'forward' strand (Fig. 2a). We next asked whether loose ends occurred preferentially at reference sequence repeats. Indeed, we found that unmappable bases were enriched near loose ends, most frequently LINE elements (Fig. 2b and Extended Data Fig. 5a). We next reasoned that some loose ends would result from the somatic fusion of mappable bases to unalignable sequences. Confirming this, we found a tumor-specific enrichment of repetitive and foreign sequences, including satellite and viral sequences, mated to reads on the forward (but not reverse) strand of somatic loose ends (Fig. 2c and Extended Data Fig. 5b).

To identify distinct classes of repetitive SVs missing from SRS whole-genome profiles, we systematically classified tumor-specific sequences fused to each somatic loose end through assembly or consensus alignment (Fig. 2d and Methods). Overall, 55% of somatic loose ends showed evidence of tumor-specific fusion to a distal sequence. For over half of these (33% of somatic loose ends), the distal sequence aligned uniquely, indicating that these were fully mapped breakends missed by the initial junction caller (Fig. 2e) (SvAbA[18]). In 23% of somatic loose ends (3% of detected breakends), the distal sequence was repetitive or foreign and could not be unambiguously placed on any reference (ambiguously mapped breakends; Fig. 2e). Finally, 45% of somatic loose ends (6% of detected breakends) did not map to any distal location (partially mapped breakends; Fig. 2e). Notably, partially mapped breakends were enriched in boundaries of large (>1-Mb) CN-unmappable regions (odds ratio (OR) = 3.8; $P < 2 \times 10^{-16}$) (Extended Data Fig. 5c), indicating that some represented CN changes shifted away from a CN-unmappable SV breakend (for example, centromeric breakends causing arm-level chromosomal changes).

Combining fully mapped breakends across both loose ends and junctions indicated that 91% of JaBbA v1 breakends could be uniquely mapped. Notably, the fraction of partially or ambiguously mapped breakends did not vary substantially across cancer types (Extended Data Fig. 5d; range of 5–33%) or established cancer drivers (Extended Data Fig. 5e; range of 0–38%), although we observed tumor types (for example, acute myeloid leukemia) and cancer genes (*SMARCB1*, *TSC2* and *FGFR3*) with higher (>25%) fractional burdens. Given the estimated recall of JaBbA v1 (~96%), these results suggest that 87% of cancer SVs in the 87% of the genome that is CN-mappable can be fully resolved by SRS.

Long-molecule validation

To orthogonally assess these SRS-derived estimates of missing somatic SVs, we profiled the whole genomes of 11 melanoma ($n = 10$) and breast cancer ($n = 1$) tumor samples and their matched normal tissues with both SRS and Oxford Nanopore Technologies long-read sequencing (LRS; median read N50 of 11 kb; median coverage of 73× and 32× for tumor and normal samples, respectively). After calling large (>10-kb) somatic SVs in CN-mappable regions (Methods), we found a strong overlap (87%, 7,258 breakends) between LRS and SRS breakends, including 77% overlap with fully mapped SRS breakends (Fig. 2f). The majority of junction calls identified by either platform had local read depth changes that were consistent with breakend topology; reciprocal breakends were copy-neutral, whereas nonreciprocal breakends showed a CN drop along their forward strand (Extended Data Fig. 6a). This analysis along with manual inspection of long and short read support at inidivudal junctions (Extended Data Fig. 6b) suggested that both SRS-only and LRS-only junctions comprise largely true positives; combining SRS and LRS breakend counts suggests that SRS missed ~12% of breakends. This result is consistent with our simulation-based estimate of recall (Fig. 1b and Extended Data Fig. 3f). Notably, we found a similar proportion of reciprocal and non-reciprocal breakends among those detected and missed by SRS (Fig. 2f), indicating that reciprocal and copy-neutral breakends do not comprise the bulk of missed structural variation in cancer genomes. These results confirm our SRS findings that most cancer SVs are nonreciprocal and copy-altered (Fig. 1e).

We next asked whether LRS improved SV event detection, which relies on the recognition of high-order patterns across multiple junctions[3,4]. Although LRS did not help identify many additional simple or complex events relative to SRS (Fig. 2g), LRS junctions also resolved breakends at complex SVs found by SRS, including for chromothripsis, pyrgo, rigma and templated insertion chains[3,4]. The incorporation of LRS junctions enabled more complete haplotype reconstruction at loci where SRS found loose ends (Fig. 2h).

As additional validation of our results, we analyzed 27 high-purity (purity of >0.5) breast cancer and matched normal samples with both SRS and synthetic LRS (sLRS) whole-genome profiles (10x Genomics linked reads, median N50 molecule length of 23 kb, median coverage of 173× and 98× in tumor and normal samples, respectively; Methods)[19]. Similar to LRS, most sLRS SV calls (Methods) overlapped with SRS breakends, showed concordant patterns of reciprocality and CN change, and

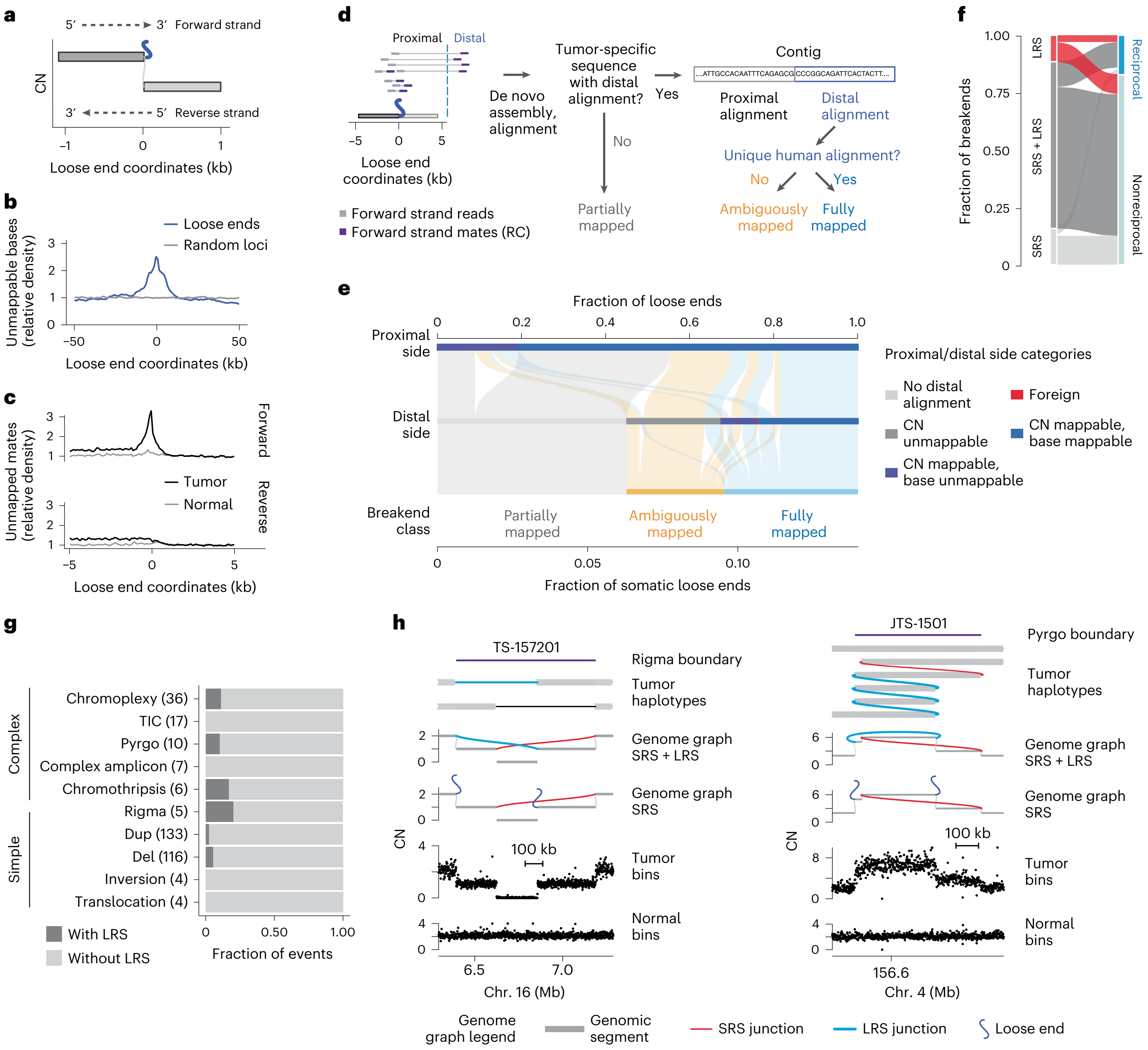

Fig. 2 | Loose ends pinpoint missing cancer SVs. a, Loose end coordinates are centered at each loose end and increase in the 5′ to 3′ direction along the forward strand. For a loose end arising from the right side of its associated reference genomic segment (that is, the side with larger reference genomic coordinates), the forward strand is the positive reference strand, that is, the strand with increasing reference coordinates along its 5′ to 3′ direction. Conversely, for a loose end arising from the left side of its associated reference genomic segment, the forward strand is the negative reference strand. **b**, Density of unmappable bases around loose ends (Methods). **c**, Density of uniquely mapping reads with unmapped (i.e. non-uniquely aligning) mates around loose ends. **d**, Subclassification of loose ends based on local assembly and consensus alignment (Methods). RC, reverse complement. **e**, Alluvial plot showing each loose end class (bottom row) and the mappability tier of the distal (middle row) and proximal (top row) ends of breakend sequences obtained through local assembly or consensus alignment. **f**, Alluvial plot comparing SRS and LRS breakend calls. The fraction of breakends identified by LRS only, SRS only and both platforms (LRS and SRS) is shown (left), stratified by whether the breakend was reciprocal to another breakend in the same sample (right). LRS breakends were taken from tumor-specific junctions found by at least two of four LRS SV callers (SVIM[49], cuteSV[50], Sniffles2 (ref. 51) and SAVANA[52]). SRS breakends comprise junction breakends and loose ends in the JaBbA v1 genome graph. **g**, Stacked barplots showing the fraction of complex SVs called from genome graphs with versus without the addition of LRS junctions. DEL, deletion; DUP, duplication; TIC, templated insertion chain. **h**, Examples of a rigma (left) and pyrgo (right) identified by LRS and missed by SRS. Tracks from top to bottom show tumor haplotype reconstructions, a genome graph with SRS and LRS junctions, a genome graph with only SRS junctions, and purity- and ploidy-transformed read depth. Pyrgo and rigma boundaries are delineated by a purple line at the top of each plot.

yielded similar complex SV calls in sLRS junction-augmented genome graphs (Extended Data Fig. 6c–e). These breast cancer and melanoma LRS and sLRS results are consistent with our pan-cancer finding that SRS captures most large cancer SVs in CN-mappable regions.

Loose ends reveal neotelomeres

We next sought to investigate specific mutational processes engendering loose ends. We observed that a fraction (4.8%) of ambiguously mapped loose ends (0.01% of all breakends) were fused to telomere

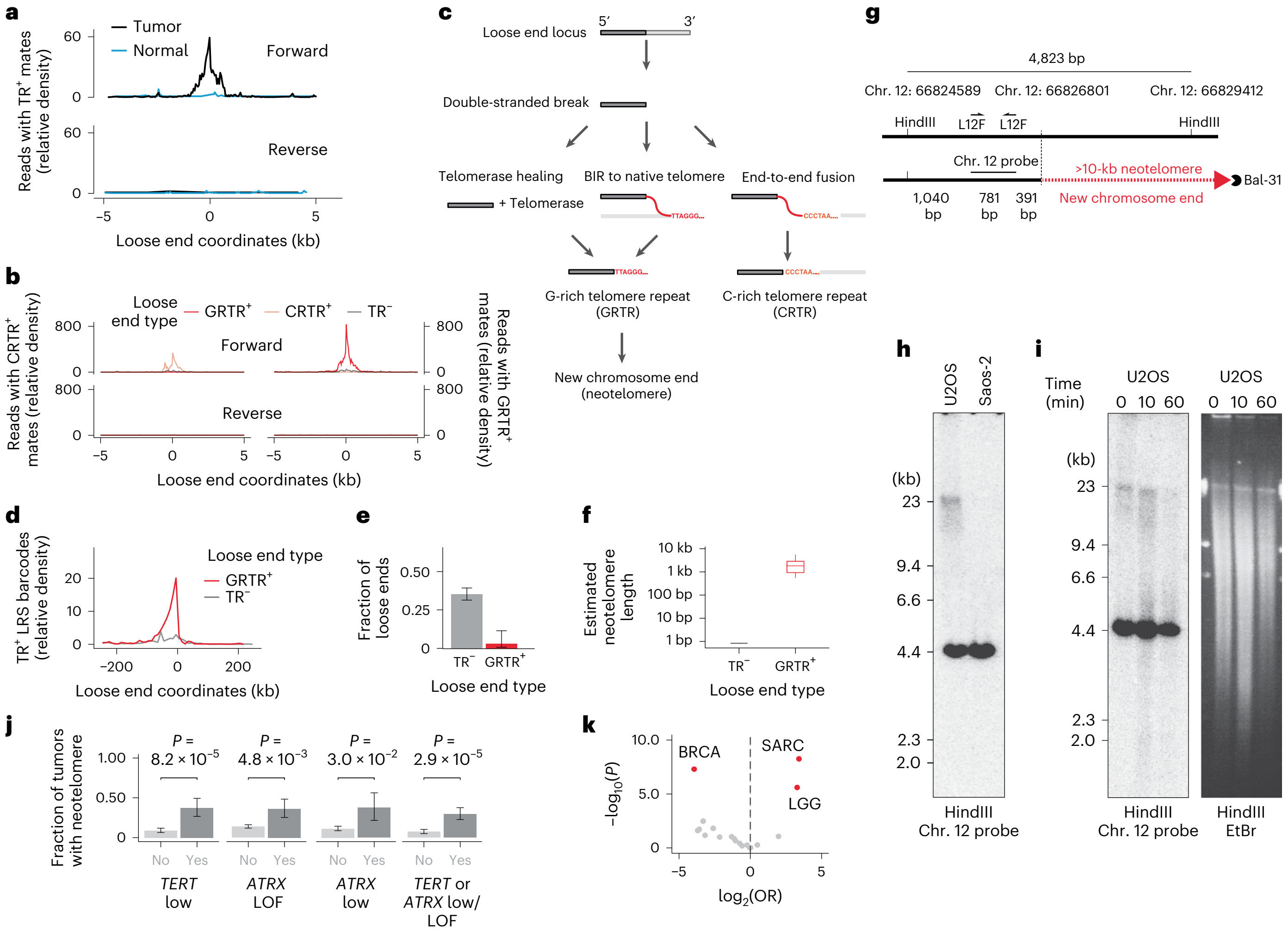

Fig. 3 | Loose ends reveal neotelomeres. a, Density of reads mated to telomere repeats near loose ends. TR, telomere repeat. **b**, Density of reads mated to GRTRs and CRTRs on the forward and reverse strands of GRTR+, CRTR+ and telomere repeat-negative loose ends. **c**, Potential etiologies of telomere repeats fused to loose ends. **d**, Density of sLRS barcodes harboring telomere repeat-positive read pairs near GRTR+ loose ends relative to telomere repeat-negative loose ends. Telomere repeat-positive read pairs are defined as a read pair in which one mate is spanned entirely by G-rich telomere repeats and the other by C-rich telomere repeats. **e**, Fraction of loose ends fused to a unique distal interstitial location via sLRS (Methods). Data are shown as mean ± 95% confidence interval (CI) (n = 71 GRTR+ loose ends and 28 CRTR+ loose ends from 14 tumor samples). **f**, Estimated telomere length at GRTR+ loose ends compared to telomere repeat-negative loose ends (n = 71 GRTR+ loose ends from 14 tumor samples). In box plots, the line represents the median, the body represents the IQR and whiskers extend to 1.5 times the IQR. **g**, Schematic of the neotelomere detection assay. **h**, Southern blot showing a diffuse ~23-kb band in U2OS cells but not in a control cell line (Saos-2). Both cell lines show the 4.4-kb HindIII control band. **i**, The U2OS-specific band disappears after exonuclease (Bal-31) digestion of genomic DNA before HindIII digestion (left) without altering the overall size distribution of DNA fragments (right). EtBr, ethidium bromide. Time refers to length of Bal-31 exposure. For **h**,**i**, the experiment was repeated four times with similar results. Panels show uncropped images of the entire gel lanes. **j**, Fraction of tumors with a neotelomere (that is, GRTR+ loose end) across the given categories. '*ATRX* low' corresponds to *ATRX* RPKM of <500 and '*TERT* low' corresponds to *TERT* RPKM of 0. Error bars are the 95% CIs on the binomial proportion. LOF, loss of function. **k**, Tumor type enrichment of GRTR+ loose ends. Tumor types with a statistically significant association with GRTR+ loose end burden are highlighted in red. See the Fig. 1d legend for definitions of the abbreviations. In **j**,**k**, P values were calculated by two-sided Wald's test on the coefficients of a negative binomial generalized linear model (Methods). In **k**, values with |log(OR)| > log(1.5) and false discovery rate (FDR) < 0.1 after Benjamini–Hochberg correction are highlighted in red.

repeats, as evidenced by telomere repeat-positive sequences mated to reads on the positive loose end strand (Fig. 3a). We refer to these breakends as telomere repeat-positive loose ends and surmised that they might represent neotelomeres, telomere-stabilized chromosome ends at previously interstitial genomic loci.

Telomere repeat-positive mates were found on the forward strand of telomere repeat-positive loose ends, but not on the reverse strand or in matched normal samples (Fig. 3a), indicating that these were neither telomere insertions[20,21] nor constitutional neotelomeres[22,23]. Deeper analysis of telomere repeats at loose ends revealed strong strand bias, with loose ends harboring either G-rich (GRTR) or C-rich (CRTR) repeats but not both (Fig. 3b). The GRTR pattern is consistent with a neotelomere, whereas the CRTR pattern is consistent with the fusion of an interstitial sequence to a native chromosome end (Fig. 3c, right). The predominance of the GRTR pattern among telomere repeat-positive loose ends, in combination with the tumor specificity and forward strand bias, suggested that somatic neotelomeres are frequent in cancer.

To better assess sequences fused to GRTR+ loose ends, we profiled three cancer cell lines (U2OS, NCI-H526 and NCI-H838) with sLRS (Methods). We found telomere repeat-positive linked reads within 5 kb of 26 of 31 GRTR+ loose ends (83.8%) (Methods). Telomere repeat-positive linked reads were found up to 50 kb upstream of each GRTR$^+$ loose end, indicating power to map distal fusion partners at these loci (Fig. 3d). In contrast to sLRS junctions and telomere repeat-negative loose ends, linked reads at GRTR$^+$ loose ends rarely (<1.5%) mapped to distant chromosomal locations, consistent with new chromosome ends (Fig. 3e). Quantitative analysis of repeat counts at linked reads mapping to these loci (Methods) revealed 2.4 ± 1.3 (s.d.) kb of telomere repeats per GRTR$^+$ locus, in line with previous estimates of native cancer telomere lengths[20] (Fig. 3f).

To confirm that GRTR$^+$ loose ends were indeed chromosome ends, we performed Southern blot analysis on restriction-digested U2OS and control (Saos-2) genomic DNA using radiolabeled probes against two U2OS GRTR$^+$ loose ends. At each locus (Fig. 3g and Extended Data Fig. 7a), we found a small (<5-kb) band consistent with an unaltered reference allele and a longer U2OS-specific diffuse band consistent with a neotelomere (Fig. 3h and Extended Data Fig. 7b). To further investigate the nature of these nonreference bands, we subjected intact genomic DNA to exonuclease (Bal-31) digestion[24]. The U2OS-specific (but not wild-type) bands disappeared with prolonged exonuclease exposure (Fig. 3i and Extended Data Fig. 7c), consistent with their origin at a chromosome end. These results establish these two U2OS GRTR$^+$ loose ends as bona fide neotelomeres.

We next hypothesized that telomerase-mediated healing of double-stranded DNA breaks might give rise to neotelomeres (Fig. 3c, left)[25]. However, neotelomeres were not found more frequently in tumors that amplified *TERT* or expressed it at high levels (CN > 2 ploidy, expression *z* score > 2). Instead, neotelomeres were enriched in samples with low or negligible *TERT* expression (reads per kilobase per million mapped reads (RPKM) = 0) (Fig. 3j). Tumors that lack telomerase may activate the alternative lengthening of telomeres (ALT) pathway, a break-induced replication (BIR) process (Fig. 3c, middle) suppressed by ATRX[26]. Indeed, we found that neotelomeres were significantly more common in tumors harboring truncating mutations in *ATRX* than in *ATRX*-wild-type cancers (Fig. 3j). Furthermore, we found that several ALT-associated cancers, including sarcomas (18%; OR = 6.47; $P = 1.95 \times 10^{-5}$) and low-grade gliomas (12.3%; OR = 3.92; $P = 4.1 \times 10^{-3}$), had the highest rate of GRTR$^+$ loose ends relative to other tumor types (Fig. 3k). These results indicate that GRTR$^+$ loose ends and neotelomeres may be a new hallmark of the ALT phenotype.

Loose ends link viral integration to amplicon formation

Surveying additional mutational processes engendering loose ends, we found ambiguously mapped somatic breakends fused to viral sequences, indicating junctional viral integration at large SVs (Extended Data Fig. 8a). While the integration of viral sequences into otherwise unrearranged loci (Extended Data Fig. 8a, left) has been widely studied in cancer[27,28], the role of viruses in causing chromosomal-scale SVs (Extended Data Fig. 8a, right) has been a topic of only recent interest[29–31]. Somatic loose ends harboring tumor-specific viral sequence (viral loose ends) were rare overall (~1% of cancers), although enriched in cancer types with viral etiology in our dataset[4]: cervical squamous cell carcinoma (CESC; 32%), liver hepatocellular carcinoma (LIHC; 13%) and head and neck squamous cell carcinoma (HNSC; 7%) (Extended Data Fig. 8b). Consistent with previously characterized viral integration patterns, we found viral loose ends fused to oncogenic HPV sequences in CESC and HNSC and hepatitis B virus (HBV) sequences in LIHC[27].

Breakends initiating complex amplifications are themselves likely to be amplified[4]. Viral loose ends were frequently amplified (CN > 7) relative to nonviral loose ends ($P = 1.7 \times 10^{-4}$; OR = 8.66) (Extended Data Fig. 8c), and HPV-16 loose ends had higher mean CN than either HPV-18 or HBV loose ends ($P = 8.2 \times 10^{-3}$ and $P = 2.2 \times 10^{-5}$, respectively, Extended Data Fig. 8d). Among these was an HNSC tumor (TCGA-4077) locus where two high-copy viral loose ends on chromosome 14 flanking an intronic region of the *RAD51B* gene were fused to opposite ends of the HPV-16 genome (Extended Data Fig. 8e). This locus is consistent with an ecDNA where HPV-16 is fused between two ends of a long-range duplication junction. This and other similar amplicon structures with high-copy viral loose ends (Extended Data Fig. 8e,f) point to HPV-16 integration as an initiating event in SV evolution, rather than a viral insertion into an existing ecDNA.

Crossover between parental homologs is rare in cancer

We next asked whether loose ends could be used to assess the contribution of aberrant homologous recombination to cancer rearrangements. Homologous recombination-driven crossover between parental homologs (allelic homologous recombination, or AHR) is a hallmark of meiosis[32]. Although AHR has been observed in somatic cells[33], its contribution to cancer structural variation is unclear. AHR crossovers lead to segmental uniparental disomy (UPD) in approximately half of segregants (Fig. 4a, left). In balanced allelic graphs, AHR crossovers manifest as reciprocal pairs of partially mapped and copy-neutral loose ends on distinct parental homologs (Fig. 4b, left, and Methods). Notably, this form of UPD (AHR-UPD) is mechanistically distinct from UPD arising through progressive acquisition of nonhomologous recombination (for example, end joining)-driven rearrangements and/or chromosomal missegregation (progressive UPD, or P-UPD; Fig. 4a,b, right).

In our simulations (Extended Data Fig. 3a and Methods), JaBbA v1 distinguished AHR-UPD from P-UPD with both high precision (84.4%) and high recall (87.4%), substantially outperforming previous allelic CN algorithms (with precision ranging from 11–44%) (Extended Data Fig. 9a,b). Analysis of segment width distributions showed that AHR-UPD was distinct from P-UPD, whose distribution closely mirrored that of other forms of loss of heterozygosity (LOH; Fig. 4c). Likewise, AHR-UPD events were large (median width of 19.8 Mb), unlike P-UPD events (median width of 0.69 Mb) and other forms of LOH (median width of 0.62 Mb), which were focal (Fig. 4c).

Although AHR was found in many cancers (24% of all tumors) and specific tumor types (for example, 55% of cases of malignant lymphoma) (Extended Data Fig. 9c), it contributed to a minority of UPD events, most of which were progressive (31% P-UPD versus 1% AHR-UPD by total width) (Fig. 4d). Overall, a small minority of detected cancer breakends (<1%) arose by AHR (including non-UPD LOH). On the basis of an approximate rate of 0.5 AHR events per tumor and 100 cell divisions in the average ancestral cancer clone, and barring effects of selection, we estimate a rate of 10^{-12} AHR events per base pair per cell division. This is four orders of magnitude lower than the rate of meiotic recombination in human gametes, suggesting that AHR events are infrequent in somatic evolution[34].

Germline but not somatic loose ends are consistent with NAHR

A second mechanism by which aberrant homologous recombination can cause large SVs is through non-AHR (NAHR), or crossover between long (>500-bp) stretches of nearly identical genomic sequences at distant haploid coordinates[32,35,36] (Fig. 4e). We reasoned that such SVs would engender pairs of loose ends with substantial (>500-bp) strand-specific sequence homology in their vicinity (Extended Data Fig. 10a and Methods)[36]. Indeed, the burden of homologous loose end pairs accurately reflected the true NAHR burden across a compendium of simulated SRS tumor whole-genome profiles (Extended Data Fig. 3a) harboring a wide range of NAHR SV fractions (1–10%) (Fig. 4f).

Analyzing breakend pairs within each tumor, we found that approximately 20% of germline loose ends (Methods) were consistent with NAHR in contrast to only about 0.5% of somatic loose ends (and 0.06% of all somatic SV breakends) (Fig. 4g). These findings are consistent with prior observations about the substantial role of NAHR in germline variation[8,37]. The somatic NAHR burden did not vary by tumor type nor was it lower in tumors harboring biallelic pathogenic mutations in DNA repair genes, including frequently mutated homologous recombination pathway mediators (*BRCA1*, *BRCA2*, *PALB2* and

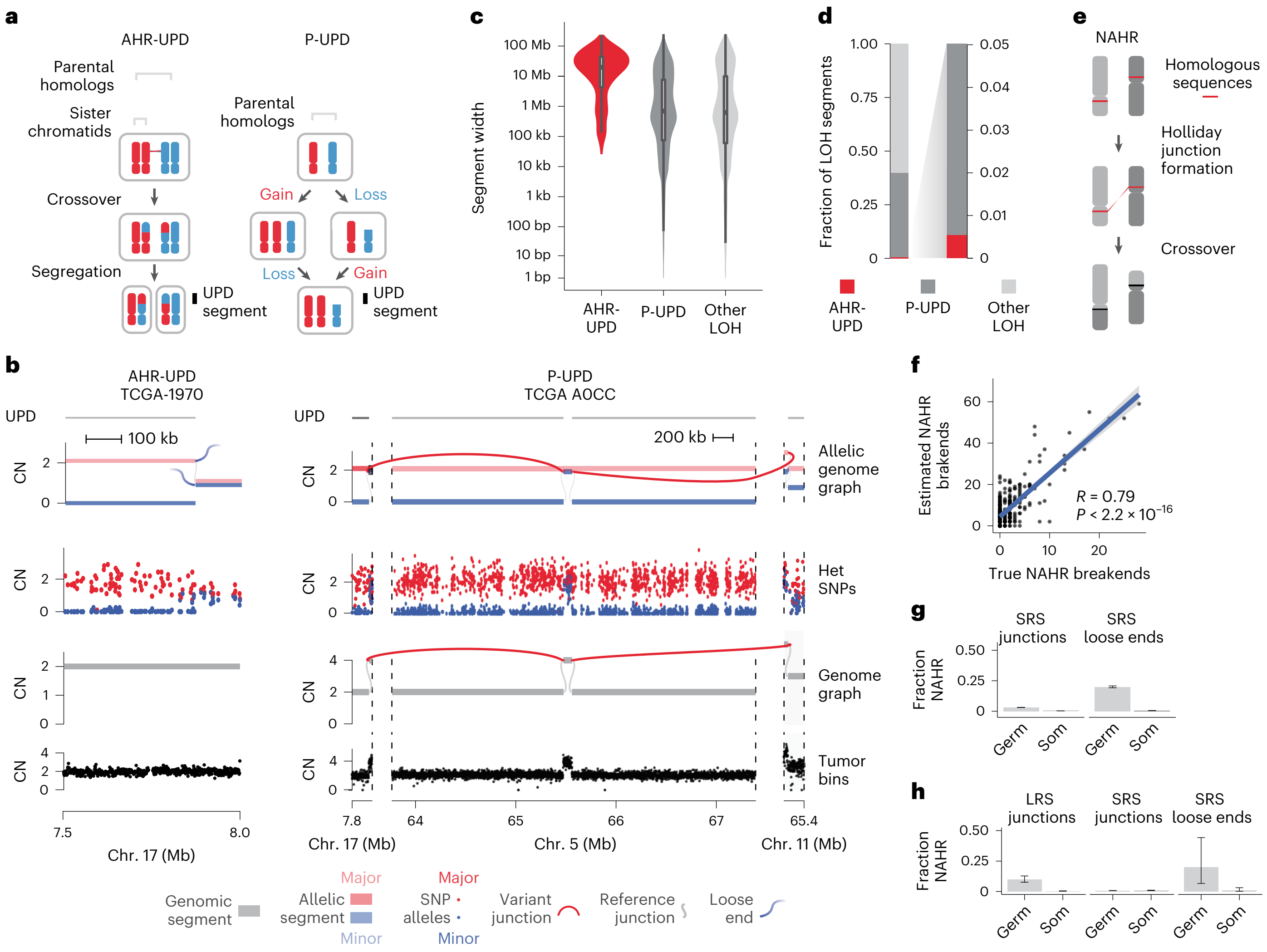

Fig. 4 | AHR rarely drives CN-mappable breakends. a, Schematic showing mechanistic differences between AHR and P-UPD, two mechanisms that give rise to segmental UPD. **b**, Examples of AHR-UPD (left) and P-UPD (right). The allelic graph (top subpanel) shows parental homolog-specific CN, which matches purity- and ploidy-transformed allelic SNP read counts (scatterplot, second subpanel from top) (Supplementary Information). The AHR-UPD locus shows no breakends in the total CN JaBbA v1 graph (third subpanel) but a pair of loose ends in the allelic graph. By contrast, the P-UPD locus does not harbor a pair of allelic graph loose ends, but rather contains a copy-altered breakend in both the allelic and total CN JaBbA v1 graphs. Het, heterozygous. **c**, Width distribution of segments produced by AHR-UPD, P-UPD and all other LOH (n = 545 AHR-UPD ranges, 39,877 P-UPD ranges and 61,469 other LOH ranges from 1,330 tumors). In box plots, the line represents the median, the body represents the IQR and whiskers extend to 1.5 times the IQR. **d**, Fractional contribution of P-UPD, AHR-UPD and other forms of LOH to the total number of LOH segments. **e**, Schematic of NAHR. **f**, Number of estimated (y axis) versus true (x axis) NAHR-mediated breakends per simulated sample (n = 500 simulated genomes). The blue line shows the line of best fit, with Pearson's correlation coefficient provided on the graph; error bands show the standard error of the prediction. The P value was calculated from the t distribution of Pearson's correlation coefficient test statistic. **g**, Fraction of somatic junctions, somatic loose ends and germline loose ends consistent with NAHR rearrangements in the SRS pan-cancer whole-genome cohort (n = 1,330 samples). Error bars represent the 95% CIs on the binomial proportion. Germ, germline; Som, somatic. **h**, Fraction of germline and somatic LRS junctions, SRS junctions and SRS loose ends consistent with NAHR in a separate melanoma and breast cancer cohort with paired SRS and LRS whole-genome profiles (n = 11 samples). Error bars represent 95% CIs on the binomial proportion.

RAD51C). In summary, given a mean of 0.16 somatic NAHR events per tumor occurring across an estimated eligible territory of 2.8×10^8 homologous position pairs, we estimate a somatic NAHR density of 6×10^{-10} events per cancer genome bp^2 (Methods).

To validate these SRS findings in long-molecule whole-genome profiles, we analyzed 38 melanoma and breast cancer cases profiled with SRS and either LRS or sLRS. Both LRS and sLRS data confirmed our SRS findings that somatic NAHR SVs were rare (<1% of LRS junction calls) while germline NAHR SV events were common (Fig. 4h and Extended Data Fig. 10b–e). Notably, we did not identify any reciprocal somatic NAHR rearrangements, a class of SVs that may potentially be missed through analysis of SRS loose ends.

Extrapolating beyond the CN-mappable genome

The analyses described above were limited to the 87% of the genome where CN could be reliably measured with SRS (Fig. 5a). The remaining 13% that is CN-unmappable comprises largely regions in or around telomeres and centromeres (Extended Data Fig. 2b). To assess the burden of large SVs here, we applied two simplifying assumptions: (1) the rate of NAHR between any two regions in the genome is proportional to the number of position pairs with substantial homology (>500 bp with >96% homology) between these regions and (2) the density of non-NAHR-driven rearrangements is uniform across the genome, and hence the burden of non-NAHR breakends in a given region is proportional to its width. Both of these assertions

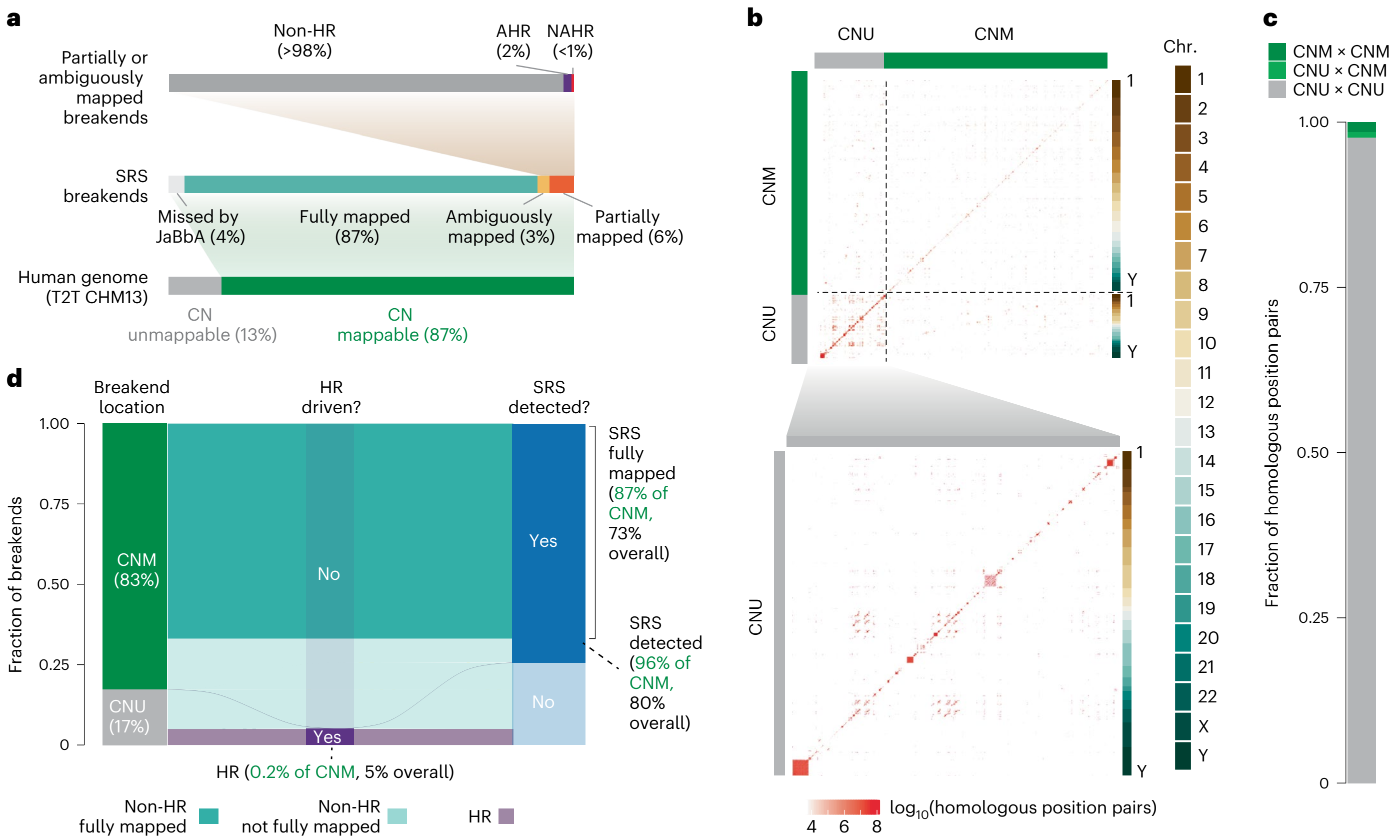

Fig. 5 | Extrapolating beyond the CN-mappable genome. a, Summary of SV breakends in the CN-mappable genome, including those predicted to be undetected by JaBbA v1 (Figs. 1 and 2). **b**, Heatmap showing the number of NAHR-eligible reference sequence position pairs, defined as pairs of reference positions >10 kb apart with ≥96% homology across 500 bp. The size of each bin in the genome-wide plot is 10 Mb (top subpanel) and 1 Mb (bottom subpanel, CN-unmappable zoom-in). CNU, CN-unmappable; CNM, CN-mappable. **c**, Fractional contribution of NAHR-eligible position pairs (see above) tallied across CN-unmappable and CN-mappable genome partitions. The number of position pairs with at least one site in a CN-unmappable region is expected to be ~100 times greater than the number of position pairs fully contained in CN-mappable regions. **d**, Alluvial plot showing the estimated fraction of SV breakends mapped by SRS across the genome. Colors stratify breakends on the basis of SRS mappability and homologous recombination versus other repair mechanisms.

hold true, to a first approximation, across the CN-mappable genome (Extended Data Fig. 10f,g).

We used the latest telomere-to-telomere build (T2T CHM13; ref. 38) to estimate the number of homologous position pairs outside CN-mappable regions (Fig. 5b). We found that CN-unmappable sequences harbored ~100-fold-greater homologous position pairs (2.7×10^{10} bp^2) than the CN-mappable portion of the T2T CHM13 genome build (2.8×10^8 bp^2) (Fig. 5c). This suggested that CN-unmappable regions harbor ~100 times as many NAHR SVs as CN-mappable regions. Integrating these measurements (Fig. 5a–c and Methods), we estimate that CN-mappable regions harbor 83% of all large SV cancer breakends, most of which are detected by SRS (Fig. 5d). Furthermore, even when CN-unmappable regions are taken into account, we estimate that homologous recombination contributes to a small proportion (~5%) of large cancer SV breakends (Fig. 5d).

Discussion

As cancer whole-genome SRS efforts scale and long-molecule genome profiling technologies mature, it is important to understand the limitations of SRS, particularly for the detection of chromosomal alterations. The conventional wisdom in the field has been that SRS misses most SVs owing to the prevalence of repeats in the human genome and the unclear contribution of NAHR to somatic structural genomic evolution[8,37,39,40]. Contrary to this prevailing intuition, we find that SRS detects and maps most large (>10-kb) somatic SV breakends in CN-mappable genomic regions. Intuitively, this is because most cancer chromosomal alterations are unbalanced and nonreciprocal (Fig. 1e), thus creating a CN footprint that SRS, when guided by mass balance approaches such as JaBbA v1, can reliably detect (Fig. 1b).

Our SRS analyses suggest that long-molecule technologies (for example, LRS and sLRS) will only modestly improve the detection of chromosomal breakends. We confirm this by jointly profiling the whole genomes of cancer samples and their matched normal samples with deep long-molecule sequencing (LRS or sLRS) and SRS. Given our findings, what additional insight into SVs can long-molecule technologies hope to offer? First, long molecules will enable the phasing of junctions to nearby somatic and germline variants. Resolution of the multi-junction haplotype structure at complex SVs may substantially inform their mechanistic interpretation and functional annotation, as in a recent study from our group[19]. Second, long molecules substantially increase the sensitivity for smaller (≤10-kb) somatic SVs, which were excluded from our analyses[41–43]. Future long-molecule studies will be needed to uncover the mutational processes and selective pressures driving the evolution of these smaller SV classes, including retrotransposition events.

Our study provides some of the most definitive evidence showing that NAHR drives a small proportion (<1%) of chromosomal alterations, at least in CN-mappable genomic regions. Our NAHR estimates in the remaining 13% (Fig. 5) of the genome assume that CN-mappable and CN-unmappable regions are subject to similar mutational processes. This assertion may require re-evaluation given recent studies investigating centromeric mutational processes[44].

Other settings where homologous recombination has been invoked, such as in the recombination of extrachromosomal DNA (ecDNA)[45,46], may similarly represent unique chromatin environments that are distinct from the remainder of the genome where homologous recombination rarely creates large SVs.

Practically, our study establishes JaBbA v1 as a state-of-the-art algorithm for cancer CN analysis, improving upon JaBbA v0.1 as well as classic 'change point'-based CN callers (Fig. 1b). The use of mass balance in the JaBbA model provides both superior performance in detecting somatic breakends and a lens into missing cancer SVs. Our study supports the use of JaBbA v1 and, more broadly, SRS in clinical cancer cytogenetics, where whole-genome SRS is poised to become routine in an era of plummeting sequencing costs[47,48].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01540-6.

References

1. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
2. de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
3. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
4. Hadi, K. et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* **183**, 197–210 (2020).
5. Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. *Nature Rev. Genet.* **23**, 298–314 (2021).
6. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
7. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
8. Malhotra, A. et al. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res.* **23**, 762–776 (2013).
9. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
10. Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. Detecting copy number variation with mated short reads. *Genome Res.* **20**, 1613–1622 (2010).
11. Greenman, C. D. et al. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* **22**, 346–361 (2012).
12. McPherson, A. W. et al. Remixt: clone-specific genomic structure estimation in cancer. *Genome Biol.* **18**, 140 (2017).
13. Aganezov, S. & Raphael, B. J. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *Genome Res.* **30**, 1274–1290 (2020).
14. Ross, E. M., Haase, K., Van Loo, P. & Markowetz, F. Allele-specific multi-sample copy number segmentation in ASCAT. *Bioinformatics* **37**, 1909–1911 (2021).
15. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
16. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
17. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
18. Wala, J. A. et al. Svaba: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
19. Setton, J. et al. Long-molecule scars of backup DNA repair in BRCA1- and BRCA2-deficient cancers. *Nature* **621**, 129–137 (2023).
20. Barthel, F. P. et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
21. Sieverling, L. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun.* **11**, 733 (2020).
22. Wilkie, A. O. M., Lamb, J., Harris, P. C., Finney, R. D. & Higgs, D. R. A truncated human chromosome 16 associated with α thalassaemia is stabilized by addition of telomeric repeat $(TTAGGG)_n$. *Nature* **346**, 868–871 (1990).
23. Morin, G. B. Recognition of a chromosome truncation site associated with α-thalassaemia by human telomerase. *Nature* **353**, 454–456 (1991).
24. Lange, T. D. et al. Structure and variability of human chromosome ends. *Mol. Cell. Biol.* **10**, 518–527 (1990).
25. Maciejowski, J. & de Lange, T. Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* **18**, 175–186 (2017).
26. Lovejoy, C. A. et al. Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of telomeres pathway. *PLoS Genet.* **8**, e1002772 (2012).
27. Zapatka, M. et al. The landscape of viral associations in human cancers. *Nat. Genet.* **52**, 320–330 (2020).
28. Cameron, D. L. et al. VIRUSBreakend: viral integration recognition using single breakends. *Bioinformatics* **37**, 3115–3119 (2021).
29. Symer, D. E. et al. Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers. *Genome Res.* **32**, 55–70 (2021).
30. Akagi, K. et al. Intratumoral heterogeneity and clonal evolution induced by HPV integration. *Cancer Discov.* **13**, 910–927 (2023).
31. Li, J. S. Z. et al. Chromosomal fragile site breakage by EBV-encoded EBNA1 at clustered repeats. *Nature* **616**, 504–509 (2023).
32. Sasaki, M., Lange, J. & Keeney, S. Genome destabilization by homologous recombination in the germ line. *Nat. Rev. Mol. Cell Biol.* **11**, 182–195 (2010).
33. Choate, K. A. et al. Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in *KRT10*. *Science* **330**, 94–97 (2010).
34. Cheung, V. G., Burdick, J. T., Hirschmann, D. & Morley, M. Polymorphic variation in human meiotic recombination. *Am. J. Hum. Genet.* **80**, 526–530 (2007).
35. Kidd, J. M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
36. Renkawitz, J., Lademann, C. A. & Jentsch, S. Mechanisms and principles of homology search during recombination. *Nat. Rev. Mol. Cell Biol.* **15**, 369–383 (2014).
37. Turner, D. J. et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.* **40**, 90–95 (2021).
38. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

39. Parks, M. M., Lawrence, C. E. & Raphael, B. J. Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biol.* **16**, 72 (2015).
40. Pascarella, G. et al. Recombination of repeat elements generates somatic complexity in human genomes. *Cell* **185**, 3025–3040 (2022).
41. Aganezov, S. et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
42. Nattestad, M. et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* **28**, 1126–1135 (2018).
43. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
44. Saayman, X., Graham, E., Nathan, W. J., Nussenzweig, A. & Esashi, F. Centromeres as universal hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. *Mol. Cell* **83**, 523–538 (2023).
45. Schimke, R. T. Gene amplification in cultured animal cells. *Cell* **37**, 705–713 (1984).
46. Rosswog, C. et al. Chromothripsis followed by circular recombination drives oncogene amplification in human cancer. *Nat. Genet.* **53**, 1673–1685 (2021).
47. Duncavage, E. J. et al. Genome sequencing as an alternative to cytogenetic analysis in myeloid cancers. *N. Engl. J. Med.* **384**, 924–935 (2021).
48. Almogy, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.29.493900 (2022).
49. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, btz041 (2019).
50. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
51. Smolka, M. et al. Comprehensive structural variant detection: from mosaic to population-level. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.04.487055 (2022).
52. Elrick, H. et al. Abstract LB080: SAVANA: a computational method to characterize structural variation in human cancer genomes using nanopore sequencing. *Cancer Res.* **83**, LB080 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.