Article

Long-molecule scars of backup DNA repair in BRCA1- and BRCA2-deficient cancers

https://doi.org/10.1038/s41586-023-06461-2

Received: 18 August 2021

Accepted: 20 July 2023

Published online: 16 August 2023

Open access

Check for updates

Jeremy Setton[1,13], Kevin Hadi[2,3,4,13], Zi-Ning Choo[2,3,4,13], Katherine S. Kuchin[2,3,5], Huasong Tian[2,3], Arnaud Da Cruz Paula[6], Joel Rosiene[2,3], Pier Selenica[6], Julie Behr[2,3,5], Xiaotong Yao[2,3,5], Aditya Deshpande[2,3,5], Michael Sigouros[7], Jyothi Manohar[7], Jones T. Nauseef[3,7,8,9], Juan-Miguel Mosquera[2,7,9], Olivier Elemento[7,9,10], Britta Weigelt[6], Nadeem Riaz[1], Jorge S. Reis-Filho[6], Simon N. Powell[1,14] ✉ & Marcin Imieliński[2,3,7,9,11,12,14] ✉

Homologous recombination (HR) deficiency is associated with DNA rearrangements and cytogenetic aberrations[1]. Paradoxically, the types of DNA rearrangements that are specifically associated with HR-deficient cancers only minimally affect chromosomal structure[2]. Here, to address this apparent contradiction, we combined genome-graph analysis of short-read whole-genome sequencing (WGS) profiles across thousands of tumours with deep linked-read WGS of 46 *BRCA1*- or *BRCA2*-mutant breast cancers. These data revealed a distinct class of HR-deficiency-enriched rearrangements called reciprocal pairs. Linked-read WGS showed that reciprocal pairs with identical rearrangement orientations gave rise to one of two distinct chromosomal outcomes, distinguishable only with long-molecule data. Whereas one (*cis*) outcome corresponded to the copying and pasting of a small segment to a distant site, a second (*trans*) outcome was a quasi-balanced translocation or multi-megabase inversion with substantial (10 kb) duplications at each junction. We propose an HR-independent replication-restart repair mechanism to explain the full spectrum of reciprocal pair outcomes. Linked-read WGS also identified single-strand annealing as a repair pathway that is specific to BRCA2 deficiency in human cancers. Integrating these features in a classifier improved discrimination between BRCA1- and BRCA2-deficient genomes. In conclusion, our data reveal classes of rearrangements that are specific to BRCA1 or BRCA2 deficiency as a source of cytogenetic aberrations in HR-deficient cells.

Cancer genomes provide a record of the genetic alterations acquired from DNA damage and DNA repair defects during normal cell development and carcinogenesis[3]. Genome-wide somatic alteration patterns in BRCA1-deficient (BRCA1d) and BRCA2-deficient (BRCA2d) cancers[2,4] are attributed to a deficiency in HR, a major pathway for the repair of double-strand breaks (DSBs) in human cells. Some of these mutational patterns could reflect specific error-prone mechanisms of DSB repair that cells use in the absence of HR[5]. Such mutational patterns can provide biomarkers of HR deficiency and help to identify clinically relevant therapeutic vulnerabilities[6,7].

Impaired DSB repair in HR-deficient (HRD) cells is thought to compromise structural genomic integrity, leading to characteristic cytogenetic alterations including radial chromosomes and chromosome bridges[1,8,9]. Confirming these cytogenetic observations, microarray and WGS studies have found loss of heterozygosity (LOH) and other megabase-scale patterns of allelic imbalance to be enriched among HRD cancers[8,10–13]. Such copy number alterations, however, are also found in HR-proficient (HRP) tumours and have not been linked to specific classes of structural variants (SVs). Paradoxically, the key genomic features that distinguish BRCA1d and BRCA2d from HRP tumours are single-nucleotide variants (SNVs), small deletions with microhomology, tandem duplications and simple deletions[2], all which have minimal effects on chromosomal structure. As a result, it is still poorly understood how aberrant DSB repair produces the associated cytogenetic phenotype in HR deficiency.

Developments in the analysis of cancer genomes allow for systematic annotation of complex SVs such as chromothripsis (chromosome shattering)[14], chromoplexy (balanced rearrangement chains)[15] and templated insertion chains (TICs)[16–18]. Previous WGS analyses of HR deficiency, however, have not considered this expanded SV taxonomy, either ignoring complex SVs[4] or treating them as a single 'clustered rearrangement' category[2]. They have also treated copy number and rearrangement independently, unlike more recently developed genome-graph algorithms that integrate these features under the principle of mass balance[18]. We reasoned that a genome-graph analysis

[1]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [3]New York Genome Center, 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, Weill Cornell Medicine, New York, NY, USA. [6]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [7]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [8]Division of Hematology and Medical Oncology, Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [9]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [10]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [11]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [12]Department of Pathology and Perlmutter Cancer Center, NYU Grossman School of Medicine, New York, NY, USA. [13]These authors contributed equally: Jeremy Setton, Kevin Hadi, Zi-Ning Choo. [14]These authors jointly supervised this work: Simon N. Powell, Marcin Imieliński. ✉e-mail: powells@mskcc.org; mski@mskilab.org

might uncover HR-deficiency-specific patterns of complex SVs, which could improve the classification of HRD tumours and provide mechanistic insights into their origin.

Genome-graph analysis of HRD tumours

To investigate the role of complex SVs in HRD cancers, we assembled a dataset of 979 predominantly primary (95%) cancer WGS profiles from four tumour types that are commonly associated with HR deficiency[19] (breast, ovarian, prostate and pancreatic cancer; referred to as BOPP, Methods and Extended Data Fig. 1). These included 24 and 36 cancers with biallelic inactivation of *BRCA1* and *BRCA2*, respectively, and 487 HRP tumours that lacked mono- or biallelic alterations in any HR-pathway gene (Extended Data Fig. 1, Supplementary Table 1, Methods and Supplementary Note 1). We then compared SV patterns between BRCA1d, BRCA2d and HRP tumours using methods that integrate copy number changes and rearrangements across genome graphs[18].

Analysing the burden of individual simple SV classes between BRCA1d, BRCA2d and HRP tumours, we confirmed the previously observed enrichment of short (1–100 kbp) SV duplications in BRCA1d cancers, and deletions in both BRCA1d and BRCA2d cancers (Extended Data Fig. 2a). Although BRCA1d and BRCA2d cancers had higher SV burdens than did HRP cancers (Extended Data Fig. 2b), we found no significant difference in the burden of simple translocations and inversions (Extended Data Fig. 2a), as has been previously noted[4].

We next asked whether HRD tumours were enriched in specific classes of complex SVs, and found no significant difference in the burden of seven previously characterized complex SV categories[18] in BRCA1d or BRCA2d relative to HRP tumour samples (Extended Data Fig. 2c). Contrary to the commonly held assumption that HRD cancers are exceptionally rearranged compared to HRP cancers, we found that they contained similar burdens of most SV classes, including complex SVs. TICs, however, were significantly enriched among both BRCA1d and BRCA2d relative to HRP tumours (Extended Data Fig. 2c). TICs arise through the copying and pasting of smaller (less than 10 kb) and genomically dispersed DNA segments in between larger (megabase-scale) segments[16].

Near-reciprocal SVs in HRD cancers

A classic reciprocal rearrangement (that is, balanced translocation or inversion) occurs without the loss or gain of genetic material and involves a pair of DNA junctions with break ends that adjoin the same break point (Fig. 1a, left). However, many rearrangements, including translocations and inversions, are near-reciprocal, with break ends that are nearby but not adjacent on the genome (Fig. 1a, middle). TICs[16] and chromoplexies[15] (chains of balanced rearrangements) are examples of complex SVs that are near-reciprocal.

Near-reciprocal SVs contain copy loss or gain of the intervening genomic region, which we call a gap segment (Fig. 1a, middle). The direction of copy loss versus gain at the gap segment is determined by its polarity, which by conservation of mass yields a copy gain when break ends join the gap segment (+ polarity) and copy loss when break ends join the flanking segments (− polarity). For a (+) gap segment, the identical locus topology and copy number profile may be equally consistent with a translocation or a simple templated insertion, in which a gap segment is copied to a distant locus and leaves the source locus unaltered (Fig. 1a, right).

Despite their enrichment in BRCA1d and BRCA2d cancers, TICs were still found in a substantial fraction (36%) of HRP tumours (Extended Data Fig. 2d). We hypothesized that a more comprehensive analysis of near-reciprocal junctions might yield uniquely HR-deficiency-specific patterns. Analysing clusters of near-reciprocal junctions linked by (+) and (−) polarity gaps (Methods and Extended Data Fig. 3a,b) in a training dataset (Extended Data Fig. 1) revealed simple paired patterns (for example, balanced translocations with short 6–7-bp (−) polarity gap segments), as well as more complex cyclic and non-cyclic reciprocal SV topologies comprising (+) and/or (−) polarity gap segments (Extended Data Fig. 3c,d).

Comparing near-reciprocal SV topologies across genotypes, we found that paired and cyclic patterns were most significantly enriched in BRCA1d (μ = 10.17 events per case, relative risk (RR) 4.72, $P = 3.6 \times 10^{-12}$, Wald test on gamma-Poisson regression) and BRCA2d (μ = 5.59, RR 4.73, $P = 5.7 \times 10^{-11}$) cancers relative to HRP cancers (μ = 1.13) (Extended Data Fig. 3e). We call these cyclic and paired patterns reciprocal pairs. Notably, we did not find genotype-specific differences in non-cyclic or higher-order reciprocal SVs comprising more complex templated insertion events and chromoplexies (Extended Data Fig. 3c–e).

Visualizing the gap segment lengths and polarities of reciprocal pairs revealed three distinct subpatterns that were specific to BRCA1 deficiency and/or BRCA2 deficiency (Fig. 1b). The first was an enrichment in BRCA1d tumours of reciprocal pairs with two 100-bp–100-kbp (+) polarity gap segments, which we call reciprocal duplications (rDups; Fig. 1c, left). The second was an enrichment in BRCA2d tumours of reciprocal pairs with two 1-bp–10-kbp (−) gap segments, which we call reciprocal deletions (rDels; Fig. 1d, left). The third was an enrichment in BRCA1d and BRCA2d cases of reciprocal pairs comprising 1-bp–100-kbp gap segments of opposite (+) and (−) polarity, which we call reciprocal deletion-duplications (rDelDups; Fig. 1e, left). Inspection of individual reciprocal pair loci (Fig. 1c–e, left) confirmed that these occurred in genomic regions that otherwise did not contain other rearrangements within a 1-Mbp vicinity. Analysis of these patterns in a validation dataset (Extended Data Fig. 1) confirmed the enrichment of rDups, rDels, and rDelDups in BRCA1, BRCA2 and HR deficiency, respectively (Fig. 1c–e, right and Supplementary Note 2).

Long-molecule WGS of HRD cancers

Identical rearrangement topologies can have very distinct chromosomal outcomes, or phases (Fig. 1a). We noted that the topology and copy number profile of reciprocal pairs were consistent with either of two outcomes: (1) the templated insertion of a small (around 1 bp–100 kbp) segment to an ectopic site; or (2) a larger quasi-balanced rearrangement such as a translocation or inversion (Extended Data Fig. 4a). We also observed that most reciprocal pairs comprised long-range (either interchromosomal or larger than 10-Mbp intrachromosomal) junctions (Fig. 1f). This indicated that, depending on the outcome, reciprocal pairs could have either a minimal or a major effect on chromosomal structure.

To resolve this aspect of reciprocal pairs, we performed linked-read (LR) and standard WGS on 46 tumours and matched normal samples that were originally found by clinical panel sequencing to have inherited or somatic mutations in *BRCA1* (27 cases) or *BRCA2* (19 cases; Extended Data Figs. 1 and 4b and Supplementary Note 3). LR WGS provided deep (median, 149×) genome-wide physical coverage of tumour and normal samples through barcoded short-read sequencing of long DNA molecules (median length, 24.4 kbp; Extended Data Fig. 4c). We reasoned that long molecules would help to resolve reciprocal pairs into phased somatic haplotypes and provide insight into the mechanistic origin and outcome of these SVs.

We identified 186 reciprocal pairs among BRCA1d and BRCA2d cases (μ = 5.17 per case). Comparison of standard and LR WGS profiles showed concordant rDel, rDup and rDelDup calls (83.5% overlap; Extended Data Fig. 4d), although LR WGS identified 29 additional reciprocal pairs. Confirming results from the BOPP short-read WGS dataset, BRCA1d tumours had higher burdens of rDups than did BRCA2d tumours ($P = 1.95 \times 10^{-4}$, RR = 49.50, Wald test on gamma-Poisson regression), and rDups were found in most (82%) BRCA1d tumours but in only one BRCA2d tumour. Similarly, rDels were present in most (71%) BRCA2d

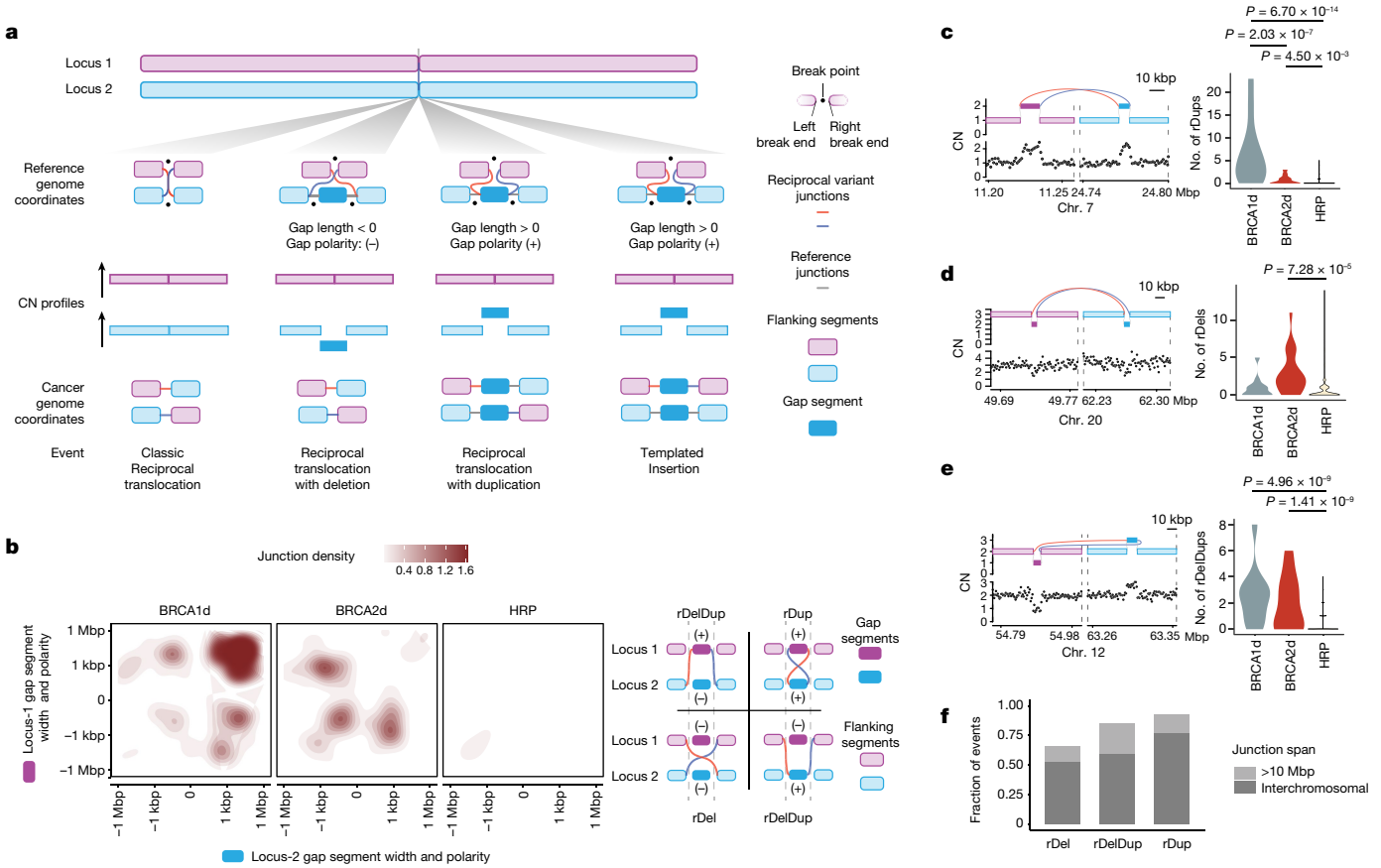


Fig. 1 | Reciprocal pairs are enriched in BRCA1d and BRCA2d tumours. **a**, Schematic of exact and near reciprocality, using translocations as an example. Exactly reciprocal junctions link break ends that adjoin the same break point (schematic on the right), giving rise to a balanced translocation. Near-reciprocal junctions are associated with a gap segment (dark blue) that is lost (middle left) or gained (right examples). The gap segment polarity refers to whether the adjoining junctions connect to the gap segment (+ polarity; right two examples) or to its adjacent segments (− polarity, middle left example). The polarity determines whether there is a copy gain (+) or loss (− polarity) of the gap segment. Both (−) and (+) gap segments can give rise to quasi-balanced translocations; however, (+) gap segments are also equally consistent with a templated insertion (right). CN, copy number. **b**, Gap segment lengths and polarities of three canonical reciprocal pair patterns (right) plotted across BRCA1d, BRCA2d or HRP cases (left). Density is calculated as a Gaussian kernel normalized by the number of BRCA1d ($n = 9$), BRCA2d ($n = 23$) or HRP ($n = 251$) cases in each plot. **c**–**e**, Examples of rDups (**c**), rDels (**d**) and rDelDups (**e**) with violin plots showing their relative burdens across 15 BRCA1d, 13 BRCA2d and 236 HRP samples, which are independent from the data in **b**. *P* values obtained by Wald test on a gamma-Poisson regression model. **f**, Distribution of junction spans associated with different classes of reciprocal pair SVs. Note that junction span is distinct from gap segment length; the former refers to the genomic distance between the two break ends belonging to a junction, whereas the latter refers to the distance between reciprocal break ends belonging to distinct junctions.

tumours but absent in all but four BRCA1d tumours ($P = 1.85 \times 10^{-7}$, RR = 11.67; Extended Data Fig. 4e).

To assess whether reciprocal pairs could be responsible for large-scale rearrangements in HRD cancers, we inferred their derivative chromosomal structure, or phase (Extended Data Fig. 4a). The specific goal of these analyses was to distinguish between *cis* (copy-paste, templated insertion) outcomes and *trans* (balanced translocation or inversion) outcomes on the basis of LR WGS alignment patterns (Fig. 2a and Extended Data Fig. 4a; Methods). After benchmarking phasing methods (Extended Data Fig. 5a,b and Supplementary Note 4), we analysed reciprocal pairs in our LR WGS data.

Phasing of 94 rDups in BRCA1d samples revealed a predominance (67/94; 71%) of *cis* phases (Fig. 2b,d), each resulting in the copying and pasting of a gap segment in between the tandem-duplicated gap segments of a distant locus. For example, given two loci ABC and DEF, this would yield an ABEBC haplotype containing the variant BE and EB junctions in tandem, and leaving the other DEF haplotype unrearranged (Fig. 2b). The remaining 29% of loci contained *trans* configurations (Fig. 2c,d), in which the BE and EB junctions were placed on discontiguous (for interchromosomal rDups) or distant (for intrachromosomal rDups) rearranged alleles. In these outcomes, two distinct derivative alleles ABEF and DEBC shared the duplicated B and E segments. This included balanced translocations with up to around 20 kbp of duplicated sequence at the junction (Fig. 2c).

Similarly, we used LR WGS to phase 46 rDelDups across 37 HRD cases (22 BRCA1d, 14 BRCA2d and one both BRCA1d and BRCA2d). As with rDups, we found both *cis* and *trans* phases at various loci (Extended Data Fig. 5c,d), although with a *trans* predominance of around 2:1 in BRCA1d tumours and a *cis* predominance of around 4:1 in BRCA2d tumours (Fig. 2d). Given ABC and DEF loci, *cis* rDelDups comprised a 'cut, copy and paste' outcome with an additional copy of E replacing B on a derivative AEC allele, with an unrearranged DEF locus containing the other E copy; by contrast, *trans* loci showed the same (1–241 kbp) E segment duplicated across two distinct DEC and AEF derivative loci. Finally, BRCA2d-tumour-specific rDels were predicted by short reads to give rise to strictly *trans* outcomes, which was confirmed by LR WGS (Fig. 2d and Extended Data Fig. 5e,f). These results show that *trans* reciprocal pairs are frequent among rDups, rDels and rDelDups and serve as a source of large-scale rearrangements in BRCA1d and BRCA2d tumours.

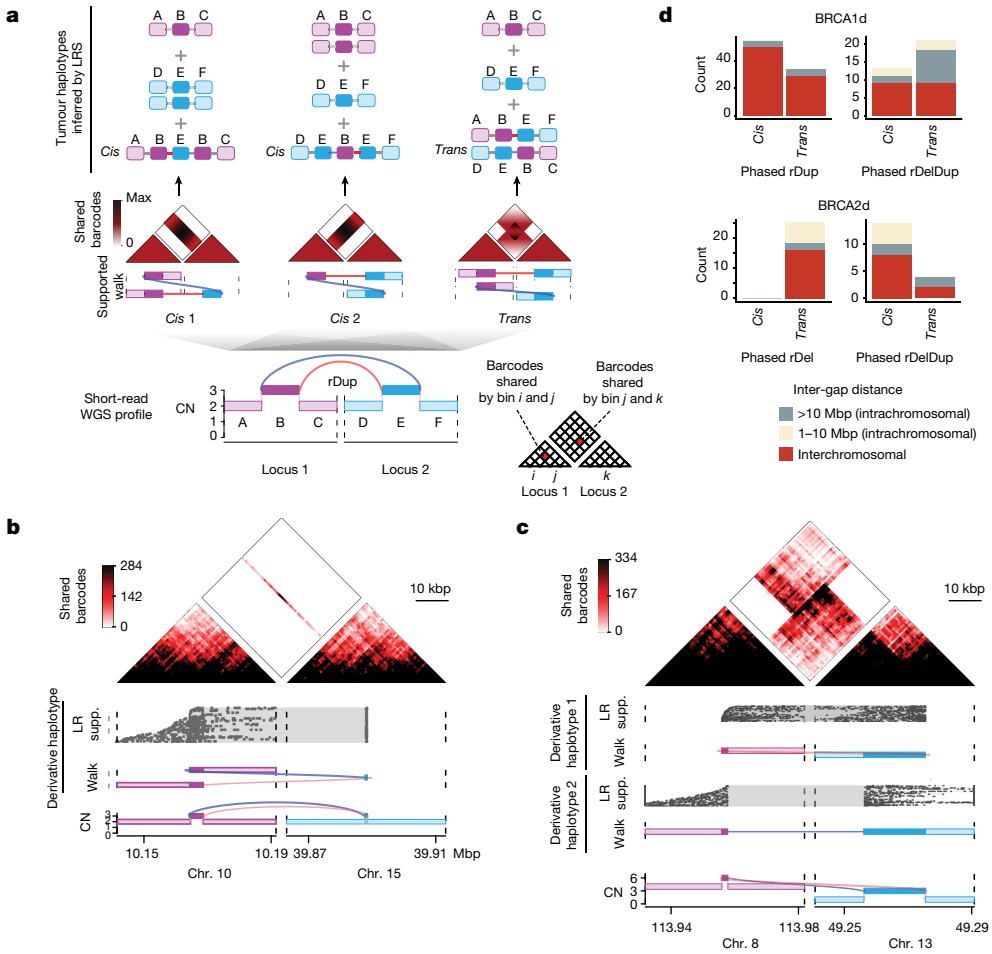


Fig. 2 | LR WGS reveals *cis* and *trans* phases for similar reciprocal pair topologies. a, Multiple phased allelic reconstructions are consistent with the rDup SV pattern observed in short-read WGS. Each derivative allele is represented as a 'walk' or oriented sequence of reference genomic segments. A phased reconstruction comprises a set of derivative alleles that together account for junction and segmental copy numbers observed in the short-read WGS genome graph. For a rDup, two of the three possible reconstructions (*cis* 1 and *cis* 2) place junctions adjacently, corresponding to the templated insertion of a distant segment between duplicated copies of a gap segment at the source locus. In the third case (*trans*), the junctions are located on discontiguous or distant alleles, consistent with a large translocation or inversion, respectively, in which each derivative allele contains a copy of both gap segments. Each reconstruction has a distinct LR WGS footprint, as visualized by a heat map in which the pixels represent LR WGS barcode sharing between rearranged loci (bottom right schematic, also applicable to **b** and **c**). LRS, linked-read sequencing. **b**,**c**, Two rDups, each phased by LR WGS, in *cis* (**b**) and in *trans* (**c**). **d**, Counts of LR WGS phased rDups, rDels and rDelDups from either BRCA1d (top; $n = 22$) or BRCA2d (bottom; $n = 14$) tumours. Reciprocal pairs are coloured by their junction span (1–10 Mbp, >10 Mbp or interchromosomal), see Fig. 1f for explanation.

Chromosomal effect of *trans* reciprocal pairs

Given the distinct chromosomal outcomes of *cis* versus *trans* reciprocal pairs, we looked for features that could distinguish these loci in short-read WGS and thus enable the study of reciprocal pair phase across a larger dataset. Deeper analysis and visualization of the phased structure of *cis* rDups (Fig. 3a, top) revealed that a short (50 bp–1 kbp) E segment was predominantly interleaved between two copies of a long (1–300 kbp) B segment in an ABEBC configuration. The trans rDups, however, comprised pairs of longer (1 kbp–300 kbp) B and E segments in ABEF and DEBC haplotypes (Fig. 3a, bottom).

Plotting these LR WGS phased *cis* and *trans* loci alongside unphased data from the short-read WGS BOPP dataset, oriented to the length of the longer gap segment in each pair (*y* axis), revealed two distinct rDup clusters. The first 'long–short' rDup cluster involved the linking of longer (1–100 kbp) and shorter (10 bp–1 kbp) gap segments and comprised exclusively *cis* events. By contrast, *trans* rDups were entirely contained in the second 'long–long' cluster (Fig. 3b). Similar length differences were found to differentiate *cis* and *trans* rDelDups (Extended Data Fig. 6a,b and Supplementary Note 5). These length differences made it possible to impute reciprocal pair phase in short-read WGS with reasonable accuracy (Extended Data Fig. 6c and Supplementary Note 6; Methods).

Because *trans* reciprocal pairs might engender long-range SVs (for example, balanced translocations and inversions), we predicted that *trans* but not *cis* events would be constrained in their chromosomal orientation. Specifically, *trans* reciprocal pairs can occur in one of two centromeric orientations: the first (type I) orientation generates only monocentric chromosomes, whereas the second (type II) generates one or more acentric derivatives (Fig. 3c). As acentric DNA fragments are prone to loss in subsequent cell divisions, a junction residing on an acentric fragment will be preferentially lost and the remaining junction will not be detected as a reciprocal pair. This will result in a type I bias for *trans* reciprocal pairs. Conversely, because templated insertions do not alter the chromosomal dosage of centromeres, *cis* loci should be agnostic to type I versus type II orientation.

Indeed, when we analysed our LR WGS phased data, we found that *trans* loci had a bias of more than 9:1 towards type I versus type II, whereas *cis* loci were equally likely to be in either the type I or the type II orientation ($P = 3.9 \times 10^{-6}$, odds ratio (OR) = 8.57, Fisher's exact test; Fig. 3d, left). We next analysed unphased reciprocal pairs in the short-read

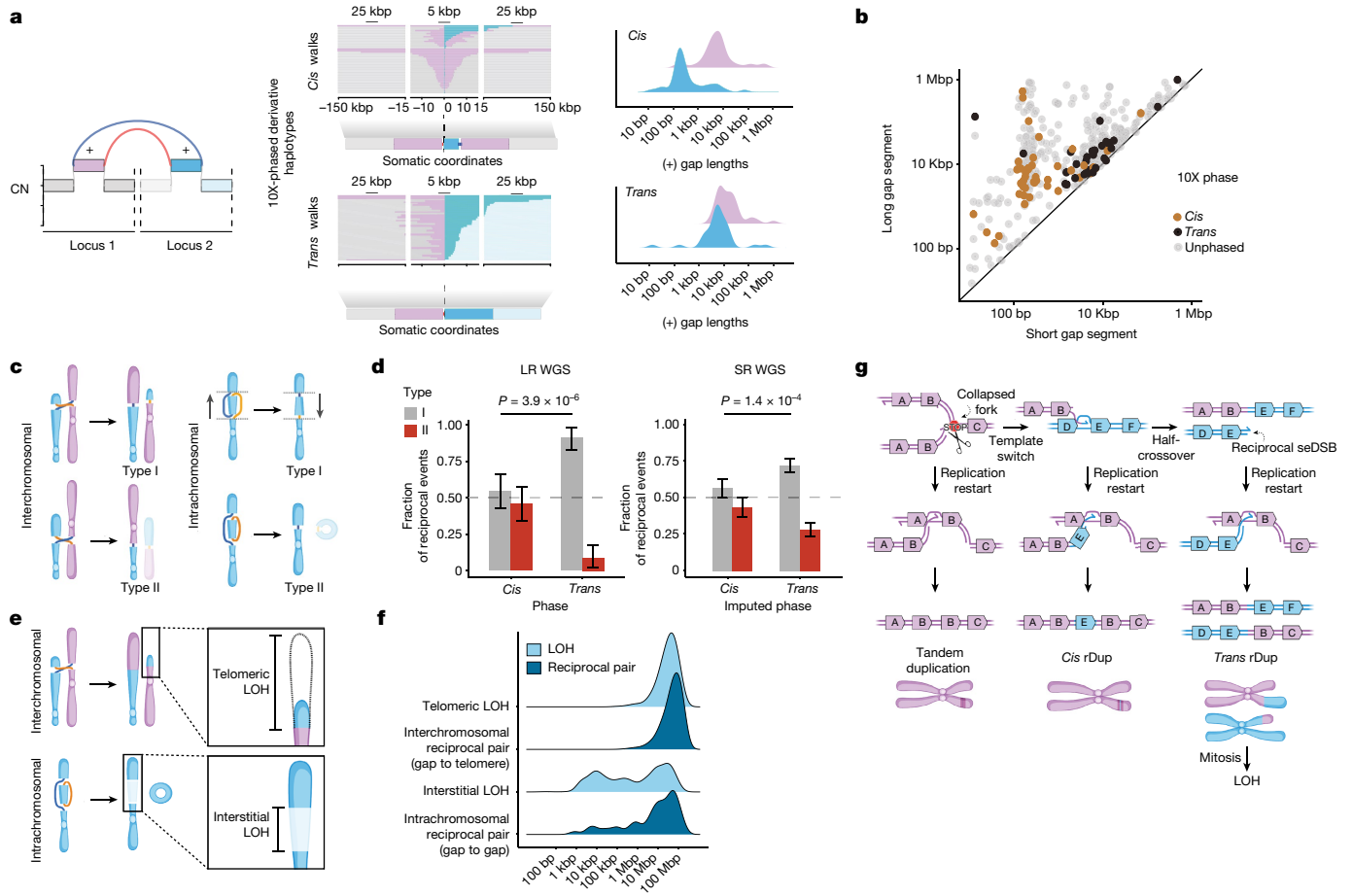


Fig. 3 | Aberrant replication-restart model links *trans* reciprocal pairs to megabase-scale chromosomal alterations. a, Derivative haplotypes (middle) for *cis* and *trans* rDups from BRCA1d ($n = 22$) and BRCA2d ($n = 14$) LR WGS profiles. Somatic chromosomal coordinates were harmonized so that 0 corresponds to the location of the first junction in the walk, with intervals coloured according to the genome-graph schematic (left). Density (right) shows the distribution of (+) gap segment lengths within *cis* and *trans* phased derivative chromosomes resulting from rDups. **b**, Scatter plot of longer and shorter (+) gap segment lengths for each rDup across BRCA1d ($n = 46$) and BRCA2d ($n = 50$) BOPP cases coloured according to LR WGS phase. **c**, Schematic illustrating two chromosomal outcomes of *trans* reciprocal pairs that either maintain centromere dosage (type I) or create an acentric derivative (type II). **d**, Fraction of type I and type II orientations among observed (left; LR WGS, $n = 131$ events) and imputed (right; short-read (SR) WGS, $n = 593$ events) *cis* and *trans* reciprocal pairs. *P* values obtained by two-sided Fisher's exact test. Error bars: 95% confidence interval on Bernoulli trial parameter. **e**, Schematic of LOH outcomes after a *trans* reciprocal pair. **f**, LOH length distributions plotted versus gap-to-telomere and gap-to-gap lengths for inter- and intrachromosomal reciprocal pairs, respectively, among BRCA1d ($n = 46$) and BRCA2d ($n = 50$) BOPP cases. **g**, Proposed replication-restart model linking tandem duplications, *cis* and *trans* rDups and LOH. Locus 1 (ABC) undergoes replication-fork collapse and may invade locus 2 (DEF), giving rise to *cis* and *trans* rDups. The latter might lead to LOH (after mis-segregation). seDSB, single-ended DSB. Variations of this model for rDelDups and rDels are shown in Extended Data Fig. 6g,h. Diagrams in **c**, **e**, **g** created with BioRender.com.

WGS BOPP dataset, imputing *cis* and *trans* phase on the basis of size and orientation (Fig. 3a,b and Extended Data Fig. 6a–c; Methods). We found the same direction of bias in these data ($P = 1.4 \times 10^{-4}$, OR = 1.96; Fig. 3d, right), despite having only imputed phases. In particular, these analyses showed that intrachromosomal *trans* reciprocal pairs yield megabase-scale inversions similarly constrained by centromere dosage.

We also found that the size distributions of interstitial and telomeric losses predicted to occur with *trans* reciprocal pairs mirrored the distribution of interstitial and telomeric LOH found in HRD cancers (Fig. 3e,f and Supplementary Note 7). Together, these results suggest that *trans* reciprocal pairs have large-scale chromosomal consequences, and thus can be implicated in cytogenetically visible aberrations that are classically associated with *BRCA1* and *BRCA2* inactivation.

Replication-restart model of reciprocal pairs

The observation of distinct *cis* and *trans* phases arising from nearly identical junction topologies (for example, rDups) suggested that they could represent distinct outcomes of a shared DNA-repair intermediate. Notably, most reciprocal pairs joined distant genomic locations (Figs. 1f and 2d) and had minimal sequence homology (Extended Data Fig. 6d), suggesting a possible homology-independent repair mechanism. The aberrant restart of a broken replication fork[20] has been implicated in the genesis of around 10–100-kbp tandem duplications in BRCA1d tumours[21]. Indeed, a key role of HR in human cells is in the repair of single-ended DSBs, which can arise at stalled replication forks[22–25].

To investigate the possibility of a shared mechanism between tandem duplications and reciprocal pairs, we analysed their distributions in our data. Notably, tandem duplications, rDups and rDelDups frequently co-occurred in BRCA1d cases (Extended Data Fig. 6e). In addition, the size distribution of the longer (+) gap segments in rDups or rDelDups, but not the shorter (+) gap segments in rDups, closely mirrored that of BRCA1d-tumour-specific tandem duplications (Extended Data Fig. 6f). Although the underlying mechanism will require experimental confirmation, we found that several simple extensions to the replication-restart model could explain the full

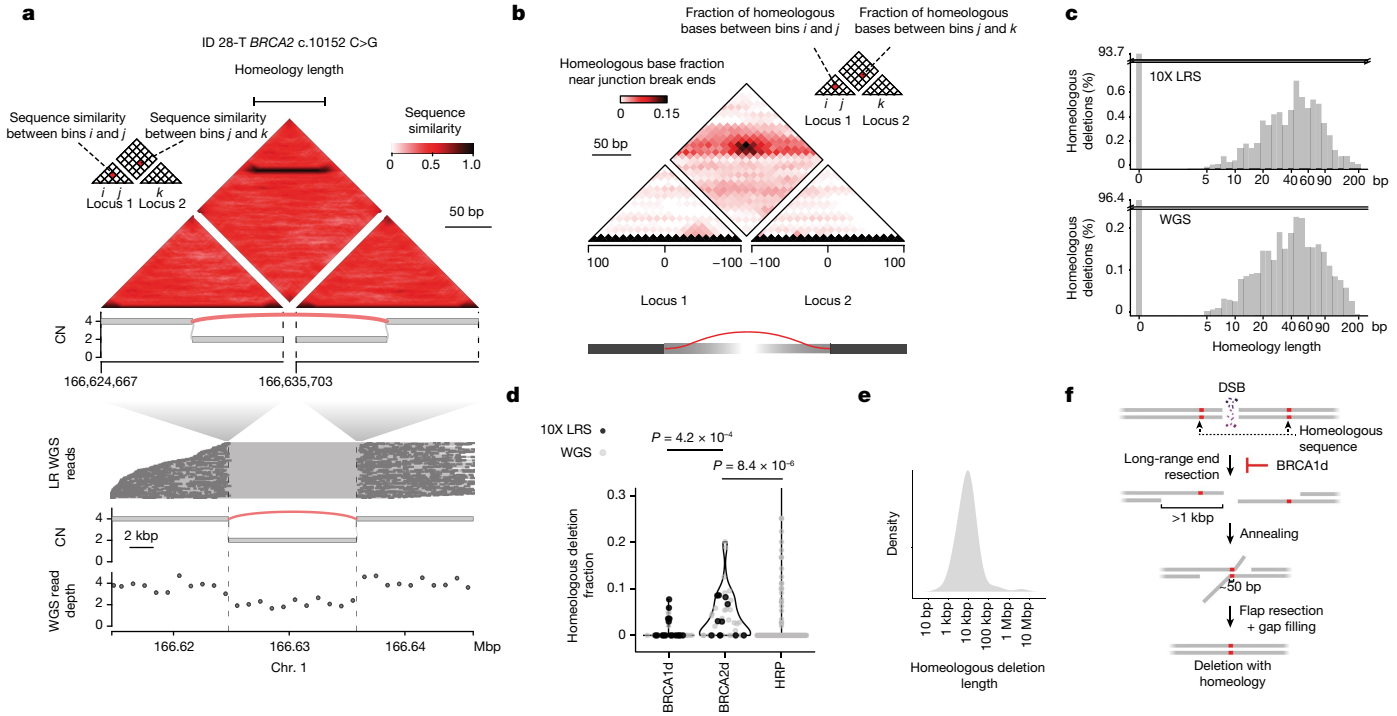


Fig. 4 | LR WGS uncovers footprints of SSA in BRCA2d genomes. a, Example of homeology (inexact sequence homology) across a deletion junction in a BRCA2d case detected by LR WGS. Heat map (top track) shows reference sequence similarity across pairs of 41-bp bins flanking junction break ends (Methods). Homeologous bin pairs are those with higher than 80% sequence similarity. The black line represents a continuous run of homeology, inferred through image analysis (Methods). Bottom track shows barcoded LR WGS read alignments supporting the homeologous deletion junction, in which each *y*-axis position represents a distinct LR barcode. **b**, Heat map showing counts of bases with homeology (sequence similarity ≥ 0.8) across all detected LR and standard WGS junctions (n = 1,240 junctions) across BRCA1d (n = 125), BRCA2d (n = 198) and HRP (n = 917) on a coordinate system defined around the location and orientation of each junction break end. Pixels are coloured according to the sequence similarity of the corresponding bin pair. **c**, Length of homeology (Methods) measured across all LR (top) (n = 34 samples, 724 junctions) and WGS (bottom) (n = 60 samples, 2,101 junctions) deletions. **d**, Fraction of homeologous junctions among simple deletions (more than 1 kbp) among BRCA1d (n = 43), BRCA2d (n = 46) and HRP (n = 374) tumours. *P* values obtained by two-sided Wald's test on gamma-Poisson regression. **e**, Junction span (distance between break ends) of homeologous LR WGS and short read WGS deletions in BRCA2d tumours (n = 46). **f**, Schematic of SSA mechanism, including dependence on long-range end resection and BRCA1 function. Diagram created with BioRender.com.

spectrum of reciprocal pair alterations, including rDups (Fig. 3g), rDelDups (Extended Data Fig. 6g) and rDels (Extended Data Fig. 6h).

Replication restart (for example, break-induced replication; BIR) mechanisms are known to be prone to template switching and half-crossovers, and have previously been linked to templated insertions in cancer cells[23,26–28]. A key decision point between template switching and half-crossover in BIR rests on the fate of the displacement loop after strand invasion[23,29]. Factors that promote displacement-loop disassembly or nascent strand displacement favour template switching with shorter replication tracts[30]. By contrast, factors that stabilize the displacement loop favour long-tract synthesis and half-crossover formation[30]. In our data, rDups and rDelDups with short ectopic (E segment) tracts were exclusively in *cis*, consistent with a template switch (Fig. 3a–d). In addition, for *cis* rDups associated with BRCA1 deficiency, the shorter E segment was always copied between two longer B segments (Fig. 3a). By contrast, only half-crossover outcomes (*trans* rDups and rDelDups) were observed alongside longer E (1–100 kbp) segments (Fig. 3a,b). In both outcomes, the distribution of rDup B-segment lengths mirrored the size distribution of tandem duplications in BRCA1d tumours (Fig. 3a and Extended Data Fig. 6f).

Together, our observations suggest that reciprocal SVs can be found in both *cis* and *trans* forms, with topological and tract-length characteristics previously associated with half-crossover and templated insertion outcomes of BIR[29,30]. We propose a provisional model invoking microhomology-mediated BIR (Fig. 3g and Extended Data Fig. 6g–h), extended from experimentally validated models established for BRCA1d-associated tandem duplications[20], that plausibly accounts for the full spectrum of reciprocal pairs.

Scars of backup repair in BRCA2d tumours

Given our LR WGS data, we posited that other scars of HR-deficiency-specific repair pathways could be detected using long-molecule mapping. Single-strand annealing (SSA) is a DSB repair pathway that involves the hybridization of approximately homologous (homeologous) repeat sequences flanking a DSB. Experimental model systems of HR deficiency have shown that SSA is active in BRCA2d but not in BRCA1d cancers[31–33]. SSA can tolerate as little as 80% sequence identity when annealing similar sequences deep inside resected break ends[33]; however, previous genome studies of HR deficiency have only analysed exact microhomology and have not examined inexact sequence identity.

To better assess the burden of SSA in LR WGS profiles, we developed and validated an algorithm (Methods and Supplementary Note 8) to detect runs of homeology, or 80% or higher sequence identity, near somatic break ends (Fig. 4a). This algorithm identified a peak of homeology around 50 bp (Fig. 4b), yielding 138 junctions with homeology greater than 10 bp across 46 LR WGS samples. Notably, most of these homeologous junctions were also detected with high efficiency in short-read WGS (Extended Data Fig. 7a,b and Supplementary Note 9), indicating that our analysis of homeology could be applied to the full short-read WGS BOPP dataset.

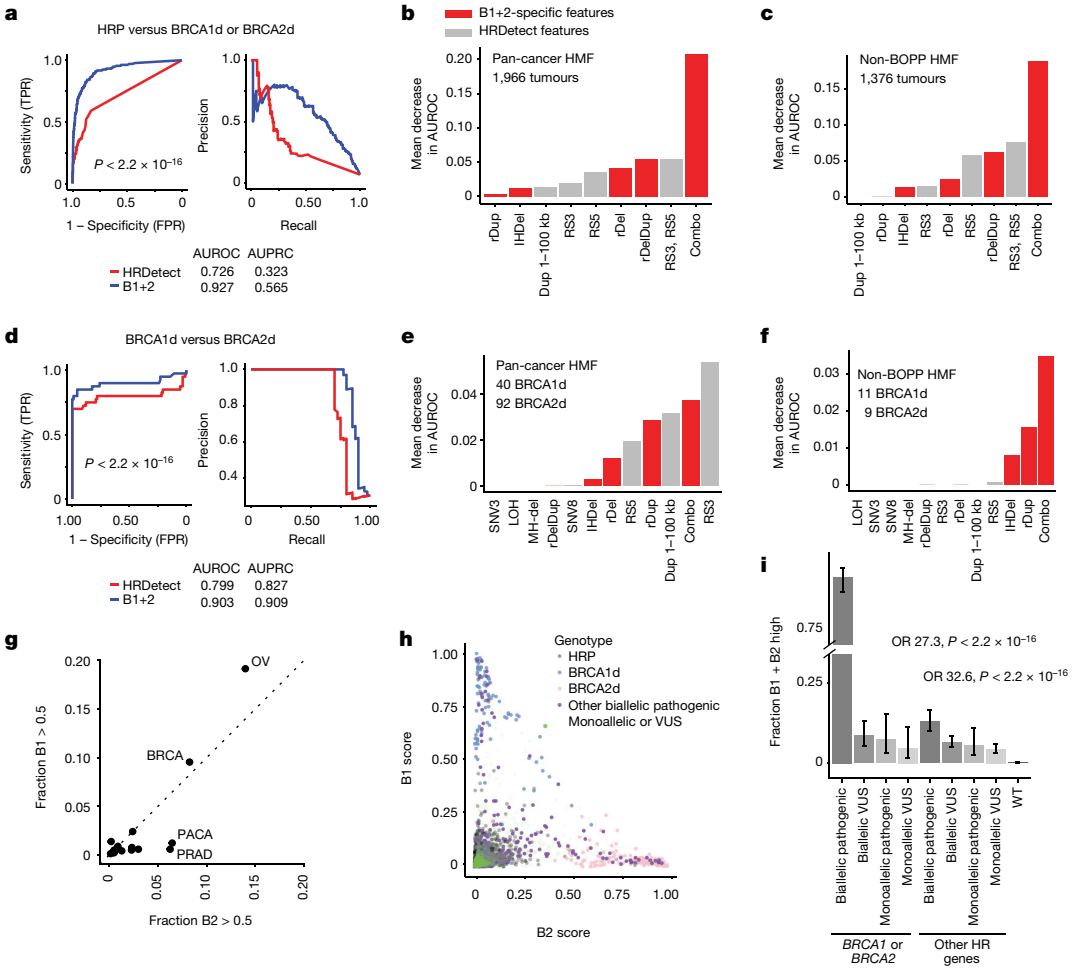


Fig. 5 | SV features distinguish between BRCA1 and BRCA2 deficiency. **a**, Receiver operating characteristic (ROC) curve and precision recall curve (PRC) comparing SV features highlighted in this study with those used in HRDetect for accurately classifying HR deficiency (either BRCA1d or BRCA2d). *P* values denote comparison of AUROC by two-tailed DeLong test. TPR, true positive rate. FPR, false positive rate. **b**,**c**, Importance of highlighted SV features in an independent pan-cancer WGS dataset (**b**) and its non-BOPP subset (**c**). 'Combo' refers to a combination of the B1+2 classifier-specific SV features highlighted in this study. IHDel, SV deletion with inexact homology. **d**, ROC curves assessing B1+2 and a random forest classifier using only the six HRDetect features in predicting BRCA1d versus BRCA2d status. **e**,**f**, Feature importance for BRCA1d versus BRCA2d classification in an independent pan-cancer WGS dataset (**e**) and its non-BOPP subset (**f**). See Extended Data Fig. 1 and Methods for training and testing dataset summary. MH-del, short deletion with microhomology. **g**, Frequency of HR deficiency, as defined by either B1 or B2 score > 0.5, among common cancer types, including samples excluded from training and testing for harbouring VUSs or monoallelic variants ($n = 7{,}918$). BRCA, breast adenocarcinoma; OV, ovarian cancer; PACA, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma. **h**, B1 versus B2 scores in the B1+2 classifier. **i**, Fraction of cases that are B1 or B2 positive (score > 0.5) for cases with a rare biallelic germline or somatic mutation in additional (that is, not *BRCA1* or *BRCA2*) HR-pathway genes ($n = 7{,}918$ tumours). Error bars show 95% confidence interval on the Bernoulli trial parameter. *P* values and odds ratios obtained with Fisher's exact test, without adjustment for multiple comparisons. WT, wild type.

Analysis of 583 tumour–normal pairs (BRCA1d, BRCA2d or HRP) revealed that 1,248 of 49,561 (2.5%) junctions were homeologous, with a distribution that mirrored LR WGS (Fig. 4c). Although the median homeology length among these junctions was 40 bp, we observed tracts as long as 128 bp. We next asked which classes of simple or complex variants contained homeologous junctions. Comparing distributions across genotypes revealed that BRCA2d tumours had a significantly higher burden and fraction of larger (more than 1 kbp) homeologous deletions relative to BRCA1d (RR = 3.93, $P = 4.2 \times 10^{-4}$, Wald test on gamma-Poisson regression) and HRP cancers (RR = 3.28, $P = 8.37 \times 10^{-6}$, Wald test on gamma-Poisson regression; Fig. 4d). Although we also observed homeologous break ends in other SV classes, the burden of these events did not correlate with HR-proficiency status (data not shown). Notably, the median size of homeologous deletions (around 10 kbp; Fig. 4e) was consistent with the length of end resection that is known to occur in BRCA2d cells[31,32], supporting the role of SSA as a backup repair pathway in human BRCA2d tumours (Fig. 4f).

SV features improve HRD subclassification

Having identified SV footprints of backup repair that are specific to BRCA1d and BRCA2d cancers, we next sought to understand whether these features could improve pan-cancer HR-deficiency classification. To assess the predictive value of the SV features highlighted in our study, we built a pan-cancer HR-deficiency classifier B1+2, which augments the six features used by HRDetect with the five highlighted in our study (1–100-kb tandem duplications, rDups, rDelDups, rDels and homeologous deletions; Methods and Supplementary Note 10). We trained the classifier on 62 BRCA1d, 64 BRCA2d and 2,536 HRP pan-cancer cases built from our BOPP and MSKCC datasets and additional publicly available samples (Extended Data Fig. 1 and Supplementary Note 11).

When applied to an independent test set of 1,966 BRCA1d, BRCA2d and HRP pan-cancer WGS profiles from the Hartwig Medical Foundation ('HMF' dataset), B1+2 showed a marginal improvement over HRDetect (B1+2 area under the receiver operator characteristic (AUROC) = 0.98, AUPRC = 0.87; HRDetect AUROC = 0.97, AUPRC = 0.86; Extended Data Fig. 8a). Indeed, as with HRDetect, classification was mainly driven by the fraction of indels with microhomology, yielding similar scores between the two algorithms (Extended Data Fig. 8b–d).

Given that HR deficiency is a disorder of DSB repair and altered genome structure, we next asked how well HR deficiency could be predicted solely on the basis of break-point-level structural genomic features. To assess this, we compared the performance of a random forest classifier trained using only the SV features in HRDetect (RS3 and RS5 signatures) and one trained with additional SV features specific to B1+2 (homeologous deletions, reciprocal pairs and 1–100-kbp simple duplications). We found substantially better performance ($P < 2.2 \times 10^{-16}$, DeLong test) with the B1+2-based SV classifier (AUROC = 0.93, AUPRC = 0.57, pan-cancer HMF) relative to the SV classifier based on HRDetect (AUROC = 0.73, AUPRC = 0.57; Fig. 5a). Although certain B1+2 classifier-specific SV features were individually relevant, the highest performance was observed when these features were used in combination (Fig. 5b,c). The performance improvement was most clearly attributable to the B1+2-specific SV features that recognized BRCA2d tumours (Extended Data Fig. 8e).

We next asked whether B1+2 could distinguish between BRCA1 and BRCA2 deficiency, which are distinct biological states, each with possibly distinct therapeutic vulnerabilities[34]. As HRDetect was not developed to address this task, we compared B1+2 to a random forest classifier trained on six HRDetect features (see above). We found that B1+2 substantially outperformed (AUROC = 0.90, AUPRC = 0.91) this HRDetect-like classifier (AUROC = 0.80, AUPRC = 0.83) in distinguishing BRCA1d from BRCA2d tumours ($P = 0.005$, DeLong test; Fig. 5d). B1+2 classifier-specific SV features were particularly important for making this distinction in non-BOPP cancers (Fig. 5e,f). We also performed similar comparisons to CHORD[4] (Extended Data Fig. 9a–c and Supplementary Note 12). As B1+2 outputs the separate probability of BRCA1d (B1 score) and BRCA2d (B2 score), we could analyse the probability of BRCA1d or BRCA2d in the tumours called HR-deficient by the classifier (B1+2 positive, B1 + B2 score > 0.5). This analysis confirmed that prostate and pancreatic cancer HR deficiency is significantly enriched in the BRCA2d phenotype relative to breast and ovarian cancer, in which BRCA1 and BRCA2 deficiency are equally likely[35,36] (Fig. 5g). Extending this analysis to non-BOPP samples, we found a lower rate (less than 5%) of HR deficiency, but with an increased bias toward BRCA2 deficiency (hepatocellular carcinoma and sarcoma; Extended Data Fig. 9d,e).

A major use of HR-deficiency genomic signatures is to uncover alternate mechanisms by which the HR pathway is inactivated and assess the pathogenicity of variants of uncertain significance (VUSs). Investigating B1+2 score distributions in cases that were excluded from our training and test data (Fig. 5h; including cases with monoallelic alterations and VUSs in *BRCA1* and *BRCA2* and/or other HR-pathway alterations; Supplementary Table 1), we found a significantly higher rate of B1+2 positivity across various strata of monoallelic and/or VUS cases (Fig. 5i and Supplementary Note 13), although this rate was substantially lower than that for cases with biallelic pathogenic alterations in *BRCA1* or *BRCA2* (95%). This included genes with distinct biases for BRCA1d (*BARD1* and *EME1*) versus BRCA2d (*PALB2* and *RAD51C*), consistent with their known roles in the HR pathway (Extended Data Fig. 9f). These results indicate that the B1+2 classifier could help to uncover and subclassify pathogenic alleles that are responsible for HR deficiency.

To further assess the relevance of classifier results, we investigated clinical outcomes for three cases with high B1+2 scores among 80 WGS cases profiled at Weill Cornell Medicine (Methods). All three cases with adequate follow-up data showed favourable responses to platinum chemotherapies and/or PARP inhibition (Supplementary Fig. 1 and Extended Data Fig. 10). This included a B2-high (B2 = 0.912) de novo case of metastatic neuroendocrine prostate cancer with an atypical (20.7 months) extracranial complete response to first-line platinum doublet (cisplatin–docetaxel) therapy and a second complete response to platinum rechallenge. The other two cases showed survival that exceeded the expectation (less than one year) for tumours of this histology and stage. Although all three cases also showed high HRDetect and CHORD scores, B1+2 provided extra certainty in distinguishing between BRCA1d and BRCA2d (Extended Data Fig. 10).

Discussion

Our LR WGS study provides one of the largest datasets so far of long-molecule whole-genome profiles in DNA-repair-deficient cancers. Long-molecule phasing allowed us to specifically link *trans* reciprocal pairs to large-scale chromosomal alterations. These results address a paradox in the field by providing a link between HR-deficiency-specific rearrangement patterns and megabase-scale cytogenetic phenotypes that are associated with HR deficiency[11,37,38].

Long-molecule data also reveal specific scars of backup repair pathways in HRD cells, including SSA, which has long been thought to help to maintain genome stability in BRCA2d cells[31]. Although SSA has been extensively studied using induced DSBs in synthetic plasmid reporter systems, it has not previously been shown to be relevant to human BRCA2d cancer genomes. Our data also provisionally extend the relevance of a second repair mechanism, homology-independent replication restart (Fig. 3g and Extended Data Fig. 6g,h), which has been previously implicated in BRCA1-deficiency-associated tandem duplications[20]. Extension of this mechanism to reciprocal pairs is most strongly supported by the existence of translocations and large inversions (*trans* rDups and rDelDups) with substantial (1–100 kbp) DNA duplication at one or both junctions. The substantial (more than 50 bp) duplications seen at *trans* rDups and rDelDups cannot be explained by simple end-joining but imply a replication-coupled repair process.

In HRP cells, BIR restarts replication when stalled and/or collapsed forks create single-ended DSBs[20,22,39]. A RAD51–RAD52-independent variant of BIR called microhomology-mediated BIR (MMBIR) can repair single-ended DSBs by invading nearby DNA duplexes in the absence of homology[26,28] and drive replication restart in HRD cells[23]. MMBIR intermediates are also exceptionally prone to template exchanges and crossovers, and thus provide the most plausible candidate for the genesis of reciprocal pairs as well as more complex SVs. In particular, factors that stabilize displacement loops and facilitate longer tracts of repair synthesis increase the likelihood of crossover products after BIR[20,22,40–42], consistent with our observation that *trans* reciprocal SVs contain larger duplications than do their *cis* counterparts (Fig. 3a and Extended Data Fig. 6f).

The ultimate criterion by which to judge HR-deficiency classifiers is their ability to predict response to genotoxic therapy. Assessment of this hypothesis beyond a few vignettes (Extended Data Fig. 10) will require large retrospective analyses of clinically annotated and WGS-profiled cases or prospective clinical trials with WGS-based classifiers as an end-point. Furthermore, the improved ability to distinguish between phenotypes of BRCA1 and BRCA2 deficiency, previously also addressed by CHORD[4], could inform future clinical trials that target BRCA1d- or BRCA2d-specific vulnerabilities[34]. As clinical WGS becomes cheaper and more practical, the routine implementation of approaches such as B1+2, which use more detailed features of BRCA1- and BRCA2-deficiency-specific SV patterns, might become an essential part of therapeutic decision-making.

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/s41586-023-06461-2.

1. Tutt, A. et al. Absence of Brca2 causes genome instability by chromosome breakage and loss associated with centrosome amplification. *Curr. Biol.* **9**, 1107–1110 (1999).
2. Davies, H. et al. HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
3. Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat. Commun.* **9**, 3292 (2018).
4. Nguyen, L., Martens, J. W. M., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **11**, 5584 (2020).
5. Setton, J., Reis-Filho, J. S. & Powell, S. N. Homologous recombination deficiency: how genomic signatures are generated. *Curr. Opin. Genet. Dev.* **66**, 93–100 (2021).
6. Robson, M. et al. Olaparib for metastatic breast cancer in patients with a germline *BRCA* mutation. *N. Engl. J. Med.* **377**, 523–533 (2017).
7. Abida, W. et al. Rucaparib in men with metastatic castration-resistant prostate cancer harboring a *BRCA1* or *BRCA2* gene alteration. *J. Clin. Oncol.* **38**, 3763–3772 (2020).
8. Yu, V. P. C. C. et al. Gross chromosomal rearrangements and genetic exchange between nonhomologous chromosomes following BRCA2 inactivation. *Genes Dev.* **14**, 1400–1406 (2000).
9. Ban, S. et al. Chromosomal instability in BRCA1- or BRCA2-defective human cancer cells detected by spontaneous micronucleus assay. *Mutat. Res.* **474**, 15–23 (2001).
10. Patel, K. J. et al. Involvement of Brca2 in DNA repair. *Mol. Cell* **1**, 347–357 (1998).
11. Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
12. Popova, T. et al. Ovarian cancers harboring inactivating mutations in *CDK12* display a distinct genomic instability pattern characterized by large tandem duplications. *Cancer Res.* **76**, 1882–1891 (2016).
13. Marquard, A. M. et al. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomark. Res.* **3**, 9 (2015).
14. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
15. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
16. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
17. Shale, C. et al. Unscrambling cancer genomes via integrated analysis of structural variation and copy number. *Cell Genomics* **22**, 100112 (2022).
18. Hadi, K. et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* **183**, 197–210 (2020).
19. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68–78 (2011).
20. Willis, N. A. et al. Mechanism of tandem duplication formation in *BRCA1*-mutant cells. *Nature* **551**, 590–595 (2017).
21. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
22. Li, S. et al. PIF1 helicase promotes break-induced replication in mammalian cells. *EMBO J.* **40**, e104509 (2021).
23. Wu, X. & Malkova, A. Break-induced replication mechanisms in yeast and mammals. *Curr. Opin. Genet. Dev.* **71**, 163–170 (2021).
24. Scully, R., Elango, R., Panday, A. & Willis, N. A. Recombination and restart at blocked replication forks. *Curr. Opin. Genet. Dev.* **71**, 154–162 (2021).
25. Feng, Y.-L. et al. DNA nicks induce mutational signatures associated with *BRCA1* deficiency. *Nat. Commun.* **13**, 4285 (2022).
26. Kockler, Z. W., Osia, B., Lee, R., Musmaker, K. & Malkova, A. Repair of DNA breaks by break-induced replication. *Annu. Rev. Biochem.* **90**, 165–191 (2021).
27. Osia, B. et al. Cancer cells are highly susceptible to accumulation of templated insertions linked to MMBIR. *Nucleic Acids Res.* **49**, 8714–8731 (2021).
28. Sakofsky, C. J. et al. Translesion polymerases drive microhomology-mediated break-induced replication leading to complex chromosomal rearrangements. *Mol. Cell* **60**, 860–872 (2015).
29. Vasan, S., Deem, A., Ramakrishnan, S., Argueso, J. L. & Malkova, A. Cascades of genetic instability resulting from compromised break-induced replication. *PLoS Genet.* **10**, e1004119 (2014).
30. Heyer, W.-D. Regulation of recombination and genomic maintenance. *Cold Spring Harb. Perspect. Biol.* **7**, a016501 (2015).
31. Tutt, A. et al. Mutation in *Brca2* stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. *EMBO J.* **20**, 4704–4716 (2001).
32. Stark, J. M., Pierce, A. J., Oh, J., Pastink, A. & Jasin, M. Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Mol. Cell. Biol.* **24**, 9305–9316 (2004).
33. Blasiak, J. Single-strand annealing in cancer. *Int. J. Mol. Sci.* **22**, 2167 (2021).
34. Samstein, R. M. et al. Mutations in *BRCA1* and *BRCA2* differentially affect the tumor microenvironment and response to checkpoint blockade immunotherapy. *Nat Cancer* **1**, 1188–1203 (2021).
35. Taylor, R. A. et al. Germline *BRCA2* mutations drive prostate cancers with distinct evolutionary trajectories. *Nat. Commun.* **8**, 13671 (2017).
36. Goggins, M. et al. Germline *BRCA2* gene mutations in patients with apparently sporadic pancreatic carcinomas. *Cancer Res.* **56**, 5360–5364 (1996).
37. Birkbak, N. J. et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).
38. Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with *BRCA1/2* inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
39. Costantino, L. et al. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**, 88–91 (2014).
40. Elliott, B., Richardson, C., Winderbaum, J., Nickoloff, J. A. & Jasin, M. Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.* **18**, 93–101 (1998).
41. Neuwirth, E. A. H., Honma, M. & Grosovsky, A. J. Interchromosomal crossover in human cells is associated with long gene conversion tracts. *Mol. Cell. Biol.* **27**, 5261–5274 (2007).
42. Panday, A. et al. FANCM regulates repair pathway choice at stalled replication forks. *Mol. Cell* **81**, 2428–2444 (2021).

Methods

Pan-cancer WGS data sources

GrCh37/hg19 BAM alignments for 2,489 primary tumour and matched normal whole-genome sequencing data were obtained as previously described[18]. In brief, 989 tumour–normal (T/N) pairs were obtained from The Cancer Genome Atlas (TCGA) Research Network (Genomic Data Commons at https://portal.gdc.cancer.gov/, accession: phs000178.v11.p8). Additional WGS data were obtained for 874 T/N pairs from the International Cancer Genome Consortium (ICGC) from multiple studies publicly available through the European Genome-phenome Archive (EGA; https://ega-archive.org). These cohorts include: 124 breast cancers[21] (EGA: EGAS00001001178), 179 melanomas[43] (EGA: EGAS00001001552), 49 lung adenocarcinomas[44] (EGA: EGAS00001002801), 422 oesophageal adenocarcinomas[45] (EGA: EGAD00001004417) and 100 malignant lymphomas (EGA: EGAD00001002123).

Additional BAMs for 121 T/N pairs from a pan-cancer cohort obtained as part of a New York City-based multi-institution collaborative research effort comprising the Memorial Sloan Kettering Cancer Center (MSKCC), New York University, Stony Brook University Hospital, Lenox Hill, Northwell Health, Columbia University, Montefiore, and Cornell, and led by the New York Genome Center, were included here and were previously described[18]. Study approval was obtained through a central institutional review board (IRB), Biomedical Research Alliance of New York, and by local IRBs, including Stony Brook University and Northwell Health. In addition, 55 prostate cancers that were previously published were obtained through dbGaP with accession phs000447.v1.p1 (ref. 15). BAMs for 80 T/N pairs were obtained from a collaborative precision oncology effort between the Weill Cornell Englander Institute for Precision Medicine (EIPM) and the New York Genome Center. This study was approved by an institutional review board (WCM IRB no. 1305013903). A total of 340 T/N pairs across 80 cases across longitudinally or spatially distinct biopsies from Barrett's oesophagus tumours were obtained as part of a previous study[46].

Call sets were obtained from 1,484 additional T/N pairs contributing additional primary tumour whole genomes from the Pan-Cancer Analysis of Whole Genomes Consortium[47] (Extended Data Fig. 1, 'PCAWG' dataset, https://dcc.icgc.org/pcawg) and 3,957 T/N pairs from metastatic whole genomes from the Hartwig Medical Foundation (HMF, https://www.hartwigmedicalfoundation.nl/), which included germline, somatic SNV or indel, and somatic SV calls[48] (Extended Data Fig. 1: 'HMF' dataset).

MSKCC cohort

LR WGS and short-read WGS were performed on a cohort of 46 cases biopsied for ductal carcinomas of breast and found to have *BRCA1* ($n = 28$) or *BRCA2* ($n = 18$) mutations on clinical panel sequencing. These cases were collected under informed consent as part of a prospective biospecimen research protocol at the Memorial Sloan Kettering Cancer Center (MSKCC, MSKCC IRB no. 16–675). T/N pairs were profiled with Illumina short-read WGS and LR WGS (see below for protocol details). Raw sequencing data from these experiments have been made available (see 'Data availability' section; Extended Data Fig. 1: 'MSKCC' dataset).

Pipelines

Harmonized variant calling was performed on 2,489 T/N BAM file pairs by adapting previously described pipelines[18]. Additional details are provided below.

SV calling

In brief, genome-wide, 200-bp binned tumour and normal read depth was calculated from alignments and corrected for GC and mappability biases (https://github.com/mskilab-org/fragCounter). Somatic SV calls were obtained with SvAbA[49] and filtered using a panel-of-normals (PON) comprising all germline SVs detected across 2,489 T/N pairs. Any somatic SV found within 500 bp of a junction within the germline SV PON with matching orientations was discarded. PCAWG consensus SVs and 200-bp binned tumour and normal read depths were obtained from PCAWG SV release 1.6 and the PCAWG data coordination centre.

HMF SV data were obtained from the Hartwig Medical Foundation through a data sharing agreement[48]. In brief, junction calls from GRIDSS[50] and 1-kbp tumour/normal coverage ratios[17] corrected for GC content were obtained for 3,957 T/N pairs.

Genome graph analysis

High-confidence junctions, binned tumour-normal read depth ratios, and purity and ploidy estimates (see below) were used to perform junction balance analysis (JaBbA; github.com/mskilab-org/JaBbA) and generate balanced genome graphs for 7,918 and 46 cases in the pan-cancer WGS and MSKCC datasets, respectively. For a detailed treatment of the formulation behind JaBbA, see a previous report[18]. Heterozygous germline single-nucleotide polymorphism (SNP) data were used to infer allelic copy number after total copy number inference was performed genome-wide as described[18].

Purity and ploidy estimation

Across Hadi dataset cases, tumour purity and ploidy were estimated using ASCAT[51]. For the 46 cases from the MSKCC cohort, a manual review of purity and ploidy was conducted to enhance downstream genotyping accuracy; ultimately, alternative manual estimates of purity and ploidy from MSKCC were chosen for 4 out of 46 cases. For PCAWG and HMF datasets, purity and ploidy estimates were obtained from the respective PCAWG (https://dcc.icgc.org/releases/PCAWG/) and HMF (https://www.hartwigmedicalfoundation.nl/en/database/) portals[48].

LR WGS SV calling

For the LR WGS profiles generated from the MSKCC cohort of 46 cases, junctions called using LinkedSV[52] were merged with SvAbA junctions called on the corresponding short-read WGS for each case. These were then input into JaBbA using short-read coverage profiles to generate genome graphs. Merging was performed using the gGnome R package (https://github.com/mskilab-org/gGnome) to determine junctions that were uniquely detected by LR WGS (LinkedSV).

Analysis of gap segment topology

Gap segments were defined as short genomic segments joining reference-consecutive break ends, each belonging to distinct junctions and occurring on opposite strands. Each gap segment was additionally associated with a polarity (+ or −) based on the topology of junctions around the gap segment; (+) polarity corresponded to a gap segment with junctions directly attached to it, (−) polarity corresponded to a gap segment with junctions attached to the two segments to the left and right of the gap segment on the reference. The length threshold to define gap segments was visually chosen as 1 Mbp after inspection of a density plot of segments lengths across gap segment candidates satisfying the above topological criteria. This threshold was confirmed through the application of a background model, in which the gap segment candidate length distribution in each sample was fit with an exponential distribution and each gap segment candidate was assigned a *P* value according to the left tail of the exponential cumulative distribution function. Short (less than 1 Mbp) gap segment candidates were found to significantly deviate from expectation (false discovery rate (FDR) < 0.10) under this model.

Applying this definition, gap segments with shared junctions were next clustered together (applying 'eclusters' gGraph method in the gGnome R package) to identify and classify reciprocal clusters. Reciprocal clusters in which every junction was connected to two gap segments was labelled as 'cyclic'. Reciprocal clusters were annotated with the number of cluster-associated junctions and gap segments. A reciprocal

pair (rPair) is a special case of a cyclic reciprocal cluster that contains two gap segments of either orientation. rDups, rDels and rDelDups each contain two (+) gap segments, two (−) gap segments and one (+) and one (−) gap segment, respectively.

Annotating known SV events

Classes of previously described[18] simple and complex SV were annotated in balanced genome graphs derived by JaBbA for both the pan-cancer WGS (n = 7,918) and MSKCC datasets (n = 46). The following simple events were annotated within each graph: deletions, duplications, translocations, inversions and inverted duplications. The following complex events were annotated: breakage–fusion–bridge cycles, double minutes, tyfonas, chromoplexy, chromothripsis and TICs. Implementation of each event classifier can be found in the 'events' function in the gGnome R package.

Variant calling and genotyping

For the 2,489 'Hadi' dataset WGS T/N pairs (Extended Data Fig. 1), somatic mutation calls were generated with Strelka1 for SNVs and indels. Germline mutation calls were obtained with Strelka2 run on alignments from blood or adjacent normal samples and filtered to remove common variants above a population allele frequency of 1% (ExAC population: ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/subsets/). SNVs and indels were filtered through a universal genome-wide mask for hg19 (https://github.com/lh3/CHM-eval) to remove artefacts due to low mappability, as described before[53]. All germline and somatic SNVs and indels were annotated with ClinVar status (ftp.ncbi.nlm.nih.gov:/pub/clinvar/vcf_GRCh37; database date 2022-07-30). The impacts of protein-coding SNVs and indels were also annotated through SnpEff (GRCh37.75 database). SNVs and indels were considered pathogenic if annotated as 'pathogenic' or 'likely_pathogenic' through ClinVar CLNSG or if their SnpEff annotation fell within the following classes: 'frameshift variant', 'start_lost', 'stop_gained', 'stop_lost', 'splice_acceptor_variant' or 'splice_donor_variant'. ClinVar annotation took precedence over SnpEff.

LOH was determined by allele-specific copy number (CN) using allele counts across germline heterozygous SNP sites. Specifically, LOH was called in regions in which minor allele CN = 0 and major allele CN > 0, using allelic copy number as inferred from short-read sequencing data (see 'Junction balance analysis'). Similarly, homozygou deletions (homdels) were called in regions in which total copy number (sum of major and minor allele CN) = 0.

Genotype was determined across samples for 48 HR-related genes (Supplementary Table 1), including *BRCA1*, *BRCA2*, *PALB2* and *RAD51C*. Eleven of these genes were highlighted in a previous study[54]. Biallelic loss was called for genes if they contained any of the following: (1) two or more germline and/or somatic pathogenic mutations (including SNVs, indels and SVs); (2) one germline or somatic pathogenic mutation along with LOH; or (3) a homozygous deletion. Within the MSKCC cohort, 22 cases were found to have biallelic loss of *BRCA1*, 14 cases were found with biallelic loss of *BRCA2*, and one case was found with biallelic loss of both *BRCA1* and *BRCA2*.

For PCAWG dataset cases, somatic SNV and indel calls were obtained from ICGC (2016 data freeze), and annotated driver mutations were obtained from the PCAWG consortium[47]. HMF provided the following for cases in their dataset: germline SNVs and indels (through GATK HaplotypeCaller), somatic SNVs and indels (through Strelka1 and annotated by SnpEff).

Short-read WGS

Short-read WGS for the 46 MSKCC donors was performed at the New York Genome Center to a target of 80× tumour and 40× normal coverage. Library preparation from genomic DNA for these new cases was performed using the NEBNext Ultra II End Repair/dA-Tailing Module, NEBNext Ultra Ligation Module (New England Biolabs) and KAPA Dual-Indexed Adapter Kit (Roche) following the manufacturers' protocols. Quality control was performed on the finished libraries with the Agilent 2100 Bioanalyzer on the High Sensitive DNA Chip platform (Agilent Technologies) and KAPA Library Quantification Kit (Roche). Quality control determined that libraries contained an average peak height (fragment size) of at least 400 bp. Libraries were sequenced on an Illumina NovaSeq 6000 System (Illumina) to generate paired-end 2 × 150-bp reads. Reads were aligned to the GRCh37/hg19 reference using Burrows–Wheeler aligner software[55], bwa mem, 0.7.10-r789). Read post-processing was done in accordance with best practices for post-alignment data processing with Picard tools (https://broadinstitute.github.io/picard/) to mark duplicates, the GATK (v.2.7.4) (https://gatk.broadinstitute.org/hc/en-us) IndelRealigner module and GATK base quality recalibration.

LR WGS

Each of the 46 *BRCA1*- or *BRCA2*-mutant cases in the MSKCC cohort was subjected to additional LR WGS. High-molecular-weight (HMW) genomic DNA (gDNA) was extracted using a Qiagen MagAttract HMW DNA Kit (Qiagen) according to the suggested protocol. In brief, approximately 1–2 million cells were obtained from each frozen tissue sample and lysed, and HMW gDNA was captured by magnetic particles. Then the magnetic particles with HMW gDNA were washed in wash buffer and eluted in EB Buffer (10 mM Tris-HCl, pH 8.5). The HMW gDNA had a mode length of 50 kbp and max length 200 kbp, as estimated on a separate 75-V pulse-field gel electrophoresis using a BluePippin 5–430-kbp protocol (Sage Science). LR WGS library preparation was performed using a Chromium Genome Library Kit v2 (10X Genomics) following the Chromium Genome Reagent Kits v2 User Guide. In brief, 1 ng of extracted HMW gDNA was used to prepare a library, with an average fragment length of 625 bp (ranging from 300 to 2,000 bp, measured with the Agilent Bioanalyzer High Sensitivity DNA Chip). Quality control for the finished libraries was performed as above for the general WGS library preparation. The prepared libraries were sequenced on an Illumina NovaSeq 6000 Sequencing System (Illumina) with an S4 flow cell, to an average read depth of about 33×. All linked reads were aligned to GRCh37/hg19 with the EMerAld aligner (v.0.6.2)[56]. Germline haplotypes were obtained from Strelka2 germline SNV calls processed using HapCut2 (https://github.com/vibansal/HapCUT2; ref. 57).

Phasing rearranged haplotypes with LR WGS

Our specific goal in somatic phasing was to distinguish SVs that placed both reciprocal junctions on the same molecule (*cis*) from those that placed junctions at distant loci (*trans*, including distinct derivative chromosomes) in the cancer genome (Extended Data Fig. 3a). Somatic phasing is distinct from parental phasing, which determines whether reciprocal break ends arose on the same or distinct parental homologue. Namely, break ends that arise on the same parental homologue (germline *cis* phase) can end up on distinct derivative chromosomes (somatic *trans* phase).

We approached somatic phasing by assessing LR WGS molecule support for derivative rearranged haplotypes. Derivative rearranged haplotypes can be deconvolved from junction-balanced genome graphs as walks[18]. Walks were derived from JaBbA graphs on the MSKCC cohort for which both LR and short-read WGS were available using the 'walks' gGraph method in the the gGnome package (https://github.com/mskilab-org/gGnome). Barcoded reads were matched against each possible walk within a 100-kbp window of the junctions to be phased (gGnome score.walks function). The walk (or set of walks) that carried the largest number of junction-supporting barcodes was considered the likeliest haplotype explaining the rearrangement.

Specifically with respect to the rDup and rDelDup patterns, two sets of possible derivative haplotypes exist: *cis* and *trans*. *Cis* haplotypes for rDup or rDelDup patterns are walks that contain both involved rearrangements consecutively, or, in other words, contain a single segment that is flanked by both junctions. *Trans* haplotypes are two separate

walks that each contain one of the two rDup or rDelDup rearrangements and would have to exist simultaneously, and thus are considered as a single set of walks. Junction-supporting LR barcodes were counted for each possible walk across every rDup or rDelDup instance in the MSKCC cohort. To assess whether the rDup or rDelDup patterns existed with the junctions in *cis* or *trans*, walk-supporting LR barcode counts were compared among the individual *cis* walks and the *trans* walks summed together. The walk (or set of walks in the *trans* case) was taken to be the derivative haplotype underlying the rDup or rDelDup pattern. Haplotypes were also validated by visually assessing the barcode sharing patterns for each rDup or rDelDup present in the dataset to confirm the haplotype as labelled by this heuristic.

Imputing short-read-sequencing reciprocal pair haplotypes

To impute the haplotype phase of reciprocal pairs identified by short-read WGS, we applied a threshold of 3.5 to the $\log_{10}$ gap length. Specifically, for rDup reciprocal pairs, the imputed haplotype phase was *cis* if the minimum of the two $\log_{10}$ gap lengths was less than 3.5, and *trans* otherwise. For rDelDup pairs, the imputed haplotype phase was *cis* if the $\log_{10}$ length of the (+)-polarity gap was less than 3.5, and *trans* otherwise. The imputed phase of all rDel pairs was *trans* because this is the only phase possible given the junction topology.

Sequence homeology

'Homeology' refers to approximate (higher than 80%) similarity between a pair of genomic sequences. To assess sequences at junction-associated break ends, we applied a sliding bin approach. For every position within a 200-bp window around each break end pair, a 41-bp bin centred at the base was queried for the corresponding hg19 reference sequence. All pairs of 41-bp bins within each junction-associated 200-bp window were then aligned to one another to construct a 200-by-200 matrix of Levenshtein edit distances. The distance matrix was the converted to a similarity matrix (Fig. 4a heat map) in which each entry *ij* indicates the sequence similarity, calculated as (1 − Levenshtein edit distance)/41, between a pair of 41-mers corresponding to bins *i* and *j* in the junction-associated window. The matrix was then converted to a binary bitmap image in which each pixel denoted sequence similarity of >0.8. Connected components of pixels in the image were annotated with the Pearson's correlation of the associated pixel indices, which was used as a measure of linearity of the pattern. Each junction was then annotated with the size (in pixels) of the largest connected component with a linearity of at least $r^2 > 0.9$. This value thus represents the longest contiguous stretch of bases with at least 80% sequence similarity. This procedure was run using the 'homeology' function in the GxG R package (see 'Code availability').

Discordant and split reads supporting junctions with homeology were realigned to hg19 using bwa mem (implemented using an R wrapper in the package RSeqLib), to obtain uniform mapping quality scores for those cases containing junction homeology within the in-house dataset in which alignments were present. Reference mappability was determined using two orthogonal means. In the first, sliding 150-mers stepping by one base were queried across hg19 and aligned to the full reference using bwa mem to determine mapping quality scores. Average mapping quality was determined for each base for hg19. The second method used GEM mappability score with a sliding 150-mer across hg19 as described previously[58].

Mutational signatures

Mutational signatures were derived from the signature.tools.lib R package suite for implementing the HRDetect algorithm[2]. In brief, SNV signatures were deconvolved using the known signature weights from COSMIC SNV signature version 2 (https://cancer.sanger.ac.uk/signatures/signatures_v2/, available through the signature.tools.lib R package[59]) with an implementation of non-negative least squares ('SignatureFit' function from the signature.tools.lib package). With the same approach, JaBbA-derived SVs were classified into the 32 SV types on the basis of size, topology and junction clustering as previously described[21], and were fit to rearrangement signatures derived from 560 breast cancers. Microhomology in small deletions was searched in 3′ flanking sequence for up to 25 bases. The HRD-LOH index was determined by the number of segments per genome larger than 15 Mbp (but under the span of an entire chromosome) containing LOH.

Classifying HR, BRCA1 and BRCA2 deficiency

To build classifiers distinguishing overall HR deficiency, BRCA1 deficiency and BRCA2 deficiency, random forests (RFs; from the randomForest R package) were trained on a dataset of pan-cancer primary tumours consisting of 62 BRCA1d, 66 BRCA2d and 2,536 controls that were confidently HRP (lacking CLINVAR pathogenic, CLINVAR VUS, truncating or missense mutation in *BRCA1*, *BRCA2*, *RAD51*, *RAD51B*, *RAD51C*, *RAD51D* and *PALB2* and LOH in *BRCA1* or *BRCA2*). The following six features were counted for each case using the R package signature.tools.lib: COSMIC SNV signatures 3 and 8; rearrangement signatures 3 and 5; HRD-LOH index; and proportion of deletions with microhomology. rDups, rDels and rDelDups were also counted after annotation on each genome graph.

To evaluate the performance of RFs, ROC curves and corresponding AUROCs were computed on an independent test set of pan-cancer metastatic tumours (HMF dataset, Extended Data Fig. 1) consisting of 40 BRCA1d, 92 BRCA2d and 1,834 HRP controls using the pROC R package (v.1.18.0, https://cran.r-project.org/web/packages/pROC/). Feature importance was determined by resampling the test set across 30 bootstraps with permutation. The decrease in accuracy after permuting each feature on the test set was calculated.

In the following two comparisons, classifier skill to discriminate overall HR deficiency from HR proficiency was analysed by using the full (Hadi, MSKCC, and PCAWG) training set and evaluating the resulting models on the full (HMF) test set (Extended Data Fig. 1). An SV-only RF was trained on rDups, rDels, rDelDups, homeologous deletions, duplications with length 10–100 kbp, RS3 and RS5 as features and compared against an RF trained on rearrangement signatures 3 and 5 as features to compare the efficacy of the classes of SVs described in this manuscript against the established SV types previously used in HRDetect[2]. A full RF consisting of currently described features and previously established features (rDups, rDels, rDelDups, homeologous deletions, duplications with length 10–100 kbp, RS3, RS5, MH-dels, SNV3, SNV8 and LOH score) was trained and then tested against the published HRDetect model (consisting of MH-dels, SNV3, SNV8, RS3, RS5 and LOH score as features) using ROC curves and feature importance metrics. HRDetect scores were obtained by running the function 'applyHRDetectDavies2017' from the signature.tools.lib R package on a feature matrix composed of test samples.

The third comparison evaluated classifier skill to discriminate BRCA1 deficiency from BRCA2 deficiency. For this test, the full RF trained with current and previous features was used to compare against a RF trained with HRDetect-only features. In contrast to the above, ROC and feature importance evaluation were performed on only the 40 BRCA1d and 92 BRCA2d cases from the test dataset (Extended Data Fig. 1).

Statistical information

All statistical analysis was performed as stated in the figure legends using the R programming language (v.4.0.2). *P* values obtained that are smaller than 2.2×10^{-16} are not accurately estimated in R and are listed as such ('$P < 2.2 \times 10^{-16}$'). Generalized linear modelling was performed using the 'glm' or 'glm.nb' function from the stats or MASS R packages. Wilcoxon rank-sum testing was performed using the 'wilcox.test' function from the stats R package. Fisher's exact test was performed using the function 'fisher.test' from the stats R package. ROC curves were generated using the function 'roc' from the R package 'pROC'. Comparison of ROC curves was done using the function 'roc.test' from

the R package 'pROC' with the argument 'method = 'delong'. Statistical methods were not used to predetermine sample size. The study design did not involve blinding or randomization.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets generated for the current study include the WGS and LR WGS data for 46 *BRCA1* and *BRCA2*-mutated cases (see 'MSKCC cohort') have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAD00001010326. Further information about EGA can be found at https://ega-archive.org (the European Genome-phenome Archive of human data consented for biomedical research). Processed data and an associated notebook for generating the main and Extended Data figure panels are provided as a GitHub repository (https://github.com/mskilab/setton_hadi_choo_2023). Source data are provided with this paper.

Code availability

Executable notebook code spanning all key analyses across main and Extended Data figure panels is provided as a GitHub repository (https://github.com/mskilab/setton_hadi_choo_2023). Analyses were performed using R v.4.0.2 with R packages available from CRAN (https://cran.r-project.org/). The following lists R packages developed by authors to perform the described analyses. Genome-wide coverages for samples for which a BAM alignment was present were calculated with the fragCounter R package (https://github.com/mskilab-org/fragCounter). Fitting of junction-balanced genome graphs was carried out using the JaBbA R package[18] (https://github.com/mskilab-org/jabba). Analysis of junction links and link clusters as well as classification of complex event types within each genome graph was performed with the function 'eclusters' in the package gGnome (https://github.com/mskilab-org/gGnome). Walk deconvolution on genome graphs was also performed using gGnome. LR WGS barcodes supporting junctions were queried using the 'score.walks' function in the skitools R package (https://github.com/mskilab-org/skitools). Visualization of genomic tracks were made with the gTrack R package (https://github.com/mskilab-org/gTrack). Analysis of sequence homeology across junction break ends is implemented with the function 'homeology' in the package GxG (https://github.com/mskilab-org/GxG). Custom tools for miscellaneous data manipulation tasks were implemented using the package khtools (https://github.com/kevinmhadi/khtools).

43. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
44. Lee, J. J.-K. et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell* **177**, 1842–1857 (2019).
45. Frankell, A. M. et al. The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nat. Genet.* **51**, 506–516 (2019).
46. Paulson, T. G. et al. Somatic whole genome dynamics of precancer in Barrett's esophagus reveals features associated with disease progression. *Nat. Commun.* **13**, 2300 (2022).
47. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
48. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
49. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
50. Cameron, D. L. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).
51. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
52. Fang, L. et al. LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. *Nat. Commun.* **10**, 5585 (2019).
53. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
54. Olivieri, M. et al. A genetic map of the response to DNA damage in human cells. *Cell* **182**, 481–496 (2020).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
56. Shajii, A., Numanagić, I. & Berger, B. Latent variable model for aligning barcoded short-reads improves downstream analyses. *Res. Comput. Mol. Biol.* **10812**, 280–282 (2018).
57. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
58. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
59. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
60. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

Acknowledgements We thank T. de Lange, S. Keeney, H. Klein, J. Skok, and B. Neel for critical feedback during manuscript preparation and C. Black and the New York Genome Center ResComp team for high-performance computing support. We thank the members of the Imielinski Lab for help with proofreading. This work was directly supported by the Starr Cancer Consortium Award I11-0051 to M.I., J.S.R.-F. and S.N.P. J.S., J.S.R.-F., and S.N.P. were also supported by NCI 1P50CA247749 to S.N.P. M.I. was also supported by the Burroughs Wellcome Fund Career Award for Medical Scientists, Doris Duke Clinical Foundation Clinical Scientist Development Award and Weill Cornell Medicine Department of Pathology and Laboratory Medicine startup funds. Certain panels in the figures and Extended Data figures were created with art from BioRender.com under academic licence terms.

Author contributions Conceptualization: J.S., K.H., N.R., J.S.R.-F., S.N.P. and M.I. Data curation: J.S., K.H., Z.-N.C., K.S.K., H.T., J.R., P.S., X.Y., J.B., A.D., A.D.C.P., J.-M.M., O.E., B.W., N.R., J.S.R.-F, S.N.P. and M.I. Formal analysis: K.H., Z.-N.C., J.R., K.S.K, P.S. and M.I. Funding acquisition: M.I., S.N.P., J.S.R.-F. and O.E. Investigation: J.S., K.H., Z.-N.C., H.T., K.S.K., P.S., J.R., J.T.N., N.R., J.S.R.-F., S.N.P. and M.I. Methodology: K.H., Z.-N.C., X.Y., J.B., A.D. and M.I. Project administration: J.S.R.-F., S.N.P. and M.I. Resources: H.T., A.D.C.P., J.-M.M., M.S., J.M, O.E., B.W., J.S.R.-F., S.N.P. and M.I. Software: Z.-N.C., X.Y., J.B., A.D. and M.I. Supervision: S.N.P. and M.I. Validation: K.H., Z.-N.C., K.S.K., P.S., J.R., H.T., J.S.R.-F. and M.I. Visualization: J.S., K.H., Z.-N.C., K.S.K. and M.I. Writing (original draft): J.S., K.H., S.N.P. and M.I. Writing (review and editing): all authors.

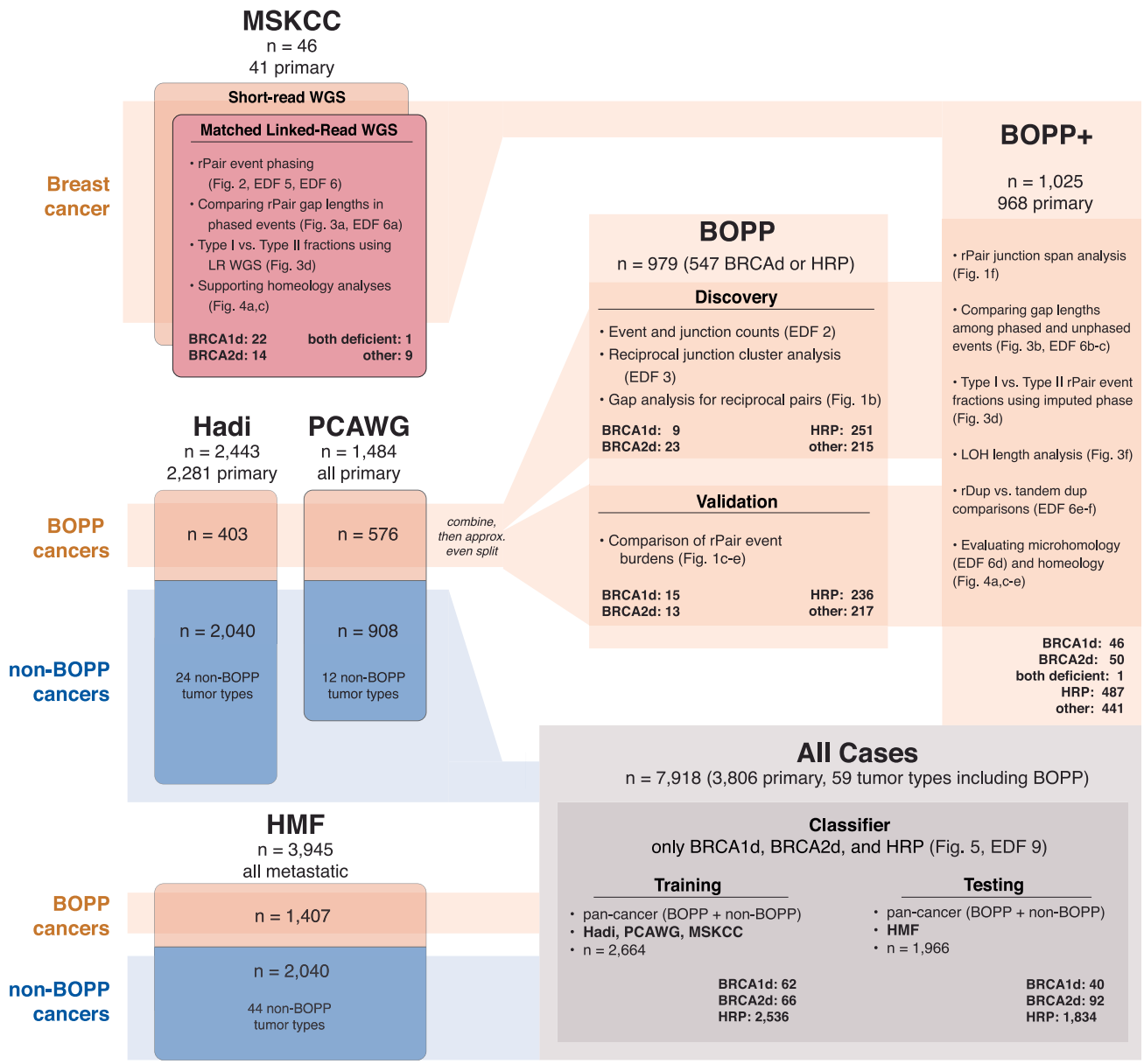
Competing interests J.S.R.-F. reports receiving personal or consultancy fees from Goldman Sachs, Bain Capital, REPARE Therapeutics and Paige.AI, membership of the scientific advisory boards of VolitionRx, REPARE Therapeutics, Personalis and Paige.AI, membership of the Board of Directors of Grupo Oncoclinicas and ad hoc membership of the scientific advisory boards of Roche Tissue Diagnostics, Ventana Medical Systems, Novartis, Genentech, Merck, Daiichi Sankyo and AstraZeneca, outside the scope of this study. B.W. reports ad hoc membership of the scientific advisory board of REPARE Therapeutics. S.N.P. reports the following: consulting fees: Varian Medical Systems, Philips Healthcare, AstraZeneca; research funding: Philips Healthcare, Varian Medical Systems. M.I. is a member of the scientific advisory board of ImmPACT Bio.

Additional information

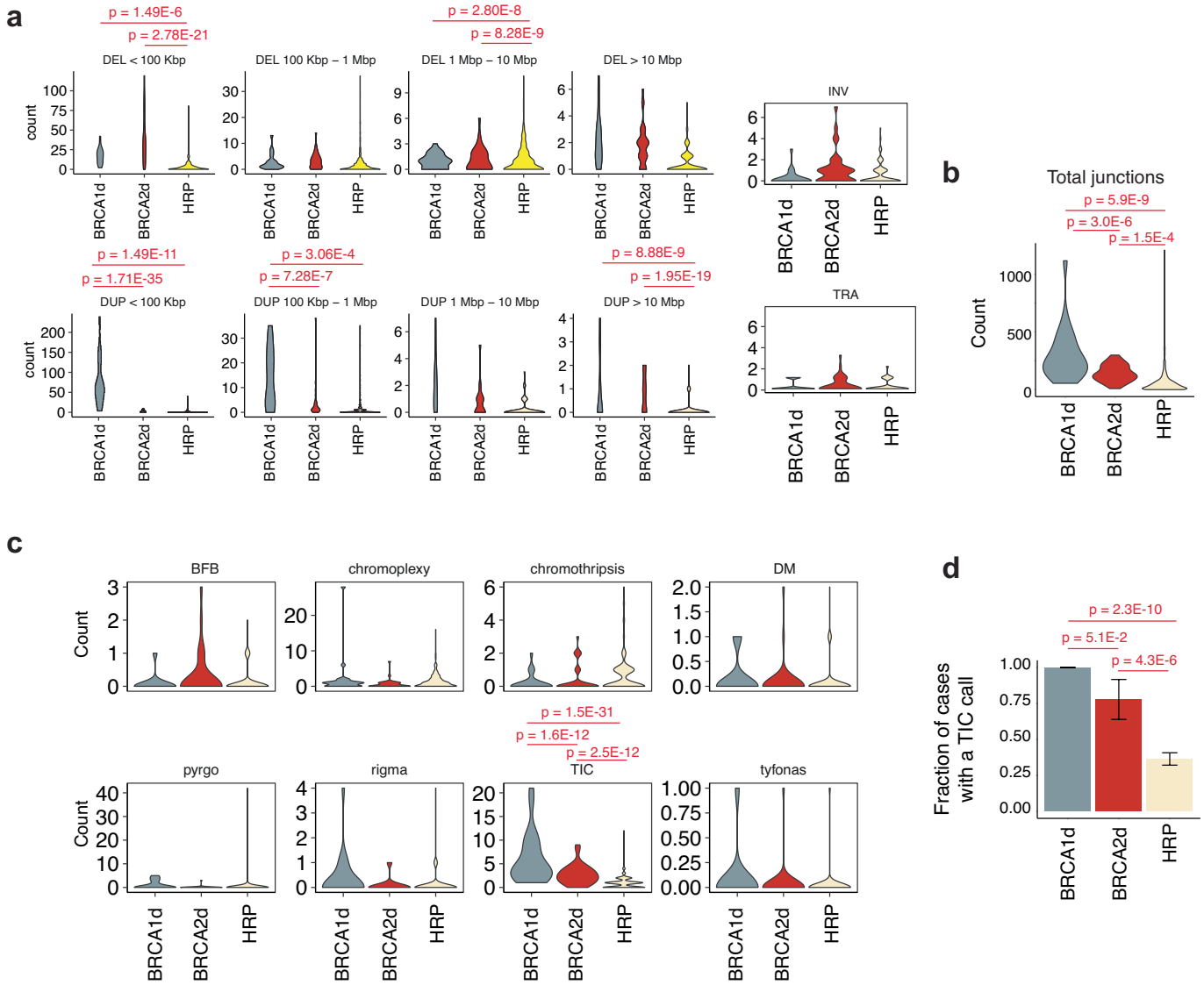
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06461-2.

Correspondence and requests for materials should be addressed to Simon N. Powell or Marcin Imieliński.

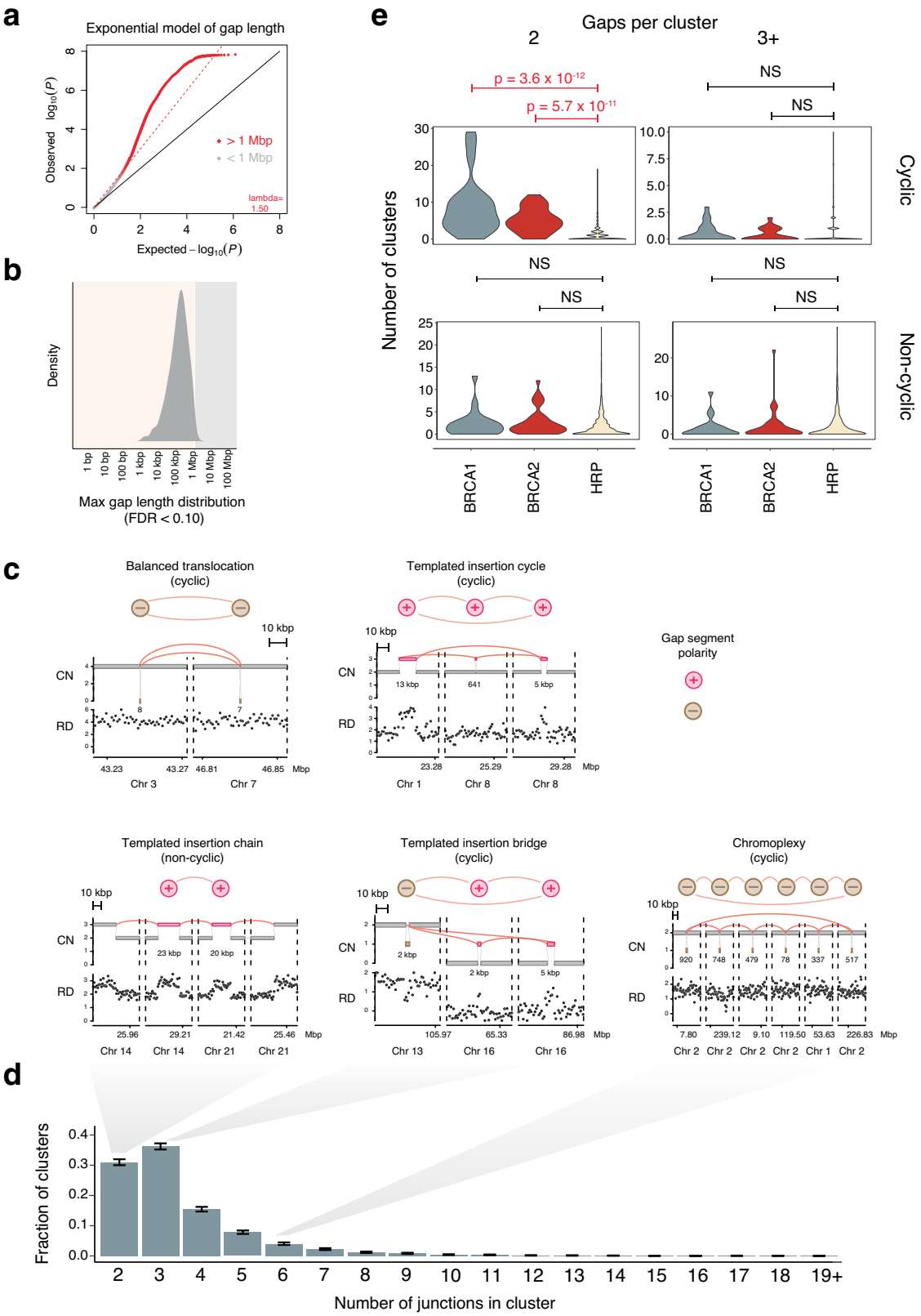
Peer review information *Nature* thanks E. Cuppen, Marcel Tijsterman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



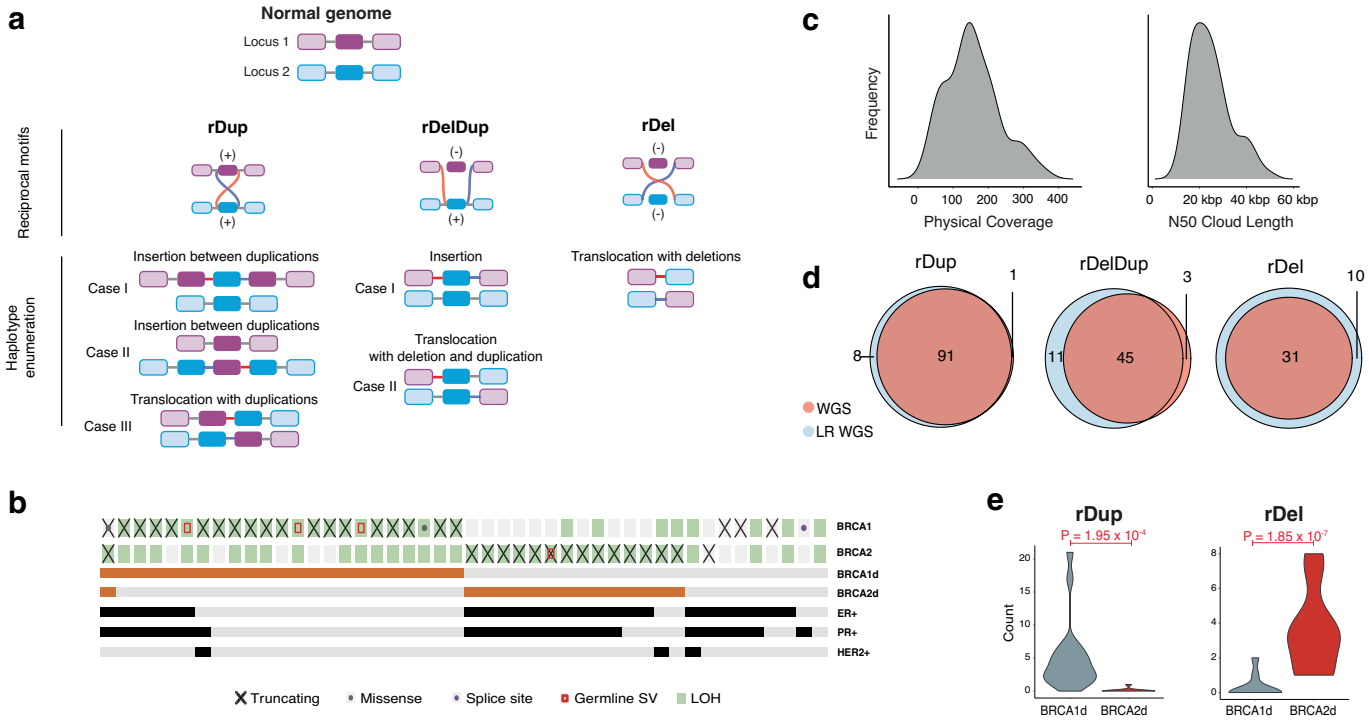
Extended Data Fig. 1 | Overview of datasets and analyses. Schematic illustrating cases included for analysis in the present manuscript. MSK, Memorial Sloan Kettering Cancer Center. PCAWG, Pan-Cancer Analysis of Whole Genomes/International Cancer Genome Consortium. HMF, Hartwig Medical Foundation. BOPP, breast, ovarian, pancreas, and prostate. Hadi refers to WGS profiles from a previously published pan-cancer analysis (ref. 18). Cohorts are mapped to key analyses in the study, denoted by main and Extended Data figure (EDF) panels.



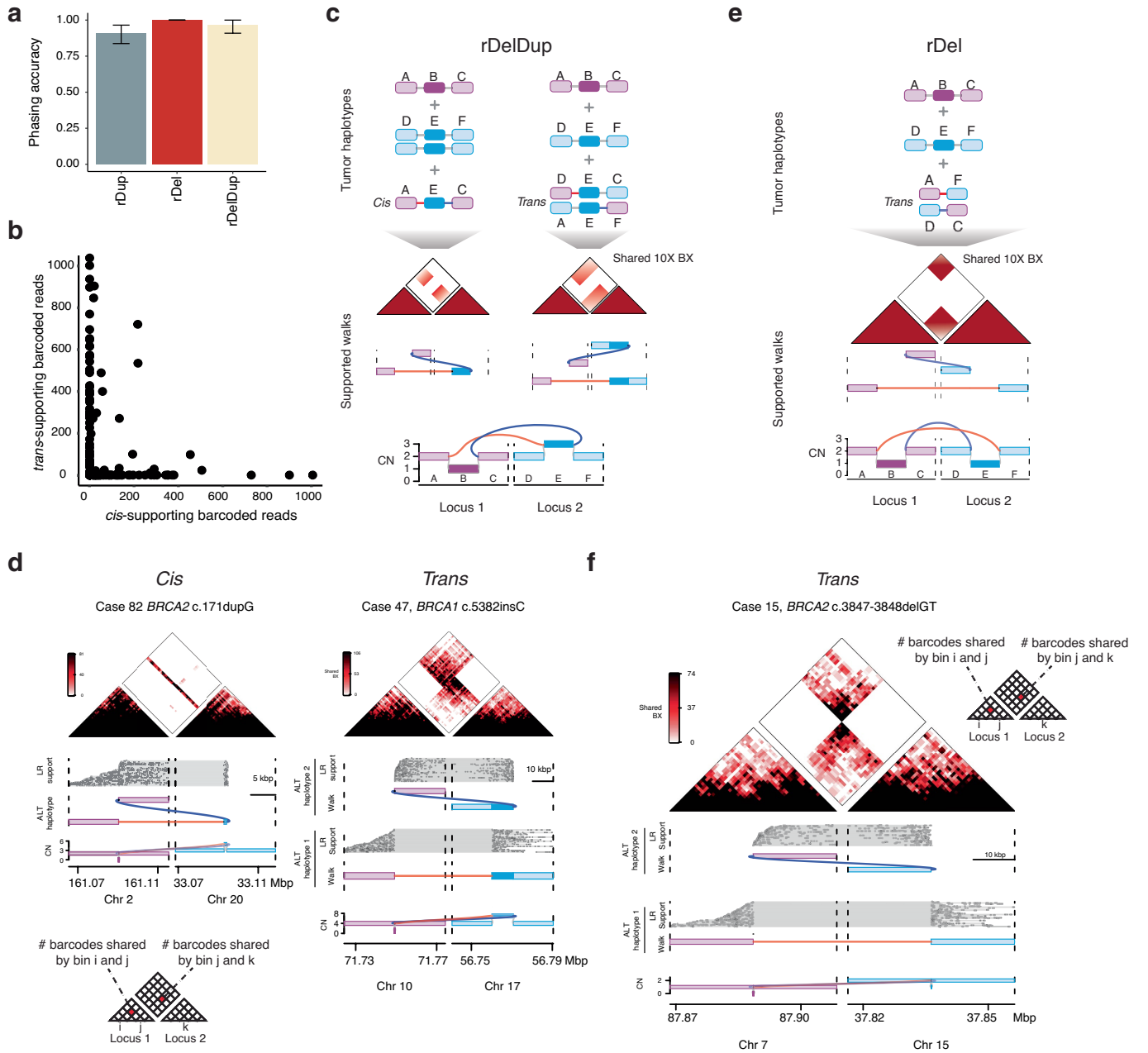
Extended Data Fig. 2 | Comparing the burden of SV classes across genotypes. a–**c**, Comparison of the burden (count) of simple SV classes (**a**), total junctions (**b**) and complex SV classes (**c**) across BRCA1d ($n = 9$), BRCA2d ($n = 23$) and HRP cases (HRP, n = 251). Templated insertion chains (TICs): BRCA1d vs HRP, RR 6.56, P = 1.5 x 10^{-31}, BRCA2d vs HRP, RR 2.96, P = 2.5 x 10^{-12}. P values obtained by two-sided Wald test on a gamma-Poisson regression model. (BFB, breakage–fusion–bridge. DM, double minute. TIC, templated insertion chain.) **d**, Fraction of cases with at least one TIC event across BRCA1d ($n = 9$ tumours with TIC/$n = 9$ tumours), BRCA2d ($n = 17$ tumours with TIC/$n = 23$ tumours) and HRP cases ($n = 90$ samples with TIC/$n = 251$ tumours). Error bars show 95% confidence interval on the Bernoulli trial parameter. *P* values and odds ratios obtained by Fisher's exact test.



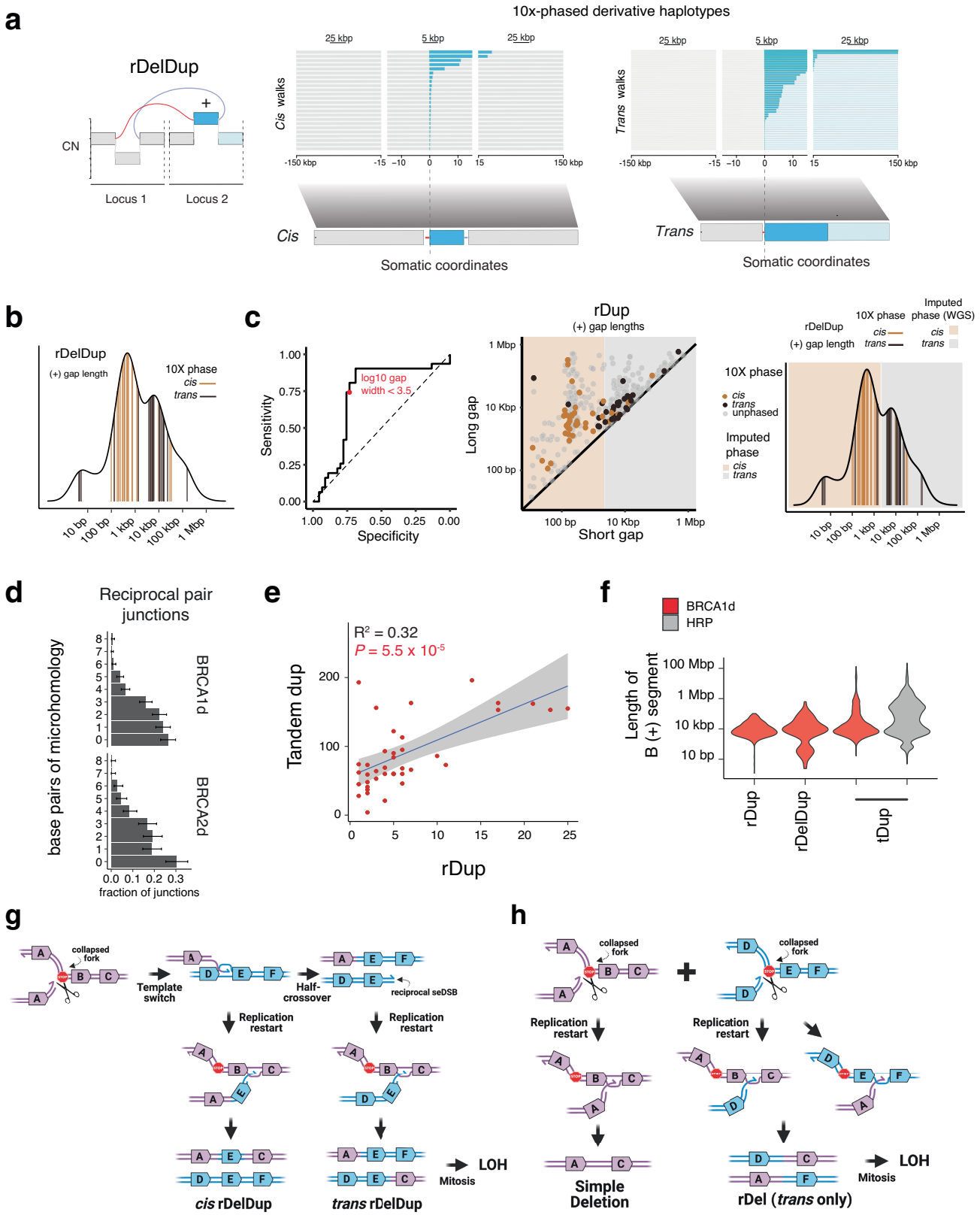
Extended Data Fig. 3 | Characteristics of reciprocal junction clusters. **a**, Quantile-quantile plot of observed $-\log_{10} P$ values obtained by evaluating 473,382 observed gap segment lengths from 283 tumour samples against an exponential null model (see Methods). The *x* axis represents $-\log_{10}$ transformed quantiles from the uniform distribution. Gap segment lengths less than 1 Mbp are shown in red. **b**, Density plot showing distribution of gap segment lengths with FDR < 0.1 across 283 tumour samples the vast majority of which are less than 1 Mbp. **c**, Examples of 2-, 3- and multi-way (based on gap segment number) reciprocal SV topologies, including a spectrum of simple and complex SV classes. **d**, Histogram of reciprocal SV cluster lengths across 283 tumour samples and 1,854 junctions. **e**, Violin plots comparing cyclic versus linear and 2-way versus higher order (3+ way) reciprocal SV topologies across 283 tumour samples. Error bars show 95% confidence interval on the Bernoulli trial parameter. *P* values obtained by two-sided Wald test on a gamma-Poisson regression model.



Extended Data Fig. 4 | Haplotype deconvolution of reciprocal pairs with LR WGS validates reciprocal pairs in *BRCA1*- and *BRCA2* mutated cancers. **a**, Enumeration of all possible phased allelic deconvolutions of each of the three reciprocal-pair classes (rDups, rDelDups and rDels). **b**, Molecular and pathological characteristics of the 46 breast cancer cases profiled with LR WGS in the MSKCC cohort. **c**, Sequencing library characteristics across 92 LR WGS profiled samples. **d**, Concordance between LR WGS and short-read WGS reciprocal-pair calls among BRCA1d ($n = 22$) and BRCA2d ($n = 14$) cases. **e**, Comparison of rDup and rDel counts among BRCA1d ($n = 22$) and BRCA2d ($n = 14$) cases within the MSKCC cohort. *P* values obtained by two-sided Wald test on a gamma-Poisson regression model.

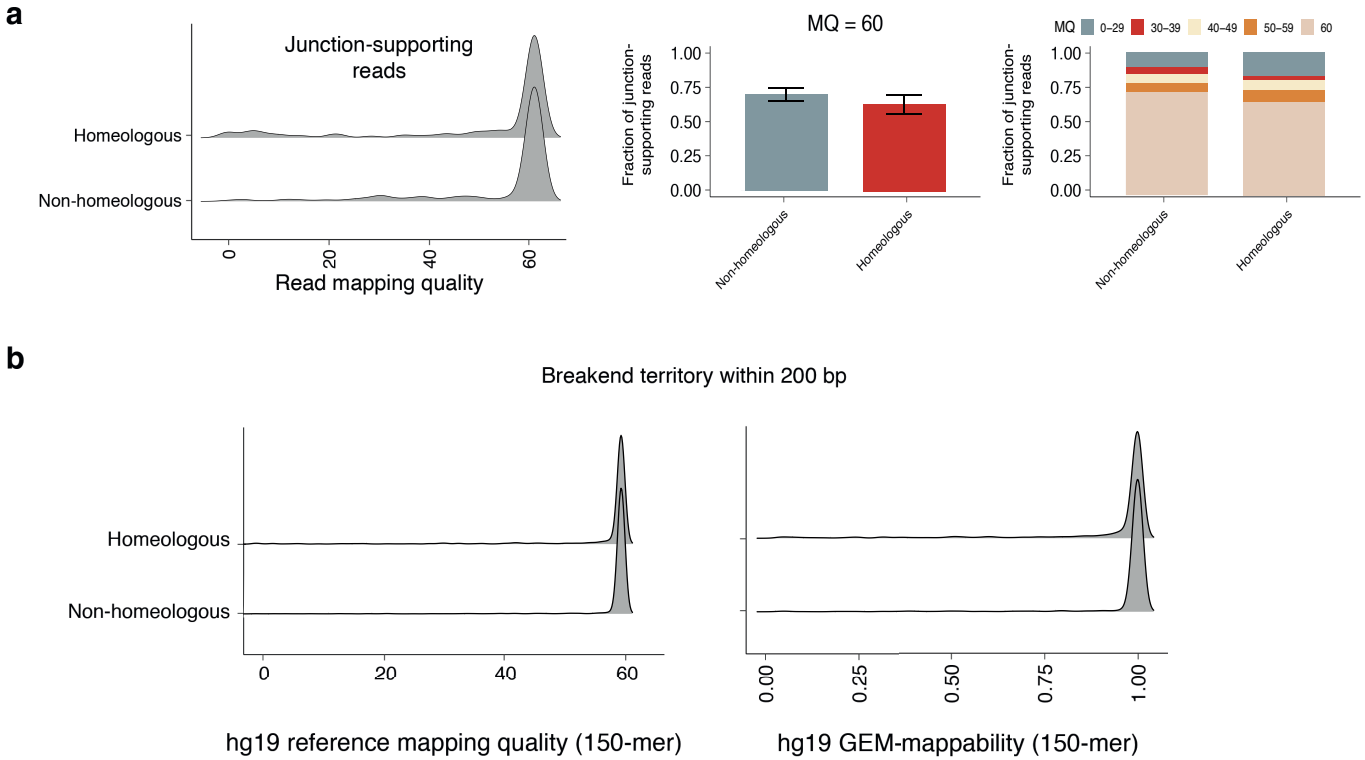


Extended Data Fig. 5 | Phasing rDels and rDelDups with LR WGS. a, Accuracy of phasing algorithm across simulated LR data sampled from *cis* or *trans* allelic configurations of 268 observed rDel, rDelDup and rDel loci. Number of correct calls/number of total events for each rPair event are as follows: rDup: 108/160, rDelDup: 61/68, rDel: 39/40. Error bars show 95% confidence interval on the Bernoulli trial parameter. **b**, Number of LR barcodes supporting *cis* versus *trans* configuration for each of the 186 reciprocal pair loci observed across 46 tumours profiled with LR WGS. **c**,**e**, Predicted LR profiles for an rDelDup (**c**) or rDel (**e**). Each plot shows a genome graph (track labelled 'CN', copy number) and possible allelic walks in reference coordinates (track labelled 'Supported walks') together with the barcode sharing heat map (middle top) and resulting *cis* and/or *trans* haplotypes in the cancer genome. In each plot, the track labelled 'Supported walks' represented enumerated paths in reference coordinates through the corresponding graph ('CN' track) with LR data supporting the barcode sharing heat map above ('LR support' track). Each heat map shows the number of barcodes shared by pairs of bins within and between locus 1 and 2. The 'Tumour haplotypes' track shows the total phased linear alleles in somatic coordinates inferred based on the LRS data. **d**,**f**, Examples of a (**d**) *cis* and *trans* rDelDup and (**f**) *trans* rDel across three BRCA1d or BRCA2d cases. CN, copy number. ALT, rearranged haplotype. BX, linked-read barcodes. LRS, linked-read WGS.

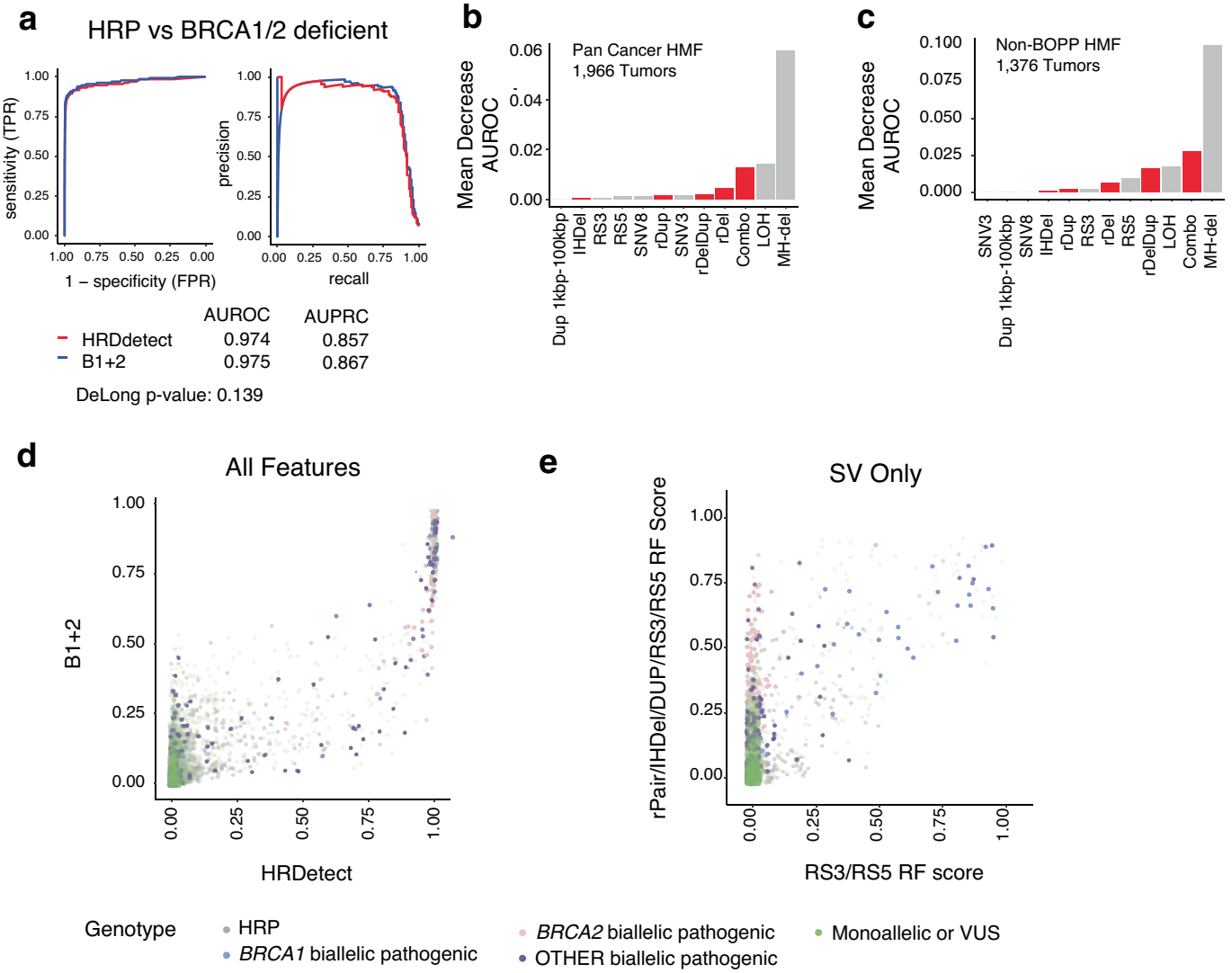


Extended Data Fig. 6 | See next page for caption.

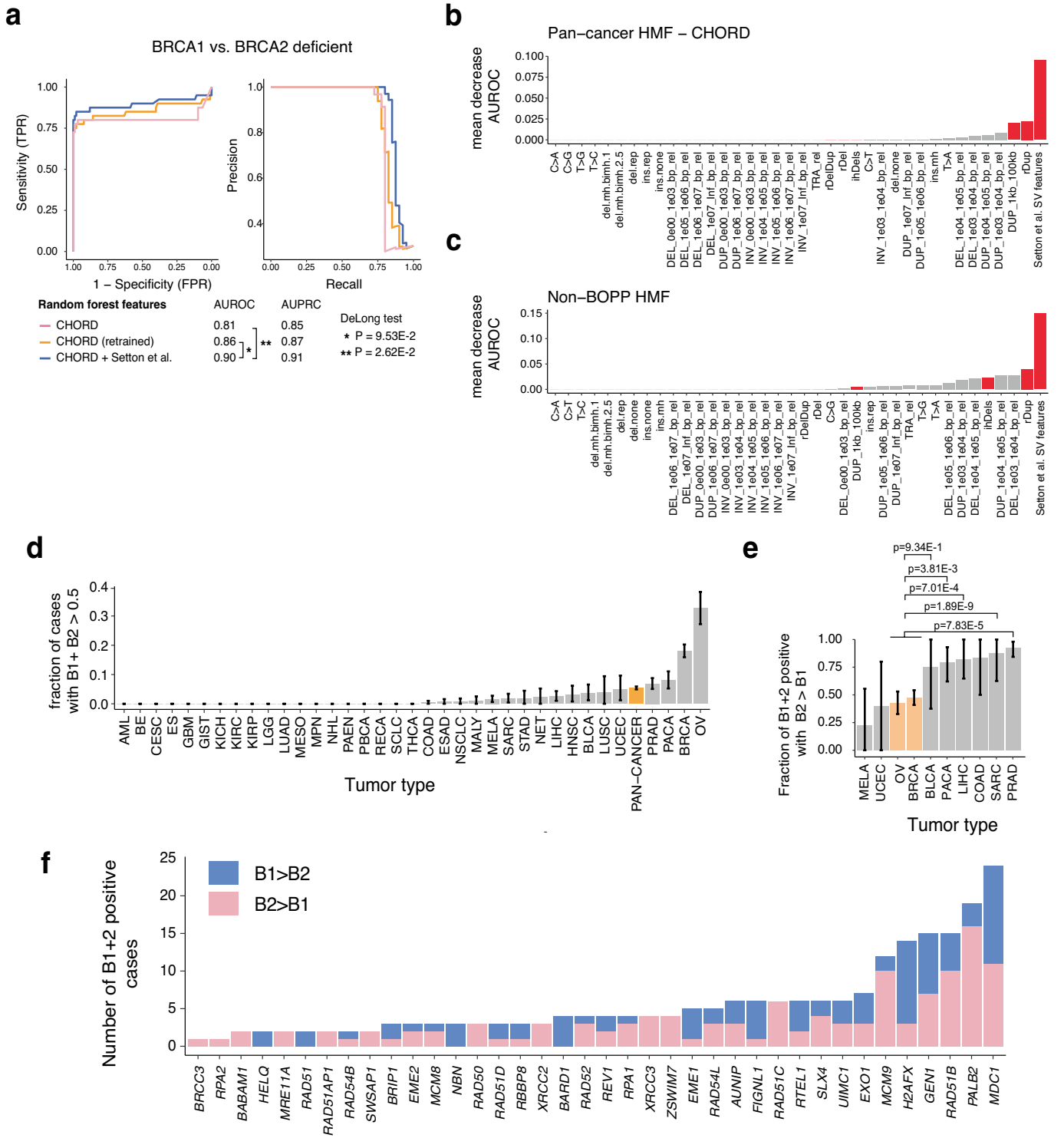
Extended Data Fig. 6 | Links between reciprocal pairs, LOH and tandem duplications. a, Schematic defining the (+) gap segment of rDelDup reciprocal pairs (left). The (+) gap segment lengths of *cis* and *trans* phased haplotypes from BRCA1d ($n = 22$) and BRCA2d ($n = 14$) samples are shown in the centre and right panels, respectively. **b**, (+) gap segment length distribution of unphased rDelDup reciprocal pairs across $n = 96$ BRCA1d or BRCA2d WGS samples with vertical lines denoting lengths of LR WGS phased events across $n = 36$ BRCA1d or BRCA2d samples. **c**, ROC curve for *cis*/*trans* phasing using gap segment length threshold (left panel) with selected value shown in red (*cis* phase is imputed if the $\log_{10}$ length < 3.5). rDup (centre) and rDelDup (right) gap segment length distributions with background colour showing the length threshold used for phase imputation. **d**, Base pairs of microhomology at 644 and 328 reciprocal-pair junctions across 46 BRCA1d and 50 BRCA2d tumours, respectively. Error bars show 95% confidence interval on the Bernoulli trial parameter. **e**, rDup and tandem duplication count per sample across 46 BRCA1d samples. R^2 and *P* value obtained by two-tailed Spearman rank correlation. **f**, Violin plots showing length of the longer rDup gap segment, (+) rDelDup gap segment and tandem duplication segment in BRCA1d (red) ($n = 46$) and HRP ($n = 487$) samples (grey). *P* values obtained by two-tailed Wilcoxon rank-sum test. **g**,**h**, Extensions of the aberrant replication-restart model for rDups (Fig. 3g) can be used to explain rDelDups (**g**) and rDels (**h**) around a locus (ABC) that undergoes replication-fork collapse and invades a second locus (DEF), resulting in distinct phased outcomes including *trans* configurations that can lead to LOH in subsequent cell cycles. Diagrams in **(g)** and **(h)** created with BioRender.com.



Extended Data Fig. 7 | Support for homeologous junctions. a, Comparison of mapping quality for reads supporting homeologous and non-homeologous junctions. Left, density plot showing mapping quality for junction-supporting reads realigned to hg19, from a subsample of 500 non-homeologous junctions and 500 homeologous junctions. Middle, bar plot of the fraction of reads with high mapping quality (MQ = 60). Error bars represent 95% confidence interval on the Bernoulli trial parameter calculated using the fraction of non-homeologous (73,871/95,934) or homeologous (16,954/31,078) junction-supporting reads that had MQ of 60. Right, bar plot of the fraction of reads stacked by intervals of mapping quality. The number of reads in each MAPQ range for homeologous junctions (31078 total) is: MAPQ 0–29: 12,025, MAPQ 30–39: 626, MAPQ 40–49: 676, MAPQ 50–59: 797, MAPQ 60: 16,954. The number of reads in each MAPQ range for homeologous junctions (95,934 total) is: MAPQ 0–29: 19,201, MAPQ 30–39: 1,052, MAPQ 40–49: 868, MAPQ 50–59: 942, MAPQ 60: 73,871. **b**, Left, reference 150-mer BWA mapping quality in the neighbourhood of homeologous and non-homeologous break ends. Reference mapping quality determined by realigning sliding window of 150-mers from hg19 stepping by 1 base to hg19 and averaging across each base pair. Right, plot of alternate mappability scores calculated as the average of the reciprocal of the number of unique locations that each 150-mer overlapping a break-end-associated base pair aligns.

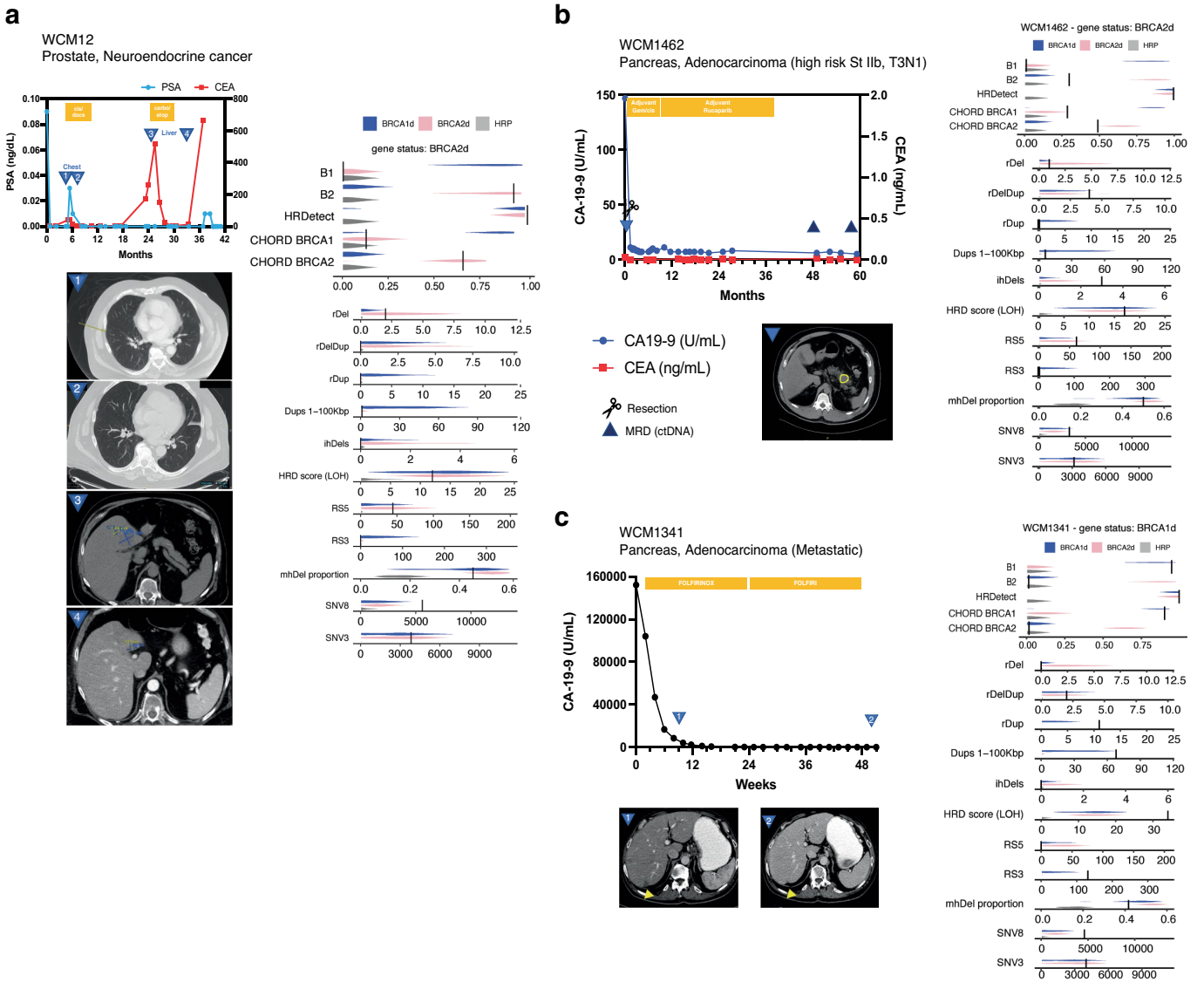


Extended Data Fig. 8 | Comparing the performance of B1+2 and HRDetect. **a**, ROC curves and PRCs comparing B1+2 and HRDetect performance on detecting HR deficiency (BRCA1d or BRCA2d) in an independent (HMF) test set. *P* value denotes comparison of AUROC by DeLong test. **b**,**c**, B1+2 feature importance for detecting HR deficiency in (**b**) an independent pan-cancer WGS dataset and (**c**) its non-BOPP subset. See Extended Data Fig. 1 and Methods for training and testing dataset summary. **d**, Scatter plot of B1+2 versus HRDetect scores across all 7,918 tumours, including those excluded from training and testing **e**, Scatter plot of scores from a random forest model trained only on SV features from the B1+2 (*y* axis) or HRdetect classifier ($n = 7{,}918$).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Distinguishing between BRCA1 and BRCA2 deficiency. **a**, ROC curves and PRCs for discriminating BRCA1d from BRCA2d tumours by CHORD and by a random forest trained on CHORD features augmented with the SV classes introduced in this manuscript. *P* values obtained by two-tailed DeLong test. **b**,**c**, Feature importance for BRCA1 versus BRCA2 deficiency classification in (b) an independent pan-cancer WGS dataset and (c) its non-BOPP subset. See Extended Data Fig. 1 and Methods for training and testing dataset summary. **d**, Tumour type and pan-cancer prevalence of B1+2 positivity (n = 7,918 tumour samples). The fraction of B1+2 positive (B1 + B2 score > 0.5) cases per tumour type is shown on the left; only tumour types with at least 20 examples are shown. Orange bar denotes the pan-cancer B1+2 positivity rate. Error bars show 95% confidence interval on the Bernoulli trial parameter. **e**. The fraction of cases with B2>B1 out of the cases that were B1+2 positive (n = 7,918 tumour samples); only tumour types with at least five B1+2 positive examples are shown. Tumour types significantly enriched for B2 positivity relative to the reference class (breast and ovarian cancer, highlighted in orange) are indicated with stars. Error bars show 95% confidence intervals on the Bernoulli trial parameter and stars indicate relative enrichment of B2 positivity within each tumour type. *P* values were obtained by two-tailed Fisher's exact test. **f**, Number of B1+2 positive (B1+ B2 score > 0.5) cases with B1>B2 or B2>B1 and harbouring biallelic or monoallelic and pathogenic or VUS variants in HR-associated genes (see main text and Methods). LOH = burden of large genomic segments harbouring loss of heterozygosity. HomeoDel = count of large deletions (>1 kbp) with homeology. del-MH = proportion of small deletions (<50 bp) with microhomology. RS3, RS5 = proportion of junctions with rearrangement signature 3 or 5 (ref. 2). SBS3, SBS8: COSMIC single base signature 3 and 8 (ref. 60).



Extended Data Fig. 10 | Clinical vignettes of patients with HRD tumours. **a**, Patient with BRCA2d neuroendocrine prostate cancer. Top left, prostate-specific antigen (PSA) and carcinoembryonic antigen (CEA) response kinetics. Time points 1–4 correspond to axial computed tomography images depicted on bottom right, illustrating favourable response to platinum-based chemotherapy. Right, classifier scores (top) and genomic features (bottom) for the highlighted patient (vertical lines) vs dataset-wide distributions stratified by genotype (violin plots; BRCA1d n = 102, BRCA2d n = 158, HRP n = 4360, HRP). **b**, Metastatic pancreas adenocarcinoma case with high B1 score (0.962), with CA19-9 response kinetics (left) and serial axial computed tomography (CT) (blue triangles and corresponding right panels) demonstrating excellent response to chemotherapy. **c**, High risk stage IIb pancreas adenocarcinoma case (WCM1462) with high B2 score (0.31), with CA19-9 and CEA response kinetics (top) and axial CT (bottom). LOH = burden of large genomic segments harbouring loss of heterozygosity. ihDels = count of large deletions (>1 kbp) with homeology. mhDel proportion = proportion of small deletions (<50 bp) with microhomology. RS3, RS5 = proportion of junctions with rearrangement signature 3 or 5 (ref. 2). SBS3, SBS8: COSMIC single base signature 3 and 8 (ref. 60).

nature portfolio

Corresponding author(s): Marcin Imielinski

Last updated by author(s): May 16 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No code used for data collection.
Data analysis	Short reads were aligned to the GRCh37/hg19 reference using Burrows-Wheeler aligner software58, bwa mem, 0.7.10-r789. Read post-processing was done in accordance with best practices for post-alignment data processing with Picard tools (https://broadinstitute.github.io/picard/) to mark duplicates, the GATK (v.2.7.4) (https://gatk.broadinstitute.org/hc/en-us) IndelRealigner module, and GATK base quality recalibration. All linked-reads were aligned to GRCh37/hg19 with the EMerAld aligner (v0.6.2). Germline haplotypes were obtained from Strelka2 germline SNV calls processed using HapCut2 (github.com/vibansal/HapCUT2). SNV signatures were deconvolved using the known signature weights from COSMIC SNV signature version 2 (https://cancer.sanger.ac.uk/signatures/signatures_v2/, available through signature.tools.lib R package with an implementation of non-negative least squares ("SignatureFit" function from the signature.tools.lib package). To evaluate performance of random forests, receiver-operating characteristic (ROC) curves and corresponding areas under the curve (AUCs) were computed using the pROC R package (v1.18.0, https://cran.r-project.org/web/packages/pROC/). Generalized linear modeling was performed using "glm" or "glm.nb" function from the stats or MASS R package. Wilcoxon rank sum testing performed using "wilcox.test" function from the stats R package. Fisher's exact test was performed using the function "fisher.test" from the stats R package. Receiver-operator curves (ROC) were generated using the function "roc" from the R package "pROC". Comparison ROC curves was done using the function "roc.test" from R package "pROC" with argument "method = 'delong'". Analyses were performed using R-4.0.2 with R packages available from CRAN (https://cran.r-project.org/). The following lists R packages developed by authors to perform the described analyses. Genome-wide coverages for samples for which a BAM alignment was present were calculated with the fragCounter R package (github.com/mskilab/fragCounter). Fitting of junction-balanced genome graphs was carried out using JaBbA R package (github.com/mskilab/jabba) (Hadi et al. 2020). Analysis of junction links and link clusters as well as classification of complex event types within each genome graph was performed with the function "eclusters" in the package gGnome (github.com/mskilab/

gGnome). Walk deconvolution on genome graphs was also performed using gGnome. l0X LR barcodes supporting junctions were queried using the "score.walks" function in the skitools R package (github.com/mskilab/skitools). Visualization of genomic tracks were made with the gTrack R package (github.com/mskilab/gTrack). Analysis of sequence homeology across junction breakends is implemented with the function "homeology" in the package GxG (github.com/mskilab/GxG). Custom tools for miscellaneous data manipulation tasks were implemented using the package khtools (github.com/kevinmhadi/khtools).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets generated for the current study include the WGS and 10X linked-read sequencing data for the 46 BRCA1&2-mutated cases (see Linked-read whole genome sequencing cohort) have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAD00001010326. Further information about EGA can be found at https://ega-archive.org (the European Genome-phenome Archive of human data consented for biomedical research).The datasets generated for the current study include the WGS and l0X linked-read sequencing data for the 46 BRCA1&2-mutated cases (see Linked-read whole genome sequencing cohort) are available for download under NCBI BioProject accession: PRJNA746293.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Cancer genomes were included for analysis in this study irrespective of their sex or gender.
Reporting on race, ethnicity, or other socially relevant groupings	Cancer genomes were included for analysis in this study irrespective of race, ethnicity, or other socially relevant groupings.
Population characteristics	Primarily European ancestry cancer genomes, see Extended Data Figure 1 for additional cohort details.
Recruitment	Consecutive breast cancer genomes with germline BRCAl/2 alterations (consented to MSK IRB 06-107, 12-245) were included in LR-sequencing cohort. Additional genomes included as described in methods.
Ethics oversight	Ethics oversight provided in setting of multi-institution collaborative research effort comprised of Memorial Sloan Kettering Cancer Center, New York University, Stony Brook University Hospital, Lenox Hill, Northwell Health, Columbia University, Montefiore, Cornell, and led by the New York Genome Center were included here and were previously described in (Hadi et al. 2020). Study approval was obtained via a central institutional review board (IRB), Biomedical Research Alliance of New York, and by local IRBs.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	See Extended Data Figure 1. No sample size calculation was performed; all available genomes were used in our analysis and for each comparison sufficient numbers were determined based on an FDR-corrected p-value and magnitude of effect size.
Data exclusions	To investigate the role of complex SVs in HR-deficient cancers, we assembled a cohort of 979 predominantly (95%) cancer WGS profiles profiles from four tumor types commonly associated with HR-deficiency (breast, ovary, prostate, and pancreas cancer; referred to as BOPP moving forward, see Methods and Supplementary Fig. 1) (Roy et al. 2011). We next sought to identify confidently BRCAld, BRCA2d, and HR-proficient cases in this BOPP cohort. We required biallelic inactivation of BRCAl or BRCA2 for a tumor to be classified as BRCAld (n=24) or BRCA2d (n=36) respectively (Riaz et al. 2017) (see Methods). We also identified 487 HR proficient BOPP samples that lacked pathogenic or rare variants in any HR-associated gene (e.g. BRCAl, BRCA2, PALB2, RADSlC; see Supplementary Table 1 for full list). We excluded the

	remaining 432 BOPP cases, which comprised tumors with monoallelic alterations and/or variants of unknown significance (VUSs) in BRCAl or BRCA2 or mutations in other HR-associated genes.
Replication	See Extended Data Figure 1. We demonstrated the robustness of SV calling by recapitulating our results with an alternative SV caller (GRIDSS) or a consensus caller, demonstrating that our results are not dependent on the choice of SV-calling algorithm.
Randomization	Not applicable as no intervention was analyzed (not possible to randomize the effect of genotype on structural variation).
Blinding	Not applicable as the outcome measured is objective (genomic structural variation), and no intervention was analyzed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging