

DNA shape and epigenomics distinguish the mechanistic origin of human genomic structural variations

Nadejda B Boev^{1,2}, Mark B Gerstein^{3,4,5,6}, Sushant Kumar^{1,2,*}

¹Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada

²Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2C4, Canada

³Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, United States

⁴Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, United States

⁵Department of Computer Science, Yale University, New Haven, CT 06520, United States

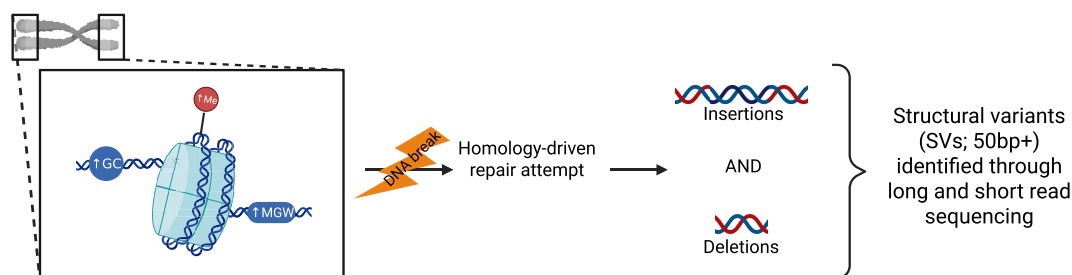
⁶Department of Statistics & Data Science, Yale University, New Haven, CT 06520, United States

*To whom correspondence should be addressed. Email: sushant.kumar@utoronto.ca

Abstract

The recent advent of long-read whole genome sequencing has enabled us to create an accurate telomere-to-telomere reference genome, construct pangenome graphs, and compile precise catalogs of genomic structural variations (SVs). These comprehensive SV repositories provide an excellent opportunity to explore the role of SVs in genotype-phenotype associations and examine the mechanisms by which SVs are introduced through double-strand break (DSB) repair. Here, we employed comprehensive SV catalogs identified through various short- and long-read whole genome sequencing efforts to infer the underlying mechanisms of SV introduction based on their genomic and epigenomic profiles. Our findings indicate that high local DNA methylation and DNA shape-related features, such as low variations in propeller twist, support the origins of homology-driven SVs. Subsequently, we utilized an active-learning-based unsupervised clustering approach, revealing that homology-dependent SVs show greater evidence of retaining ancestral recombination patterns compared to their homology-independent counterparts. Finally, our comparison of inherited and *de novo* SVs from healthy populations and rare disease cohorts showed distinct upstream H3K27me3 levels in *de novo* SVs from individuals with ultra-rare disorders. These findings highlight genome-wide characteristics that may influence the choice of repair mechanisms linked to heritable SV origins.

Graphical abstract



Introduction

Over the last decade, there has been a continuous increase in large-scale genomic sequencing studies aimed at investigating genotype-phenotype relationships across healthy populations and various common and rare diseases [1, 2]. Most of these studies have focused on characterizing short genomic variants, including single-nucleotide variants (SNVs) and small insertions/deletions (INDELs) under 50 bp. However, structural variants (SVs), operationally defined as variants longer than 50 bp, represent a significant source of genomic variation, as they induce more nucleotide-level changes than both SNVs and INDELs combined [3]. SVs are further categorized as insertions, deletions, inversions, and translocations based on their orientation and modifications to DNA segments rel-

ative to a reference genome. Due to their longer length, SVs can simultaneously influence multiple coding and regulatory elements[4, 5] and are implicated in various diseases[6–8].

Current genomic studies primarily focus on interpreting SNVs and INDELs due to our superior ability to identify these variants accurately. Short-read whole genome sequencing (sr-WGS) platforms are commonly employed in large-scale genomic studies; however, they struggle to produce a precise SV catalog. This challenge stems from SVs occurring in or near highly repetitive regions of the human genome, resulting in multi-mapping and/or mis-mapping of sequencing reads, which leads to inaccurate or non-specific SV calling [9, 10]. The recent emergence of long-read whole genome sequencing (lrWGS) aims to address these challenges. Implementing

Received: April 24, 2025. Revised: September 24, 2025. Accepted: October 22, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

lrWGS in large-scale studies has generated comprehensive SV maps for healthy populations and various diseased individuals, deepening our understanding of population diversity [11–13] and uncovering the pathogenic role of SVs in diseases [7, 14]. These new SV repositories also present a valuable opportunity to investigate the mechanistic origins of SVs.

Previous studies have linked the origin of SVs to double-strand breaks (DSBs) resulting from either effective or ineffective repair attempts [15–18]. These mechanisms include non-homologous end joining (NHEJ), alternative end-joining (aEJ), single-strand annealing (SSA), and homologous recombination (HR), which require increasing amounts of local homology and resection during their repair process [15]. Importantly, faulty repair mechanisms can lead to various diseases. For example, global DSB repair deficiencies are associated with LIG4 syndrome [19], Fanconi anemia [20], and an increased risk of multiple cancers [21]. Additionally, defective non-allelic homologous recombination (NAHR) can result in 15q13.3 microdeletion syndrome [22]. Finally, evaluating DSB repair efficacy can help predict therapeutic responses in patients with inherited or acquired HR deficiency (HRD) when receiving treatment with PARP inhibitors for breast, ovarian, prostate, or pancreatic cancers [23, 24].

Considering their clinical utility, computational and experimental studies have sought to understand DSBs and the mechanisms by which DSB repair can introduce variants (including SVs). For instance, previous research has employed repair-associated features to construct simple decision trees for categorizing germline SVs found in the 1000 Genomes Project (1KG phase 1) [25, 26] and variants in HuRef [27]. While this approach is sound, these methods mainly rely on sequence homology and repetitive element annotations to define an SV mechanism class. Moreover, these studies have only characterized a small subset of SVs derived from low-coverage srWGS data, thereby overlooking the comprehensive landscape of SVs across the genome. In contrast, in recent years, numerous experimental studies have revealed complex interactions among various genomic, epigenomic, and biophysical features during DSB repair [28, 29]. For example, multiple low-throughput reporter assays (such as BLESS [30] or TRIP [31]) suggest that DNA accessibility is a relevant repair feature [16, 32, 33]. However, these studies are inherently biased [34], as DSB formation is influenced by the selected restriction enzyme or CRISPR [35, 36], variability in mechanism frequency due to the lack of cell cycle synchronization [37], and biased repair usage depending on the cell types and states [38, 39]. The inherent limitations of experimental approaches have resulted in inconsistencies across these studies. Furthermore, structural biology studies have indicated that the biophysical features of DNA near the breakpoints, such as flexibility, kinking, or minor groove width (MGW), may lead to preferential repair by repair complexes, although these studies are limited in throughput [40, 41]. Thus, it remains unclear whether there are strict rules governing DSB repair choices. For example, whether euchromatin or heterochromatin, including facultative and constitutive, is consistently and faithfully repaired via HR or another homology-based repair mechanism (i.e. SSA or aEJ) [42].

Considering these knowledge gaps and the rise of lrWGS-based accurate SV catalogs, sufficient data is now available to revisit and explore the relationship between SVs and features that support DSB repair on a genome-wide scale. Therefore, we utilized accurately identified SVs from multiple lrWGS

efforts (including the Human Genome Structural Variation Consortium [13] (HGSVC) and 1KG-ONT [43] study, as well as high-coverage srWGS (1KG phase3 extended project) [44] to comprehensively characterize the features underlying repair mechanisms. We began similarly to previous work [13, 25, 26] by employing a straightforward homology-based workflow to categorize SVs into distinct mechanism classes. Then, we systematically investigated the underlying discriminatory sequence, biophysical, and epigenomic patterns associated with different homology-based SV categories identified by our simple workflow. For instance, we found that local DNA shape and certain genome accessibility features can differentiate SV classes. Subsequently, we trained active-learning-based unsupervised machine learning models to exploit these discriminative features, going beyond sequence homology to refine a given SV's underlying repair mechanism class. Next, we applied these models to rare SVs from the high coverage 1KG project to recapitulate the established relationship between recombination rates and repair mechanisms in an ancestry-specific manner. Finally, a methodical comparison of *de novo* SVs from rare diseases and the 1KG dataset revealed differential levels of the H3K27me3 histone modification. Overall, our work develops and applies new methods to provide extensive insights into SVs and their mechanisms of introduction.

Materials & methods

Relevant terms and abbreviations provided (Supplementary Table 1). We utilized resolved germline SVs from the HGSVC2, 1KG, and 1KG-ONT projects in this work. All these studies employed the hg38 reference genome to identify variants. We note that sequencing, assembly algorithms, and variant calling strategies were project specific. HGSVC2 serves as a resource of haplotype-specific SVs identified with Phase Assembly Variant (PAV) [13] and merged across 35 individuals sequenced with PacBio (unique SVs: $n_{\text{INS}} = 65\,836$; $n_{\text{DEL}} = 40\,496$). The pilot 1KG-ONT study identified SVs from Oxford Nanopore sequencing involving 100 individuals. The SVs were merged using Jasmine [83] and subject to five distinct SV calling strategies (unique SVs: $n_{\text{INS}} = 62\,929$, $n_{\text{DEL}} = 49\,435$). Finally, the 1KG study contains SVs from 3202 individuals sequenced with Illumina, which includes 602 trios. We used the high-confidence variants generated by integrating SVs using GATK-SV [84], svtools [85], and Absinthe (<https://github.com/nygenome/absinthe>; unique SVs: $n_{\text{DEL}} = 57\,243$). For HGSVC2 and 1KG-ONT, we utilized both insertions and deletions, while for the 1KG project, we only used deletions to mitigate quality issues from assembling insertions from short-read sequencing.

Building a homology-based workflow

We initially labeled SVs solely based on their local homology. Using the reference genome, we extracted 1975 bp sequences upstream and downstream of the breakpoints, which were concatenated to the first or last 25 bp of the SVs, respectively. Next, we performed a gapless alignment between these two sequences with a word size of 5 and a minimum identity of 80% using BLAST. We required alignments to have opposing orientations and to include ± 2 bp of the breakpoint. In this way, we considered alignments that captured plausible repair events via the introduction of the SV. Finally, we selected the alignment result with the highest bit score. Our

goal was to create SV classes reflective of DSB repair pathways, including HR, SSA, aEJ, and NHEJ [46–49]. Therefore, we classified SVs with long alignments (a minimum length of 100 bp out of 2000 bp) and high-quality alignments (a minimum identity of 90%) into the high local homology (HLH) class, inspired by HR. In contrast, we designated SVs to the intermediate local homology (ILH) class if they had shorter alignments (length between 20 and 100 bp) and slightly lower quality alignments (minimum identity of 80%). Next, we labeled variants for which BLAST could not identify homology as having no local homology (NLH), representative of SVs associated with NHEJ. Lastly, we categorized SVs where homology was recognized but did not meet the HLH or ILH thresholds as undefined. To describe the concordance between this workflow and Breakseq v2 [25], we assigned HGVC2 SVs using both methods.

To demonstrate the efficacy of this workflow, we utilized somatic SVs from cancer cohorts. We hypothesized that since extreme variant patterns are not anticipated within the germline variation profile of a healthy genome, utilizing samples with known deficiencies or proficiencies in relevant repair pathways would provide a better proxy for evaluating our framework. Therefore, we obtained SVs sequenced using Illumina and called with Manta [86] from the 100KG Project [50]. We retrieved mutational signature contributions previously computed using COSMICv2 [87] in tandem. We established cohorts to represent positive and negative controls based on evidence of HR deficiency (HRD) or proficiency (HRP) and mismatch repair deficiency (MMRD) or proficiency (MMRP), as their mechanisms differ in their utilization of homology. We calculated the proportion of deletions assigned to NLH within a sample. We considered the workflow successful if HRD and HRP samples exhibited a noticeable shift in %NLH, while MMRD and MMRP samples did not. We selected breast invasive carcinoma (BRCA) and colon adenocarcinoma (COAD) because of their higher rates of HHRD and MMRD, respectively [51], where we utilized signatures 3 and 6, respectively. We defined a sample as deficient in a repair mechanism if the relevant signature's contribution was above the 95th percentile within the cancer study. Conversely, samples with contributions of 0 in the appropriate signature were considered “proficient.” Consequently, we identified 152 samples with HRD, 909 with HRP, 103 with MMRD, and 155 with MMRP.

Active-learning-based unsupervised clustering strategy

We aimed to develop a model that better capitalizes on the features underlying DSB repair, extending beyond just local homology. Therefore, we annotated SVs, their breakpoints, and their 2000 bp upstream and downstream flanking sequences with features including sequence characteristics, epigenomics, replication timing, repetitive elements, non-B DNA structures, and DNA shape. Given the chromosome and length biases across these features, we generated 100 simulations using SURVIVOR [54], which matched the real SVs in type, chromosome, and length. Subsequently, we annotated both the observed and simulated SV datasets.

The sequence features we calculated for the SVs and flanks included means and standard deviations in %GC content, stability, flexibility, and complexity (Shannon's entropy with 10 bp sliding windows). When annotating germline SVs, we used data from H1-hESC as a proxy due to its spatial

and temporal similarities in introducing SVs in the germline [53]. We estimated epigenomic profiles via Chip-Seq (CTCF, H2AFZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K79me2, H3K9ac, H3K9me3, and H4K20me1), DNase-seq, and WGB-seq experiments provided by ENCODE. We downloaded datasets from the ENCODE portal [88] (<https://www.encodeproject.org/>) with the following identifiers: ENCBS718AAA and ENCBS111ENC. We calculated the mean and standard deviation of the signal for the flanks (insertions and deletions) and the deleted regions using pybigwig (<https://github.com/deeptools/pyBigWig/>). We used a 16-stage Repli-seq experiment [89] from the 4D Nucleome Data portal [90] to quantify the replication timing at SV breakpoints. This data was further processed using the RepliSeq package [91], which provides an S50 value. S50 represents the S phase fraction, between 0 (early) and 1 (late), indicating the fraction of cells in a region that have undergone replication. To describe repetitive elements, we counted the number of LINEs, low complexity regions, LTRs, satellites, simple repeats, and SINE motifs in the flanks and SV, as determined by RepeatMasker. Similarly, we counted the number of non-B DNA structures (G-quadruplexes, Z-DNA, A-phased repeats, direct repeats, mirrored repeats, and short tandem repeats) within the flanking sequences using the nonB-GFA tool [92]. Finally, we calculated the mean and standard deviation of DNA shape (minor groove width (MGW), helical twist (HelT), propeller twist (ProT), roll, and electrostatic potential (EP)) within the SV and flanking sequences using DNASHapeR [93]. Briefly, the DNA shape features produced by DNASHapeR were derived from pentamer nucleotides that underwent all-atom Monte Carlo (MC) simulations. This pentamer dataset was compiled by conducting MC simulations of 2121 distinct DNA fragments of varying lengths (12–27 bp) sourced from DNA and protein-DNA complex structures available in the Protein Data Bank. Upon completing the annotations, we matched the real SVs to their simulations to calculate a z-score normalized value for all the features. These z-scores thus describe the extremeness of each real SV's feature relative to their simulations. Subsequently, we applied these normalized features to train an unsupervised model using an active learning approach.

To begin, we developed insertion and deletion models independently. We then randomly withheld SVs in chromosomes 3, 5, 14, 17, 20, and 22 from HGVC2 during training. We started with all possible flanking, breakpoint, and SV feature annotations to train our model. Subsequently, we reduced the number of features to include only those that displayed statistical significance across any workflow labels based on an ANOVA test. Furthermore, we retained only one of two features if they had correlations exceeding 0.95 to avoid redundancy and to reduce model complexity. We also performed hyperparameter tuning to obtain optimized parameters for our unsupervised clustering algorithm, producing clusters that align but do not strictly require homology thresholds. We chose Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [59] to maintain the gradient-like relationships observed when comparing HLH, ILH, and NLH, while effectively handling noise and density variations and alleviating the need to select the number of clusters. We optimized various parameters, including the number of dimensions for reduction using PCA (2, 5, 10, 20), the minimum number of points in a cluster (300, 500, 700, 900, 1100, 1300), and the distance metric (Bray-Curtis, Canberra, Chebyshev, Euclidean, Manhattan). We identified

“high quality” points assigned to each cluster for every combination of parameters. These points are defined as those with a cumulative confidence higher than 0.5 (i.e. not noise), and those where the difference between the highest and second-highest cluster assignment probability exceeds 0.25 (i.e. SV is confidently assigned to a specific cluster). For each parameter combination we considered, we required that at least 50% of the SVs from each workflow class be classified as “high quality.” In tandem, we assigned clusters a pseudo-label based on the greatest proportion of the workflow label represented (i.e. HLH, ILH, NLH, Undefined). Using the workflow labels and their assigned pseudo-labels, we applied Fisher’s Exact tests between the clusters with the greatest proportion of HLH and NLH SVs. Our optimal parameter combination yielded the greatest odds ratio (OR). Together, these requirements ensured large-scale and high-quality clustering while leveraging but not explicitly requiring homology as the sole feature.

Next, we used each cluster’s subset of “high quality” points to train a simple k-nearest neighbor (KNN) model using Euclidean distance. We chose this approach since HGSVC2 SVs are considered the highest quality; therefore, we assumed the patterns underlying these clusters would be the most reliable and discriminative. To find the optimal number of neighbors to consider, we used the following expressions,

$$kvalues = \{k | k = k_{start} + 10n, n \in N, k_{start} \leq k \leq k_{end}\};$$

$$k_{start} = \sqrt{\#SVs} - 50 \text{ and } k_{end} = \sqrt{\#SVs} + 50$$

We initialized with $\sqrt{\#SVs}$ and used a step size of 10, up to 50 (represented by n). We conducted 6-fold cross-validation to identify the optimal $kvalue$, using distinct subsets of chromosomes in each fold. Since the mean accuracy remained high across all the folds, we selected the $kvalue$ with the least variation in the accuracies. After training an insertion and deletion model, we froze the optimal PCA preprocessing and KNN models, allowing them to be applied to independent datasets without retraining. Following clustering, we considered the z -scores of the alignment lengths. For each cluster, we defined homology-driven (HD) and nonhomology-driven (nHD) sub-clusters to discriminate between SVs with z -scores above and below 2, respectively. We then applied these models and sub-cluster classifications to the held-out HGSVC2 SVs and SVs from 1KG-ONT and 1KG.

Downstream analyses

Beyond building a homology-based and an unsupervised approach for assigning DSB repair mechanisms to label SVs, we performed a series of downstream analyses to characterize the distinct repair classes. For instance, we annotated SV breakpoints using the 18-state H1-hESCs Chrom-HMM [52] annotations to understand epigenetic state differences among various homology-based SV categories. Furthermore, we calculated the standardized residuals (r) derived from χ^2 tests to quantify the difference in expected contributions. We also calculated the mean enrichment score (ES) as a complementary approach to quantify enrichment or depletion patterns of SVs overlapping with functional genomic elements, relative to SV simulations. Using the observed and matching simulated SVs, we used bedtools [94] to intersect insertions’ breakpoints or the deleted regions (minimum of 1bp) with several genomic features. As such, SVs were intersected with features including coding regions (CDS), candidate *cis*-regulatory elements (cCREs) from SCREEN [95] (including proximal and

distal enhancer-like signatures), topologically associated domains (TADs) [96], T1 and T2-lamina associated domains (LADs) [97], common and rare fragile sites [98], and regions mapping to > 95% Segmental Duplications. We selected this collection of elements to strategically assess the relationships between global negative selection pressures and DSB repair strategies. With these features, we calculated a fraction, then an enrichment score (ES) and empirical P -values;

$$fraction = \frac{\#SVs \text{ with } 1bp \text{ intersection within feature}}{Total \text{ SVs}}$$

$$ES = \frac{\sum_{i=1}^{100} \frac{Fraction_{Real} - Fraction_{th, simulation}}{Fraction_{th, simulation}}}{100}$$

$$\text{if } ES > 0, p - value = \frac{N_{(Fraction_{Real} \geq Fraction_{simulations})} + 1}{101}$$

$$\text{if } ES \leq 0, p - value = \frac{N_{(Fraction_{Real} < Fraction_{simulations})} + 1}{101}$$

Given that ES is a direct comparison between observed and simulated SVs, an $ES > 0$ represents an enrichment of real SVs within a genomic element. In contrast, an $ES < 0$ represents a depletion pattern. ES was calculated by dataset and label, along with an agglomerated ES (“All together”). We used a similar strategy to assess gene sets, including house-keeping genes and pathways from Reactome [99]. In this case, we aimed to probe potential relationships between gene function and DSB repair choice, whereby the selected pathways are alluded to in Ebert et al. [13].

We then used the previously published machine learning model, Orca [61], to investigate the role of repair mechanisms in genomic folding. We filtered for HGSVC2 deletions exceeding 4000 bp at a 1 Mb resolution, where one pixel represents 4000 bp. The results included predicted wildtype (wt) and mutant contact maps, representing pairwise interactions by $\log_e(fold \text{ over distance based background})$. We used the matching wt and mutant 0.5 Mb up and downstream contact maps when measuring change. We employed the structural similarity index measure (SSIM) and absolute Spearman’s correlation coefficient as small- and large-scale change-to-folding metrics, respectively [62]. SSIM is a measure to describe structural changes as per the perceived similarity of two contact maps. In contrast, Spearman’s correlation can quantify changes in the intensity rank and can further be processed to indicate the A/B compartment. We further demonstrated changes in folding by selecting two SVs of similar length and in proximity, chr8-7194444-DEL-8515 and chr8-7283532-DEL-15242. We calculated the difference in 3D interactions by directly comparing their wt and mutant contact maps.

To strategically uncover DSB repair choice inclination in the genome, we investigated SV hotspots by cluster label. We filtered for rare SVs ($AF < 1\%$) from the 1KG project to dissociate regional repair bias from common SVs. For each individual, we counted the number of CI_{HD} and CII_{nHD} SVs within a 1Mb window across the genome. For each population, we calculated a “hotspot threshold” using the 90th percentile from the SV counts, since baseline variation differs by population. For each population, window, and subcluster class, we then counted the number of individuals who met this SV count threshold, therefore describing individuals with an abundance of SVs in a specific window in a cluster-specific manner. We used the CI_{HD} and CII_{nHD} cluster values to conduct a Fisher’s Exact test for each population and window. An $OR > 1$ indicates a window enriches CI_{HD} SVs, while an $OR < 1$ indicates an enrichment of CII_{nHD} SVs. Finally, we applied a

genome-wide Benjamini/Hochberg P -value adjustment for the tests at a population level. For visualization, ORs exceeding five were capped to five.

We expect regions with rates of greater homology-driven repair to display high recombination rates. Therefore, we assigned population-specific recombination rates to the breakpoints of rare SVs from 1KG. For each individual and subcluster class, we calculated the mean recombination rate of their SVs and then determined a cluster and population-specific median recombination rate. To assess population structures, we calculated the absolute difference in median recombination rates for each pair of populations. Upon combining *CI* and *CII* subclasses, when two populations exhibit a large difference in recombination rates for a given class, it indicates greater divergence in the deployment of recombination. For visualization, the populations were organized according to the hierarchical clusters identified by Spence *et al.* [63].

To formally investigate the role of DSB repair choice and pathogenicity, we filtered the 1KG SVs based on those identified in probands. We created three groups of SVs from healthy individuals, which included common ($AF > 5\%$) and rare SVs ($AF < 1\%$) with evidence of inheritance, as well as rare SVs lacking evidence of inheritance, which are thus considered *de novo* for the proband. Next, we obtained *de novo* SVs from various rare disease cohorts in 100KG using the *de novo* variant dataset, filtering for deletions greater than 50 bp, along with variants that passed the *altreadparent*, *abration*, and *patch* filters. This resulted in cohorts of individuals with neurological and neurodevelopmental disorders (NND; $n = 209$), ultra-rare disorders (URD; $n = 37$), cardiovascular disorders (Cardio; $n = 24$), renal and urinary tract disorders (RUT; $n = 21$), endocrine disorders (Endo; $n = 20$), and other disorders ($n = 85$). Following the application of the KNN deletion model and z -score-based threshold, we compared the proportion of unique SVs assigned to *CI* and *CII* across these cohorts. Finally, we examined all the z -score-based features that underlie the cluster labels (i.e. *CI* and *CII*) across cohorts. Our goal was to determine whether there was evidence of a feature that could distinguish inherited from *de novo* SVs, as well as SVs from healthy and diseased individuals. We believe this highlighted the importance of studying underlying patterns in repair when considering repair mechanisms, modes of inheritance, and the impact of SVs. To achieve this, we calculated the absolute Cohen's d effect size across all z -score standardized features, comparing all cohorts and both clustering classes, *CI* and *CII*. From previous work, Cohen's d values of 0.5 and 0.8 indicate medium and large effect sizes, respectively [58]. To evaluate phenotypic differentiation, we focused on participants with URDs, given the notable pattern we observed in local H3K27me3 levels, along with their complex and multisystem phenotypes. We conducted Fisher's Exact tests for reported phenotypes by comparing individuals with *CI* and *CII de novo* SVs, along with the presence or absence (including 'No' and 'Unknown') of a documented phenotype.

Results

Application and validation of local homology-based workflow using several SV catalogs

In our work, we included SVs from two lrWGS (HGSVC [13] and 1KG-ONT [43] and one srWGS (high-coverage 1KG [44]

project) studies to capture germline variant diversity at the population level. Prior research has shown that local homology near breakpoints can influence repair choice [45]. Therefore, we extracted genomic regions corresponding to 1975 bp upstream or downstream of the SV, along with the first and last 25 bp within the SV breakpoint junction. Subsequently, we performed pair-wise sequence alignment between these regions. We only considered alignments with opposing orientations relative to the breakpoints and those that are proximal to the SV breakpoint. Finally, we assessed the alignment length and quality to categorize SVs into distinct homology groups (Fig. 1A). For example, SVs with high local homology (HLH) required a minimum alignment length of 100 bp and sequence identities above 90%. In contrast, intermediate local homology (ILH) required alignment lengths between 20 and 100 bp, with identities greater than 80%. In this context, the ILH class SVs resembled variants induced by SSA and aEJ pathways. While these pathways are functionally distinct, they remain largely understudied and difficult to differentiate using strict homology thresholds [46]; therefore, we grouped these SVs together. SVs could also exhibit no local homology (NLH) or be categorized as "undefined" if their length or alignment quality did not meet these thresholds. Our homology length and quality thresholds reflect experimentally derived values from proteins involved in HR, SSA, aEJ, or NHEJ [46–49]. Upon applying this workflow, we found that NLH represents the largest proportion of SVs, regardless of the SV dataset (Fig. 1B; Supplementary Fig. S1). For instance, we observed that NLH, ILH, HLH, and undefined classifications accounted for 52.6%, 26.0%, 7.6%, and 13.8% of unique insertions from HGSVC2, respectively. We found similar contribution patterns for 1KG-ONT insertions, classified as NLH (47.3%), ILH (27.1%), and HLH (5.3%) by our homology-based workflow. These proportions remained consistent among deletions. We found that the distribution of SV lengths among NLH, ILH, and HLH SVs was similar; however, Undefined SVs skewed toward longer length values (Supplementary Fig. S2).

We subsequently utilized the SVs from the HGSVC2 project to compare our results to another sequence-based SV mechanism tool, Breakseq2 [25]. Although Breakseq2 employs distinct homology metrics and incorporates additional repeat annotations, we observed a general alignment between the two methods. We found that 89.9% of HLH-labeled insertions were classified as NAHR, while 7.6% were classified as non-homologous recombination (NHR; Supplementary Fig. S3A). This pattern also applied to deletions, with 74.3% and 2.2% of HLH SVs categorized as NAHR and NHR, respectively (Supplementary Fig. S3B).

We found minor differences in the label assignment proportions when analyzing the lrWGS dataset (HGSVC2) and the srWGS dataset (1KG project). Therefore, we compared these datasets using SVs at an individual level, using individuals common to both studies ($n = 34$). We assumed HGSVC2 SVs are of the highest quality, as they were identified from haplotype-specific assemblies derived from lrWGS (PacBio) sequencing data. As expected, we found that among deletions, the per-individual proportion of HLH SVs in HGSVC2 ($\mu_{HGSVC2} = 5.85\%$) was significantly higher than in 1KG ($\mu_{1KG} = 1.11\%$; P -value = $1.59e-36$; Fig. 1C). In contrast, the NLH proportions per individual were significantly lower for HGSVC2 ($\mu_{HGSVC2} = 48.5\%$) compared to the 1KG dataset ($\mu_{1KG} = 60.87\%$; P -value = $1.58e-29$).

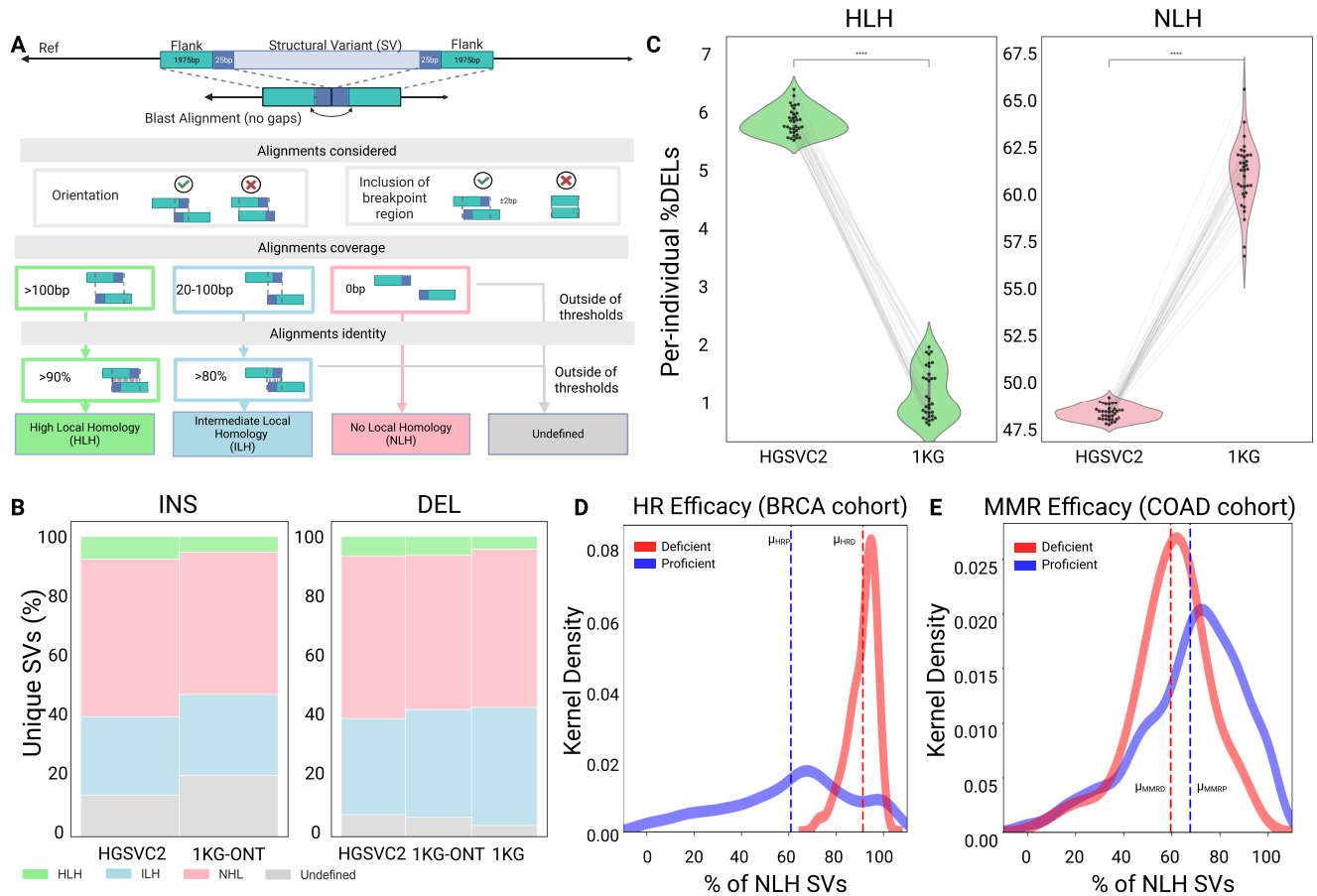


Figure 1. The homology-based framework to describe SV introduction as per DSB repair mechanisms. **(A).** Schematic of the homology-based label workflow. The 1975 bp upstream and downstream sequences local to deletions (depicted) and insertions are aligned with 25 bp of the SV without gaps using BLAST. Alignment, orientation, and the inclusion of the breakpoint are assessed. Quality and length determine the SV class. **(B).** Proportion of workflow labels applied to unique SVs in IrWGS (HGSVC2 & 1KG-ONT) and srWGS (high coverage 1KG phase 3) datasets. **(C).** Per-individual proportion of deletions labeled with the workflow when applied to individuals in both HGSVC2 and 1KG ($n = 34$). The proportions of individuals are connected with lines in the violin plots. (Paired t-tests). **(D-E).** The distribution of resolved somatic deletions (50 bp+) labeled as HLH among the 100KG project cohorts, visualized with kernel density estimate (KDE) plots. The selected cohorts demonstrate distinct patterns of repair to validate the workflow. The breast cancer (BRCA) group served as the positive control to monitor HR efficacy **(D)**, while the colorectal cancer (COAD) cohort served as the negative control to observe MMR **(E)**. Repair-deficient cohorts are shown in red, while repair-proficient cohorts are in blue. Dotted lines depict a cohort's mean %HLH.

To validate our overtly simple homology-based workflow, we utilized SVs from samples with known extreme repair patterns. Therefore, we constructed sample cohorts of short-read somatic SVs from the 100 000 Genomes (100KG) [50] project to demonstrate changes or the absence of change in the proportion of NHL SVs. We used breast and colorectal cancer samples (BRCA; COAD) to leverage HR and mismatch repair (MMR) proficiency, respectively [51]. We found that HR-deficient samples tend to have more NHL deletions (91.2%) than HR-proficient samples (60.8%; P -value = 2.3×10^{-29} ; Fig. 1D). In contrast, although still statistically different (P -value = 1.2×10^{-8}), the MMR-deficient (59.5%) and proficient samples (67.5%) exhibited a much smaller difference in NHL assignment (Fig. 1E). These findings illustrate how the workflow uniquely captures SVs introduced by homology-based repair mechanisms.

Genomic elements are uniquely enriched among homology-based workflow labels

Our homology-based workflow connects DSB repair mechanism patterns with SVs. However, our simple workflow does

not consider additional sequence and epigenetic features that likely influence DSB repair choice. These patterns are particularly interesting since previous studies exploring epigenetic differences among DSB repair strategies have revealed inconsistencies regarding whether certain DSBs in heterochromatin regions tend to be repaired via HR, aEJ, or NHEJ. Variations in experimental models and their inherent biases may explain these discrepancies [34]. Therefore, we utilized our extensive catalog of genome-wide SVs to better understand global genomic and epigenomic patterns in repair. To accomplish this goal, we applied the 18-state ChromHMM [52] annotations from H1-hESC (H1) cell lines as a proxy for the germline [53]. We found that upstream breakpoints of HLH deletions favor heterochromatin regions (standardized residual (r) $r_{HLH} = 6.5$; Supplementary Table 2A, Fig. 2A), while ILH breakpoints overlap with weak ($r_{ILH} = 11.7$) and strong transcription sites ($r_{ILH} = 7.4$). Furthermore, NHL breakpoints predominantly occupy epigenetically quiescent regions ($r_{NHL} = 10.6$) within the genome. We observed a similar pattern of association between chromatin states and homology profiles for flanking regions downstream to breakpoints in deletions and insertions (Supplementary Table 2B; Supplementary Fig. S4).

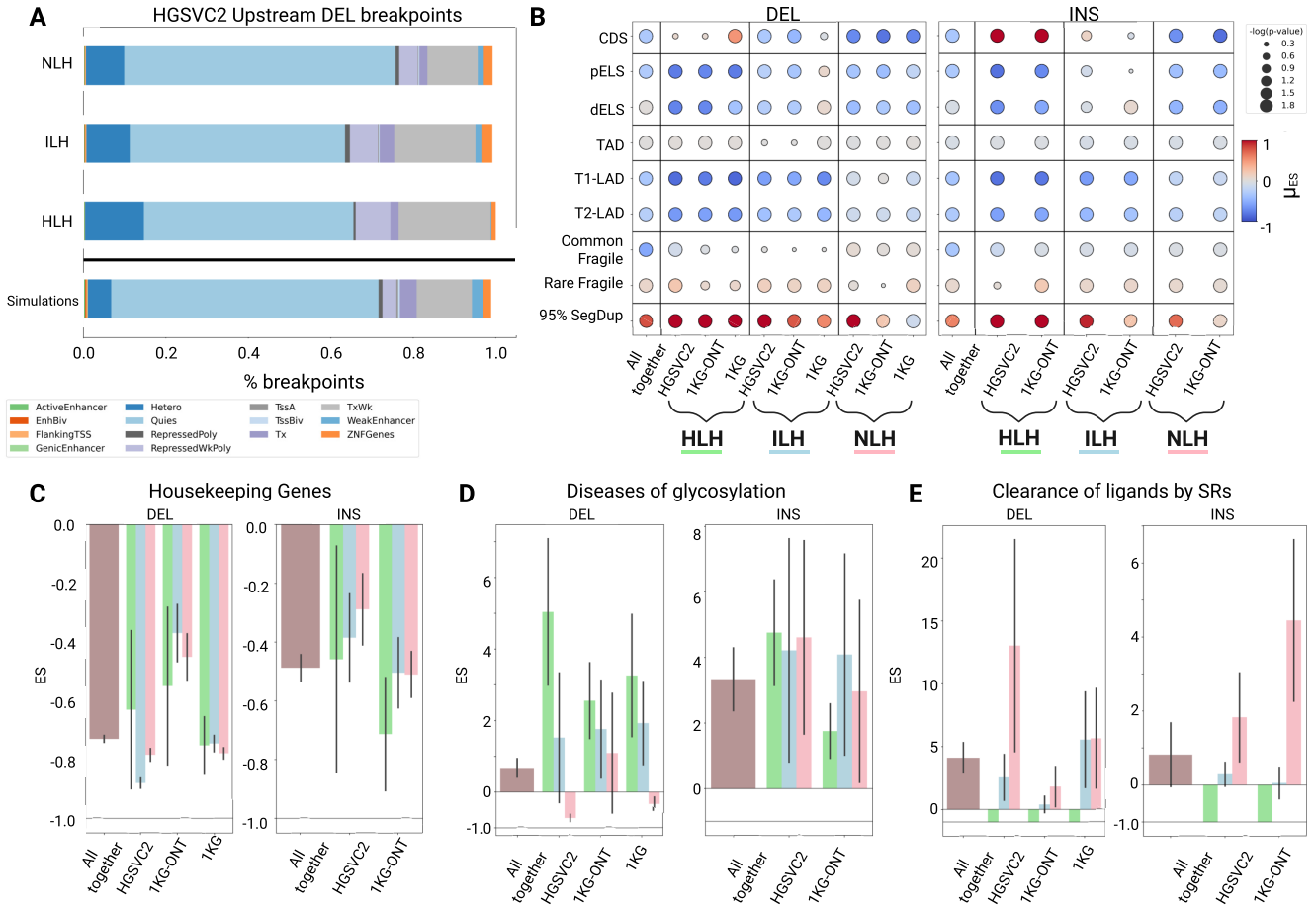


Figure 2. Enrichment of genomic elements, highlighted as per the homology-based workflow. **(A)** Stacked bar plots illustrate the proportion of upstream deletion breakpoints from HGSVC2 that fall within ChromHMM states. The bottommost bar represents these proportions across all simulations, while the bars above show the breakdown by label (HLH, ILH, NLH). **(B)** Enrichment and depletion patterns of SVs across genomic elements are depicted. Each marker indicates the magnitude, direction, and statistical significance of the mean enrichment score (ES). Magnitude and direction are differentiated by color, whereas statistical significance (represented as $-\log(\text{empirical } P\text{-value})$) is indicated by the marker size. The leftmost column demonstrates the aggregated ES among deletions and insertions, while the other columns are specific to workflow labels and datasets. The elements considered include coding regions (CDS), proximal and distal enhancer-like elements (pELS, dELS), topologically associated domains (TADs), T1 and T2 lamina-associated domains (LADs), common and rare fragile sites, and segmental duplications above 95% sequence identity. For visualization, the ES color intensity was capped between -1 and 1. **(C-E)** Bar plots illustrate the ES for deletions and insertions subject to CDS perturbation in gene sets. Mean ES (bar height and direction) and standard deviations (error bars) reveal the aggregated enrichment patterns, along with label and dataset specifics. Gene sets include housekeeping genes (HRT Atlas v1.0; **C**), Diseases of glycosylation (R-HSA-3781865; **D**), and binding and uptake of ligands by scavenger receptors (SRs; Reactome; R-HSA-2173782; **E**).

Beyond chromatin state differences, we examined the variability in selection pressures observed among distinct categories of SVs (based on the simple workflow classes) affecting various genomic elements. We generated 100 simulated SVs for each observed SV using StructUral Variant Majority Vote (SURVIVOR) [54] tool to quantify these selection pressures. Briefly, SURVIVOR utilizes a specified length, class, and reference genome to generate simulated SVs that match in length, SV type, and chromosome. We then calculated a mean enrichment score (ES) using the expected and observed SV profiles. Overall, we observed significant depletions of SVs overlapping with various functional elements in the genome (Fig. 2B), regardless of the homology-based label, aligning with prior studies [55]. However, we found variability in the magnitude and direction of these selection pressures among distinct SV categories. For example, among candidate *cis*-regulatory elements (cCREs) such as proximal enhancer-like signatures (pELS), the depletion strength for deletions in HGSVC2 is much stronger

among HLH SVs ($ES_{\text{HLH}} = -0.779$, $P\text{-value} = 0.009$) compared to ILH ($ES_{\text{ILH}} = -0.246$, $P\text{-value} = 0.009$) and NLH SVs ($ES_{\text{NLH}} = -0.345$, $P\text{-value} = 0.009$). This pattern was generally consistent across other genomic elements, including distal enhancer-like signatures (dELS), lamina-associated domains (LADs), and rare fragile sites. Interestingly, HLH insertions are enriched in segmental duplications (SD; $ES_{\text{HLH}} = 1.89$, $P\text{-value} = 0.009$). However, for deletions, there was no significant enrichment/depletion pattern in HLH SVs at CDSs ($ES_{\text{HLH}} = 0.08$, $P\text{-value} = 0.326$). Therefore, we found that the workflow-based labels align with paradoxes in the literature, where homology-based repair appears functional in both heterochromatin and euchromatin regions.

Considering the surprising enrichment pattern of HLH insertions overlapping with CDS elements, we investigated whether a gene's function could be driving this observation. Therefore, we quantified the enrichment score of SVs overlapping with specific gene sets, selected based on the

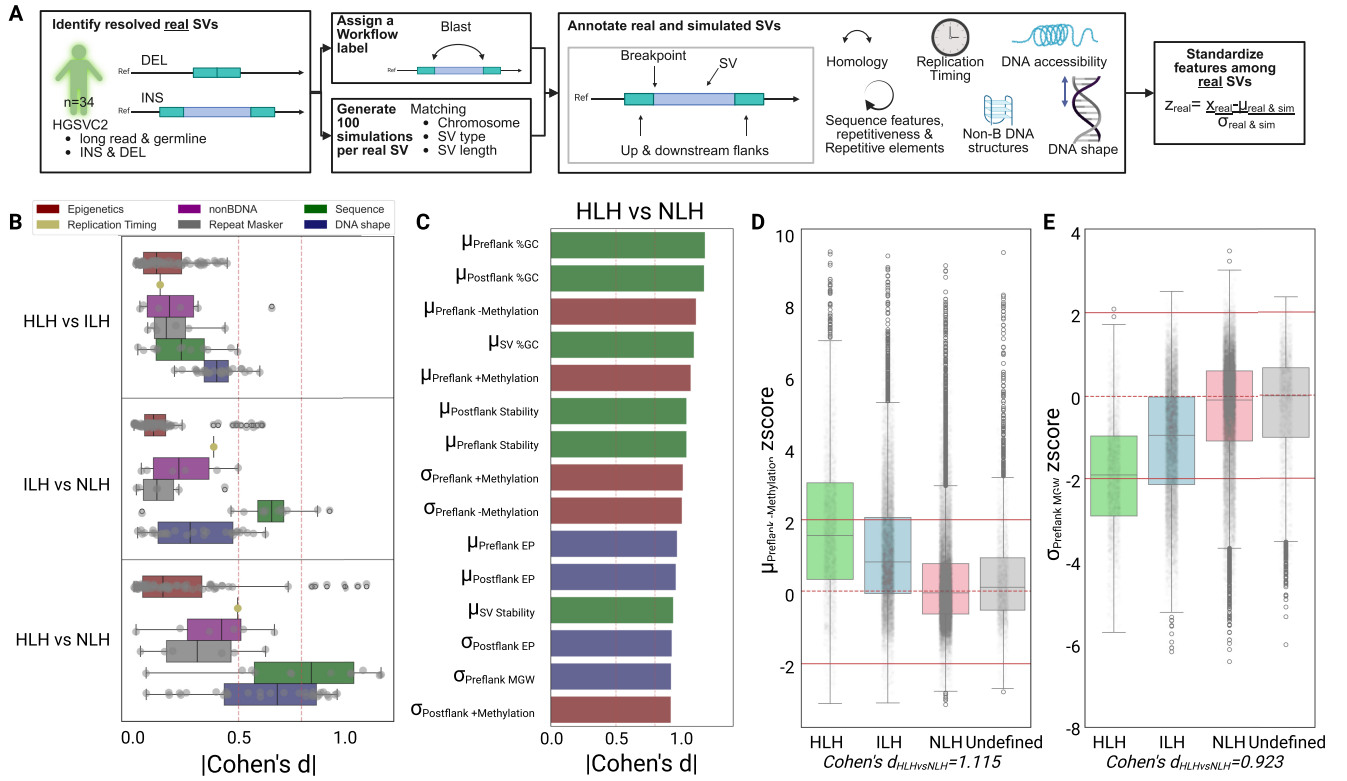


Figure 3. Probe into the underlying features that discriminate homology-based labels. **(A)** A schematic depicting the strategy to standardize SV, with 2000 bp flanks and breakpoint features. HGSC2 SVs are labeled with homology-based annotations. In tandem, 100 simulated SVs are generated using SURVIVOR for each SV. These simulated SVs match the observed SVs' chromosome, SV type, and SV length. Each observed and simulated SV is then annotated with features describing their local homology, sequence (including repetitive elements), replication timing, presence of non-B DNA motifs, epigenomics (derived from ChIP-Seq, DNA methylation, and DNAase-seq), and DNA shape. Finally, a standardized z-score is calculated for each feature. **(B-E)** Underlying patterns supporting HGSC2 deletions. **(B)** Boxplots depicting the absolute effect sizes of deletions. Each boxplot illustrates a pairwise comparison between two workflow labels for a feature type, with effect sizes measured using Cohen's d. Medium and large effect sizes are indicated with dotted lines at 0.5 and 0.8, respectively. **(C)** A barplot highlighting the top 15 most discriminatory features based on effect sizes between HLH and NLH. The color of the bar corresponds to the feature type. **(D-E)** Boxplots illustrating z-scores by workflow label according to the mean level of upstream DNA methylation **(D)** and the variation in upstream minor groove width (MGW; **E**). For SVs with z-scores close to 0, the distributions of the observed and simulated SVs for the given feature are similar, as depicted by a dotted line. Extreme z-scores, those above 2 or below -2, are shown with solid lines.

original HGSC2 study [13]. First, we quantified the enrichment pattern of SVs affecting housekeeping genes and expectedly found negative selection pressure, irrespective of the SV class (HGSC2 deletions: $ES_{\text{HLH}} = -0.629$; $ES_{\text{ILH}} = -0.879$; $ES_{\text{NLH}} = -0.783$; Fig. 2C). In contrast, we observed that HLH and ILH deletions are especially enriched in the diseases of glycosylation gene set (HGSC2 deletions: $ES_{\text{HLH}} = 5.03$; $ES_{\text{ILH}} = 1.52$; $ES_{\text{NLH}} = -0.721$; Fig. 2D), while ILH and NLH SVs are more likely to impact genes in the scavenger receptor binding and uptake pathway (HGSC2 deletions: $ES_{\text{HLH}} = -1.00$; $ES_{\text{ILH}} = 2.55$; $ES_{\text{NLH}} = 13.03$; Fig. 2E). However, when considering insertions at these receptor binding genes, the strength of this enrichment was weaker and less divergent for NLH SVs (HGSC2 insertions: $ES_{\text{HLH}} = -1.00$; $ES_{\text{ILH}} = 0.291$; $ES_{\text{NLH}} = 1.83$). While these results seek to connect repair mechanisms to gene function, the presence or absence of local homology may be an inherent characteristic of the surveyed genes. For instance, mucins in the glycosylation gene set are filled with variable number tandem repeats [56], while the variation in immune-related genes can be introduced via NHEJ [57]. Therefore, for these gene sets, there may be a built-in increase in the enrichment of HLH and NLH SVs, respectively.

DNA shape and epigenomics-associated features can discriminate between workflow labels

Our observations of differences in chromatin states and selection patterns for distinct homology-based SV classes indicate the role of additional genetic and epigenetic features that may facilitate DNA repair. Thus, we comprehensively characterized the features that underlie SVs, including sequence attributes, DNA shape, repetitive motifs, epigenetics (histone profile and DNA accessibility), and replication timing. We calculated the z-score of these features for upstream and downstream flanks, breakpoints, and within SV regions using the observed and simulated datasets (Supplementary Table 3, Fig. 3A). We obtained 135 and 109 features for deletions and insertions, respectively. Using the HGSC2 dataset and workflow-based labels, we employed Cohen's d effect sizes for pairwise comparisons (HLH vs. ILH, HLH vs. NLH, and ILH vs. NLH) to identify features contributing to their discrimination. Cohen's d measures the standardized difference between the means of two groups, where effect sizes of 0.5 and 0.8 are considered medium and large effects, respectively [58]. We observed small effect sizes for most features when comparing HLH to ILH (Fig. 3B & Supplementary Fig. S5). However, there was greater separation between HLH and NLH SVs for

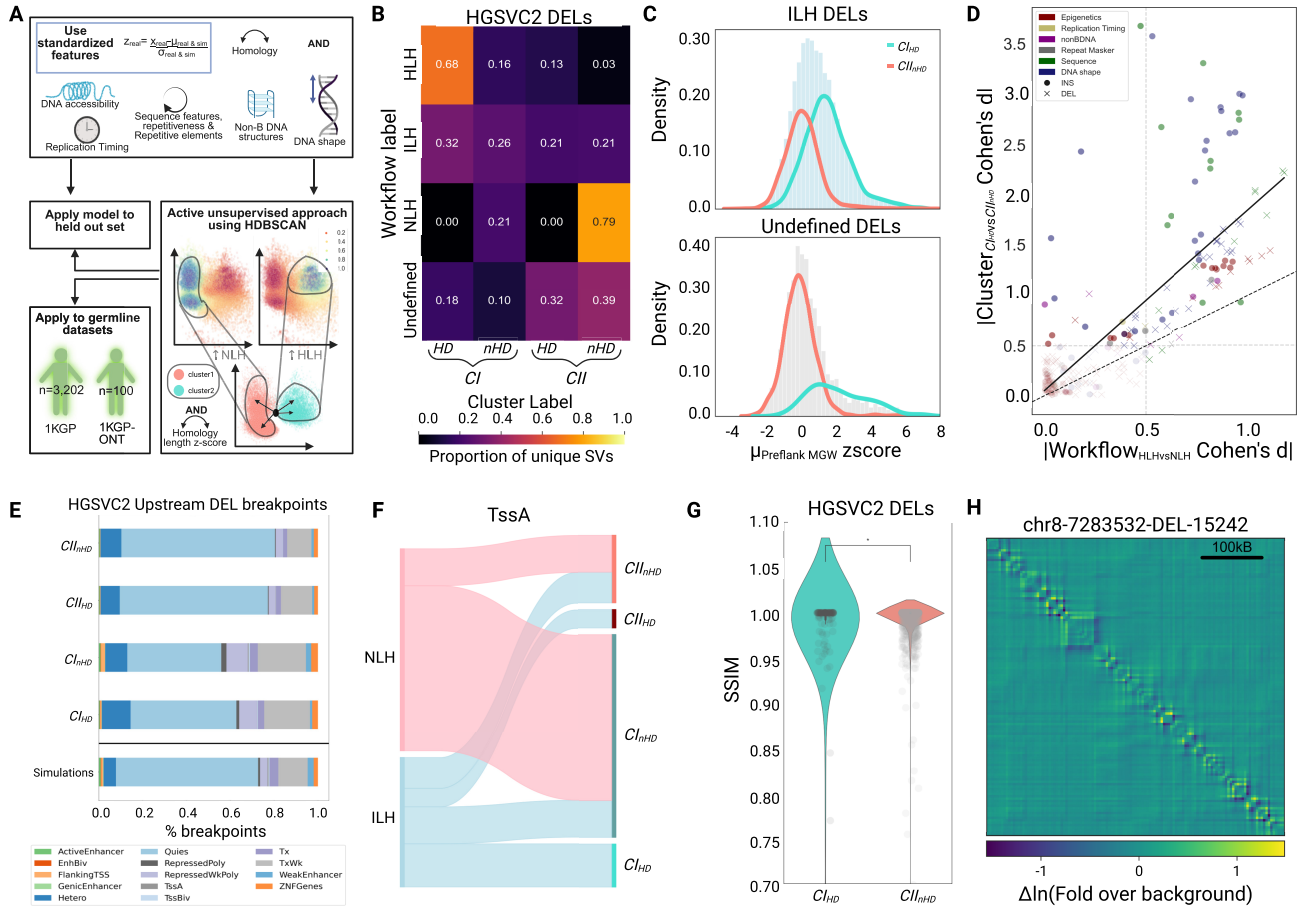


Figure 4. An active unsupervised clustering approach that exploits but does not strictly require thresholds for local homology. **(A)** Graphic of the active unsupervised clustering approach used to exploit features underlying SVs. SVs from HGSVC2 in chromosomes 3, 5, 14, 17, 20, and 22 are held out. An insertion and deletion model was independently developed. The z-score standardized features are filtered. Upon optimizing preprocessing and parameters for Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), clusters representing an enrichment in HLH and NLH SVs are identified. SVs with the highest quality cluster assignment were used to train a k-nearest neighbour (KNN) model. The models were frozen and applied to the held-out HGSVC2 SVs and SVs in 1KG and 1KG-ONT. Subsequently, SVs with z-scores above 2 in alignment length were labeled as homology driven (HD), while others were labeled nonhomology driven (nHD). **(B)** Heatmap visualizing the concordance between the homology-based and subclustering-based labels among the HGSVC2 deletions. Proportions are row-wise. **(C)** Histograms and KDE plots depicting the distribution of z-scores from the upstream mean minor groove width (MGW). The top plot depicts HGSVC2 deletions labelled as ILH, while the bottom plot contains Undefined deletions. On both plots, and for both workflow-based labels, KDE plots show the distribution of z-scores among the subcluster-based labels, CI_{HD} and CI_{nHD} . **(D)** A scatterplot comparing the effect sizes based on the workflow-based label (HLH versus NLH) and the cluster-based labels (CI_{HD} versus CI_{nHD}). A linear regression was fitted after excluding the homology-based features and is shown with a solid line. The dotted $y = x$ line serves as a comparison, demonstrating identical performance. Features with consistently low effect sizes, falling below 0.5 in both labeling strategies, are shown in grey. The remaining points are colored by feature type, and the shape of the markers corresponds to SV type. **(E)** Stacked bar plot depicting the breakpoints of upstream breakpoints from deletions falling within ChromHMM states according to the cluster-based labels. **(F)** Sankey plot illustrating insertions at ChromHMM's active transcription sites (TssA). The left side displays the homology-based label, while the right side corresponds to the cluster-based label. **(G)** Boxplots comparing the local changes in genomic folding based on the structural similarity index measure (SSIM) between CI_{HD} and CI_{nHD} SVs (Mann-Whitney). **(H)** The change in local DNA interactions with the CI_{HD} SV, chr8-7283532-DEL-15242. The 500 Mb wildtype and mutant windows downstream of the SV were directly compared.

features related to sequence, with a mean Cohen's d of 0.623. Closer examination revealed that the top 15 most discriminatory features between HLH and NLH deletions included %GC content, DNA methylation, local electrostatic potential (EP), and minor groove width (MGW; Fig. 3C). We visualized the average methylation of upstream flanking sequences and local MGW variation in deletions using homology-based labels (Fig. 3D and E). Notably, HLH SVs showed high normalized mean methylation and low normalized MGW variation. Additionally, the z-scores for NLH SVs were near 0, indicating similarity between observed and simulated SVs, suggesting a background repair mechanism. Furthermore, for most features, including methylation and MGW, undefined SVs had

z-scores similar to NLH SVs. These findings were generally consistent across insertions (Supplementary Figs S6 and 7). Altogether, our analyses suggest that a broader inventory of features may be utilized to differentiate between homology-based labels, thereby providing evidence that these features may support DSB repair.

Harnessing discriminatory features beyond homology to infer the mechanistic origin of SVs

Given that additional features beyond local sequence homology may describe patterns in repair choice, we trained an active-learning-based unsupervised model to classify SVs

(Fig. 4A). We reasoned that an unsupervised approach was most appropriate, as strict homology thresholds might be prohibitive for capturing the complete landscape of features that could contribute to an SV's introduction. Nonetheless, strict homology-based high-confidence data points can serve as pseudo-labels to guide the unsupervised model toward our biological expectations, thereby emphasizing the importance of local homology. In this approach, we first used hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [59] to perform high-quality SV cluster assignments. Subsequently, we quantified cluster enrichment in NLH and HLH SVs using the odds ratio (OR) and required that a substantial proportion of SVs be represented in the resulting clusters. Several models were evaluated by altering parameter combinations for feature dimension reduction, cluster size, and distance metrics. We then selected the model that reflected the highest OR, thereby identifying clusters with the most significant enrichment of NLH and HLH. Next, the high-quality SVs and their pseudo-label assignments were passed on to k-nearest neighbors (KNN) models. We trained separate models for insertions and deletions and utilized randomly held-out chromosomes (chr 3, 5, 14, 17, 20, 22) for model evaluation. Thus, our method employs additional underlying features while leveraging the original workflow-based (i.e. HLH, ILH, NLH) labels without imposing thresholds.

Optimal clustering for deletions involved reducing feature dimensions to five components, using Bray-Curtis distance, and a minimum of 700 SVs per cluster ($OR_{HLH \text{ vs } NLH} = 125.59$; Supplementary Table 4A). This approach utilized 121 features (Supplementary Table 5A) to generate two clusters. For insertions, we reduced the components to two, used the Canberra distance, and required 1300 SVs per cluster ($OR_{HLH \text{ vs } NLH} = 54.41$; Supplementary Table 4B), resulting in four clusters based on 95 features (Supplementary Table 5B). These clusters were then assigned pseudo-identities to train and validate a KNN model. The optimized KNN models for insertions and deletions used 118 and 96 neighbors, respectively. Next, we sub-categorized SVs within each cluster into homology-driven (HD) and non-homology-driven (nHD) based on high and low normalized homology-based alignment length scores, respectively. Overall, we generated four and eight sub-clusters for deletions and insertions, respectively.

The trained KNN models and normalized alignment length thresholding were subsequently applied to the held-out SVs from HGSVC2 (Supplementary Fig. S8), 1KG-ONT (Supplementary Fig. S9), and 1KG (Supplementary Fig. S10) datasets. We calculated the concordance between the homology-based workflow and clustering labels using HGSVC2 SVs (Fig. 4B & Supplementary Fig. S11). Among the deletions, we identified two clusters (*CI*, *CII*), and for insertions, we identified four clusters (*CI-IV*). In order, *CI*, *CII*, *CIII*, and *CIV* had the highest proportions of HLH, NLH, ILH, and Undefined SVs, respectively. Furthermore, we subdivided these four clusters into HD and nHD subcategories/sub-clusters based on the homology criterion. This method allowed us to maintain rules regarding the expectation of homology in repair (Supplementary Fig. S12) while capturing underlying patterns. For example, the HLH and *CI* groups both show high standardized mean upstream flanking %GC, whereas the NLH and *CII* groups show the opposite (Supplementary Fig. S13). Overall, we observed strong agreement between the workflow and clustering-based meth-

ods. For instance, our approach labeled 83.9% of HLH deletions and 78.9% of NLH deletions as *CI* and *CII*, respectively. Interestingly, most undefined deletions were assigned to *CII* (*CII_{HD}* at 32.4%, & *CII_{nHD}* at 39.3%).

Next, we examined specific cases where clustering improved the characterization of SVs compared to homology-based labeling. For example, the ILH deletions, chr1-248746661-DEL-59 and chr1-246387882-DEL-312, are in proximity to each other, with equivalent alignment length and quality (Supplementary Fig. S14AB). However, chr1-248746661-DEL-59 was assigned to *CI*, while chr1-246387882-DEL-312 was assigned to *CII*. A close inspection of their local contexts suggested that chr1-248746661-DEL-59 is nested within a direct repeat (DR) with high %GC and consistent MGWs. In contrast, only the downstream flank of chr1-246387882-DEL-312 falls within a DR in a low %GC region with fluctuating MGWs (Supplementary Fig. S14CD). This indicates that although these SVs share significant homology at their breakpoint junctions through clustering, their local contexts may also influence the differentiation of underlying repair mechanisms. Similarly, to explore other ambiguous cases, we analyzed the clustering of ILH and Undefined deletions. First, we examined the distribution of features such as the mean MGW of the upstream flank. Previous studies indicate that MRX/N, the complex that initiates DSB processing, can bind at the minor groove [60], and even favors DNA sequences with greater local MGW [40]. We observed that, expectedly, the local mean of MGW z -scores from HLH ($\mu_{MGW} = 1.94$), ILH ($\mu_{MGW} = 1.03$), and NLH ($\mu_{MGW} = 0.75$) deletions, decreases. We observed that following sub-clustering, ILH (*CI_{HD}* = 1.52, *CII_{nHD}* = 0.08), and Undefined (*CI_{HD}* = 2.14, *CII_{nHD}* = -0.11) deletions could be further partitioned using mean MGW z -scores (Fig. 4C). This was particularly evident, as the distribution of Undefined SVs MGW z -scores has a long right tail, which is subsequently labelled to the *CI* cluster. To confirm the generalizability of these improvements, we directly compared the effect sizes between HLH and NLH to those between the *CI_{HD}* and *CII_{nHD}* subclusters (Fig. 4D). Although some features were maintained or even decreased in their ability to differentiate the classes, excluding homology-based features resulted in a significant performance improvement ($\beta = 1.787$, P -value = 1.878×10^{-49} , $R^2 = 0.601$). We found that the clustering-based strategy enhanced the effect sizes of several features, including sequence, DNA shape, replication timing, and epigenetics, specifically for DNA methylation, chromatin openness (as measured by DNase-seq), and H3K27me3 features.

Subsequently, we recharacterized ChromHMM states using our clustering labels (Fig. 4E; Supplementary Fig. S15; Supplementary Table 6AB). We found that the upstream breakpoints of *CI_{HD}* deletions are enriched in weak transcription regions ($r_{CI_{HD}} = 12.1$ vs $r_{CII_{nHD}} = -15.2$) and weak repressed polycomb sites ($r_{CI_{HD}} = 9.6$ versus $r_{CII_{nHD}} = -14.0$). In contrast, *CII_{nHD}* SVs displayed very strong enrichment in quiescent regions ($r_{CI_{HD}} = -12.7$ versus $r_{CII_{nHD}} = 17.8$). Most interestingly, we found that bivalent transcribed transcription start sites ($r_{CI_{nHD}} = 7.5$) were uniquely enriched among *CI_{nHD}* SVs. We then examined the ChromHMM state of SVs, demonstrating inconsistency between the clustering and workflow methods. For instance, we found that NLH SVs in active transcription start sites (Fig. 4F) and those in repressed polycomb complex regions (Supplementary Fig. S16) were primarily as-

signed to *CI*. These findings indicate that the clustering-based strategy learns features beyond local homology when assigning cluster identities.

Given the striking enrichment of *CII* SVs in quiescent regions, we hypothesized that the breakpoints of *CII* SVs in these areas would cause minimal changes to 3D genomic folding. We employed a deep learning method, Orca [61], which predicts 3D genome contact maps based on genomic sequences. We generated mutant and wild-type contact maps using HGSVC2 deletions of length 4 Kbp or greater. Subsequently, we calculated the absolute correlation and the structural similarity index measure (SSIM) to quantify global and local folding perturbations, respectively [62]. We found no difference in global folding changes when comparing the effect of *CI_{HD}* and *CII_{nHD}* SVs on the 3D-genome organization (Supplementary Fig. S17A). However, we observed significant local changes (P -value = 0.0279; Fig. 4G; Supplementary Fig. S17B). We visualized the interaction differences between the mutant and wild-type maps to demonstrate these changes. We used a 0.5 Mbp downstream window from a *CI_{HD}* (chr8-7283532-DEL-15242; Fig. 4H) and *CII_{nHD}* (chr8-7194444-DEL-8515; Supplementary Fig. S17C-D) SV. These SVs were selected due to their similar lengths and proximity. Although we found no global rewiring in the 3D structure due to the SV, the *CI* SV introduces small-scale changes across the entire window. Together, this suggests a potential relationship between repair choices and alterations to 3D-genome organization.

Applications to population structure & de novo SVs from rare disease cohorts

We applied our unsupervised method to characterize and compare SVs with distinct underlying repair patterns across diverse populations. Given the vast number of ancestries included, we utilized deletions from the 1KG study for this analysis. We hypothesized that SV (*CI_{HD}* and *CII_{nHD}*) hotspots, shared across ancestries, may indicate regional biases for a particular repair mechanism, thereby explaining patterns of variation in this region. Since we did not expect significant deviations among different ancestry groups for common SVs, we only included rare SVs (Allele Frequency (AF) < 1%) in this analysis. To identify these hotspots, we divided each chromosome into non-overlapping 1 Mb windows for each individual and calculated the total number of deletions annotated as *CI_{HD}* or *CII_{nHD}* SVs. We then computed the 90th percentile for these counts within each ancestry group, using this as a threshold to determine how many individuals exceed this SV count for a given window. We quantified a window's association with the *CI_{HD}* or *CII_{nHD}* clusters by applying Fisher's Exact tests. We found very few consistent *CII_{nHD}* windows across all ancestries (Fig. 5A, Supplementary Figs S18–19). However, we observed many consistent *CI_{HD}* hotspots, typically located in sub-telomeric regions, gene-dense areas, and within segmental duplications, largely coinciding with regions of strong transcription.

Since it is well established that sub-telomeric regions are hotspots for recombination [63, 64], and we did not use recombination rate as a feature during clustering, we confirmed significantly higher recombination rates amongst *CI_{HD}* ($\mu = 3.65e - 0.08$) compared to *CII_{nHD}* SVs ($\mu = 1.68e - 0.08$; P -value = $1.06e-147$ Fig. 5B). Furthermore, we investigated recent population-specific dynamics in variation, such as geographical clines and migratory patterns, by utilizing the 1KG

population information [63]. We found no significant difference in recombination rates within subclusters (Fig. 5B); therefore, we combined the *HD* and *nHD* subclusters to investigate population-specific differences among *CI* and *CII* SVs. We found pairwise differences in recombination rates by cluster after calculating the median of a population's mean recombination rate. We visualized these differences using a previously determined population structure [63] and observed that population dynamics were pronounced only in *CI* (Fig. 5C). Most notably, the recombination rates of genomic regions affected by *CI* SVs revealed patterns that align with a north-south European cline and an east-west African cline, while no such associations were observed for *CII* SVs. We similarly observed that the recombination rates of *CI* SVs uniquely demonstrated Puerto Rican admixture. These results might indicate the potentially prominent role of regions affected by *CI* SVs in recent population divergence compared to those influenced by *CII* SVs. This further suggests a potential role for both homology and local sequence contexts in influencing recombination rates.

To support our investigation into repair and variation, particularly concerning human disease, we focused on mechanistic differences between common and rare SVs, including *de novo* SVs in healthy populations and rare disease cohorts. For this purpose, we utilized trios (parents and children) from the 1KG project, where we identified SVs that were common (AF > 5%) and inherited, rare (AF < 1%) and inherited, as well as *de novo* to a proband. Next, we compared cluster assignment frequencies and the underlying DSB repair mechanism patterns associated with germline *de novo* variants from individuals with rare diseases in the 100KG project [50]. The participants included those with neurological and neurodevelopmental disorders (NND; $n = 209$), ultra-rare disorders (URD; $n = 37$), cardiovascular disorders ($n = 24$), renal and urinary tract disorders (RUT; $n = 21$), endocrine disorders ($n = 20$), and other disorders ($n = 85$). We posited that since *de novo* variants can significantly contribute to clinical phenotypes, the frequency of the repair mechanisms and the underlying features of SVs may differ between *de novo* SVs from individuals with rare diseases and both *de novo* and inherited SVs from healthy individuals.

Interestingly, we found an increasing proportion of common (31%), rare (44%), and *de novo* SVs (59%; Fig. 6A) belonging to *CI* in the 1KG dataset. However, in most rare disease cohorts, the frequency of *CI* SVs was low, except for those with cardiovascular disorders (59%). Upon further examination of specific SVs, we discovered that the *CI_{nHD}* variant, chr13-94711392-DEL-185, was present in multiple individuals, most of whom exhibited abnormal developmental phenotypes. Notably, this SV is located within *SOX21*, a transcription factor known to regulate neuronal differentiation [65]; further, the SV perturbs pELS and CTCF sites (Fig. 6B). However, we note that this SV was not identified as a causal variant for the diseases in these individuals. We then compared the underlying patterns of repair across these cohorts. Upon calculating the effect size between the cohorts using z -score normalized features, we found that when using the H1-hESC as a proxy, the upstream level of the histone marker H3K27me3 is distinct among individuals with URD (Fig. 6C). We observed a pattern of decreasing effect sizes from inherited to *de novo* 1KG SVs. Since patients with URDs can display clinically complex phenotypes, we compared the reported phenotypes from the participants with *CI* and *CII de novo* SVs. We found that

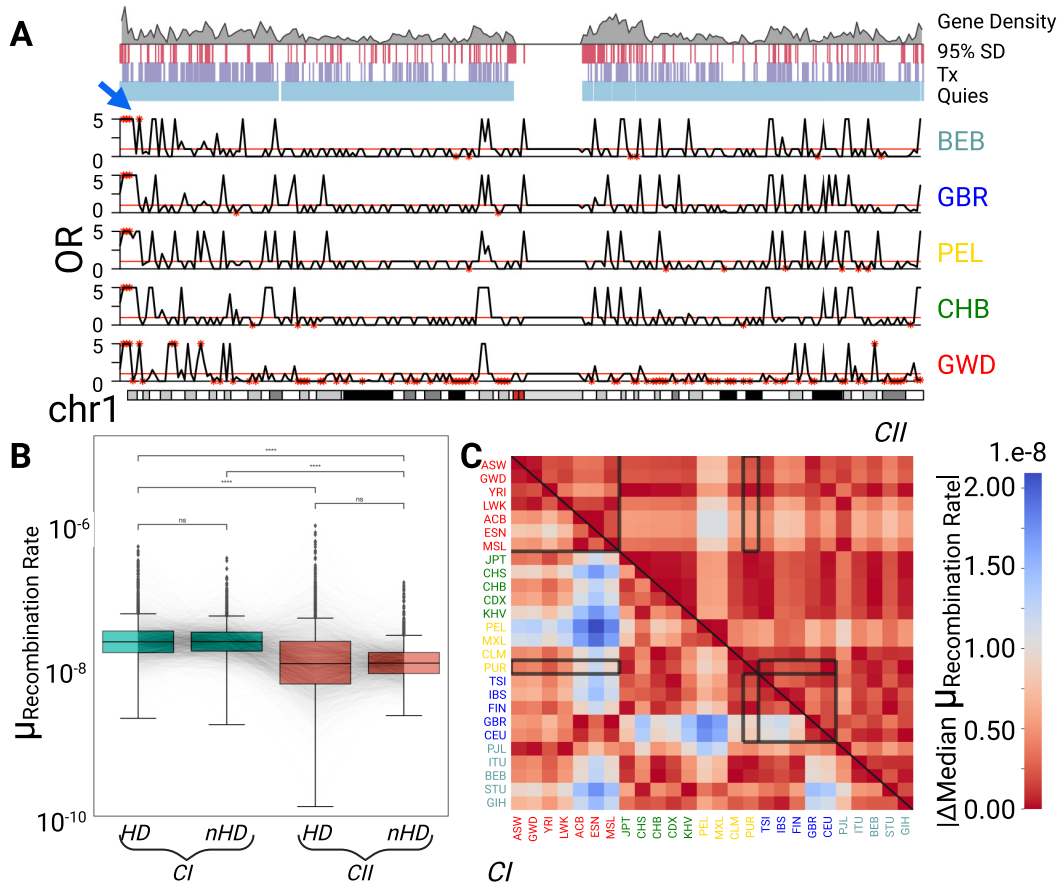


Figure 5. Application of clustering-based strategy to explore ancestry. **(A)** Karyoplot of SV hotspots across 1 Mb windows among rare deletions from 1KG on chromosome 1. The chromosome's ideogram is displayed at the bottom. Above this, using one population from each superpopulation, the odds ratio (OR) illustrating the enrichment of CI_{HD} SVs is indicated. Horizontal lines mark an OR of 1, showing no enrichment in CI_{HD} or CI_{nHD} SVs in the window. Furthermore, windows with statistical significance, after P -value adjustment, are represented with stars. An arrow points to windows showing consistent and significant hotspots across ancestries. Above these plots are genomic elements aligned to the chromosome. These elements include gene density, 95% segmental duplications (SD), as well as strong transcription (Tx) and quiescent (Quies) sites from ChromHMM. The populations included Bengali in Bangladesh (BEB) from South Asia, British in England and Scotland (GBR) from Europe, Peruvians from Lima (PEL) classified as admixed American, Han Chinese in Beijing (CHB) from East Asia, and the Gambians in Western Division- Mandinka (GWD) from Africa, in order. Text colors correspond to superpopulation. **(B)** Box plot comparing an individual's mean population-specific recombination rate among the four subclusters. Individuals' mean rates are connected by lines in the boxplots. (Paired t-tests). **(C)** Heatmap illustrating the magnitude of the pairwise differences in recombination rate. For each population, the median of the individual-level mean recombination rates was calculated for CI and CII SVs. For each population pair, the difference is visualized. The upper and lower triangles display the CII and CI differences, respectively. Lines in both the upper and lower triangles highlight the east-west African clines, Puerto Rican admixture, and the north-south European cline.

URD individuals with CI SVs displayed a non-significant enrichment in their abnormal genitourinary system and liver. In contrast, individuals with CII SVs have a non-significant enrichment in having abnormal skeletal and endocrine systems. We believe that larger sample sizes may further bolster clinical distinction between patients with URDs. This set of results highlights that while the repair mechanisms employed may not differ between healthy and diseased individuals, DNA accessibility associated features that contribute to effective repair may be altered, potentially reinforcing observed phenotypes.

Discussion

Recent advances in lrWGS and genome assembly algorithms have established accurate SV catalogs. Genomic consortia, including HGSC [13], Human Pan Genome Reference Consortium (HPRC) [66], AllofUs [67], and 1KG-ONT [43] projects,

are leveraging these technical advancements to describe SVs from diverse human populations. Beyond extending our understanding of how SVs influence genotype-phenotype relationships, these catalogs can be further utilized to investigate the mechanistic origins of SVs. In this work, we characterized the underlying repair mechanisms likely associated with the mechanistic origin of SVs in diverse human populations and some rare disease cohorts (Supplementary Table 7).

First, we applied a simple homology-based workflow to classify SVs into three categories (HLH, ILH, and NLH), inspired by DSB repair pathways (HR, SSA, aEJ, and NHEJ). We focused primarily on high-quality, *de novo* assembled haplotype-specific lrWGS SVs identified in the HGSC2 study. Additionally, we utilized SVs from orthogonal datasets from lrWGS (1KG-ONT) and srWGS (high-coverage 1KG project) platforms to systematically identify patterns in SVs. Our observation that most SVs belong to the NLH class is consistent with previous reports that NHEJ occurs at the highest fre-

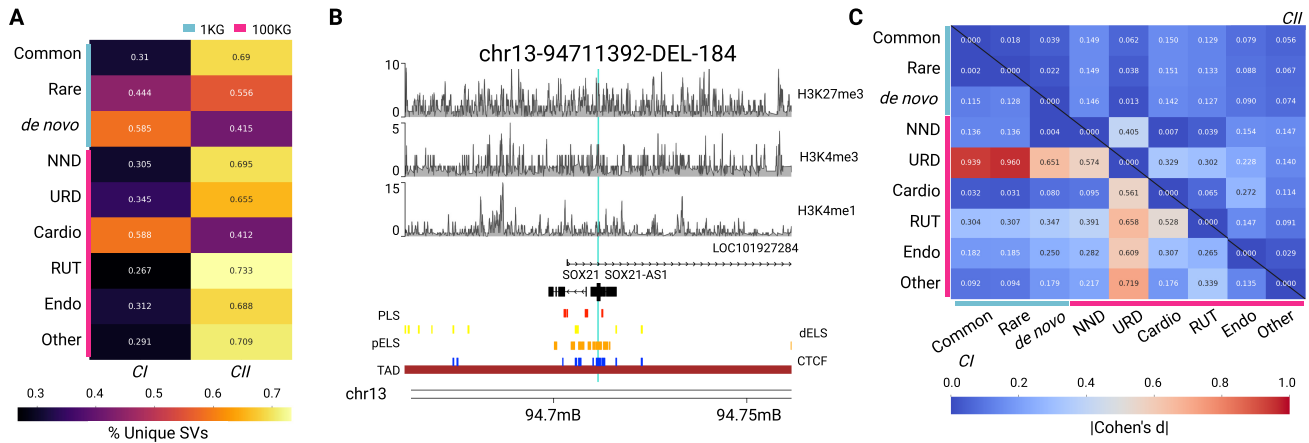


Figure 6. Application of clustering-based strategy to explore pathogenicity. **(A).** Heatmap showing *CI* and *CII* assignments across healthy (1KG) and diseased (100KG) cohorts, differentiated by turquoise and red bars. The healthy cohorts are represented by SVs from 1KG trios, including those that are commonly inherited, rare, or *de novo*. The diseased cohorts comprise *de novo* deletions from individuals with rare diseases, which include neurological and neurodevelopmental disorders (NND), ultra-rare disorders (URD), cardiovascular disorders (Cardio), renal and urinary tract disorders (RUT), endocrine disorders (Endo), and other disorders. Proportions are calculated row-wise. **(B).** Karyoplot of the *de novo* SV, chr13-94711392-DEL-184, found in multiple individuals across the 100 KG project cohorts. The chromosome's ideogram is located at the bottom. Above are genomic elements, including the presence of a TAD, CTCF sites, pELS, dELS, and promoter-like signatures, along with the CDSs that the SV affects. The epigenetic context associated with this SV is represented above, as per H3K27me3, H3K4me3, and H3K4me1 in H1-hESCs. **(C).** Heatmap illustrating the magnitude of the effect sizes across healthy and diseased SV cohorts for the local upstream levels of H3K27me3. The upper and lower triangles display the *CII* and *CI* differences, respectively.

quency, regardless of the cell cycle. However, we found that lrWGS can uncover more homology-driven SVs than srWGS due to its superior ability to detect SVs that span repetitive regions in the genome.

Subsequently, we quantified the propensity of SVs to affect distinct genomic elements and expectedly observed that most are under strong negative selection. However, we noted interesting deviations from this pattern for SVs affecting genes, specifically those belonging to certain pathways. For instance, we found that HLH/ILH and NLH SVs are significantly enriched for affecting genes involved in glycosylation, which include mucins, and the clearance of ligands by SR pathways, respectively. The relationship between homology-based SVs and mucin genes can be attributed to the high number of variable number tandem repeats present in these genes [56]. In contrast, the enrichment of NLH SVs affecting genes involved in the clearance of ligands by SRs (i.e. lipoproteins and immunoglobulins) may indicate an association between gene function and repair mechanisms. It is well established that variation in immunoglobulins is introduced through V(D)J recombination, primarily conducted by NHEJ [57]. However, these SVs are not expected in this germline context since V(D)J is a mechanism for introducing somatic variation. Therefore, this observation potentially points to a recognized technical artifact from sequencing lymphoblastoid cell lines (LCLs) characterized by a high rate of variants in IG loci [68]. Regardless, future studies should investigate the association between repair choice, the introduction of SVs, and genes with high repeat element variability.

Furthermore, we observed significant differences in DNA shape, epigenomic, and sequence feature profiles among the simple homology-based SV classes. For example, we found that HLH SVs exhibited less variation but greater flanking MGW and more pronounced local methylation. This suggests that these regions are more homogeneous and rigid with high GC content. We therefore postulate that a region's local con-

text may influence DSB repair effectiveness and/or efficiency, and/or detection [69]. Such contexts could be mediated by DNA accessibility and/or biophysical properties (for example, flanking MGW). These observations align with a prior study showing a higher DSB burden among Alzheimer's patients, where genomic loci affected by SVs had greater DNA accessibility [70].

We also noted significant differences in additional features underlying SVs belonging to distinct homology-based categories. Therefore, we trained an active-learning-based unsupervised model to classify SVs, which considered homology, DNA shape, epigenetics, replication timing, and sequence profiles (including non-B DNA, GC content, and repeat annotations) local to or within the SV. To achieve this, we applied and optimized the HDBSCAN algorithm, demonstrating improved discrimination across these features while preserving homology patterns captured by two distinct sub-clusters CI_{HD} (enriched among HLH SVs) and CII_{nHD} (enriched among NLH SVs). Next, we characterized the ChromHMM states of SVs in distinct sub-clusters derived from our unsupervised approach. Interestingly, we found that CI_{nHD} SVs are enriched in bivalently transcribed transcription start sites, which are relevant to neural development [71, 72]. The importance of bivalent states and SV introduction may align with paradoxes outlined in previous studies [32, 33, 42], whereby it remains unclear whether homology-based repair mechanisms exclusively prefer to engage in repair in heterochromatin or euchromatin. This may also suggest that homology-based repair mechanisms, such as HR, SSA, or aEJ, preferentially target "operational" regions of the genome rather than quiescent regions.

Furthermore, we observed that deletions in the CI_{HD} sub-cluster significantly disrupt local genomic folding compared to deletions in the CII_{nHD} subcluster. These findings align with previous work showing that homology-based repair is enhanced when homology searching is limited by physical

proximity [73–75]. Proximity may play a role in repair kinetics since some DSB repair processes benefit from chromatin mobility [73]; this can necessitate coordination between the repair machinery, the broken ends, and the nuclear periphery [76]. Therefore, we posit that upon a repair attempt, DSB breakpoints for homology-driven deletions may start with greater physical proximity. However, following the introduction of a deletion, the flanking regions may require more significant relocation or reorganization. Further experimental work will be necessary to verify DNA interactions before and after HR and NHEJ to elucidate repair kinetics.

Next, we utilized rare deletions from the 1KG project to investigate genomic regions characterized by a high density of SVs associated with the *CI* or *CII*. For this analysis, we focused on genome-wide SV hotspots that are consistent across individuals from distinct ancestries. Interestingly, we discovered that *CI_{HD}* SVs are enriched among genomic intervals primarily located in sub-telomeric regions. When considering sub-pathways of HR, sub-telomeres are often repaired using the error-prone Break-Induced Replication (BIR) pathway [64, 77], which is further relevant due to its clinical implications [69, 78], rather than the more accurate Synthesis-Dependent Strand Annealing (SDSA) pathway. Recent work has shown that pathogenic variants in the sub-telomeric region of chrX can be mediated through inverted repeats (which give rise to hairpins, a non-B DNA structure) and were likely introduced via BIR [69]. Moreover, fork stalling and template switching/ microhomology-mediated BIR (FoSTeS/MMBIR) mechanism has been linked to rearrangements found on chromosome 17 among individuals with Potoockli-Lupski microduplication syndrome (PTLS) [78]. Similarly, there are several disorders with etiological links to faulty NAHR, for example, 15q13.3 microdeletion syndrome [22], Charcot-Marie-Tooth disease type 1A [79], X-linked ichthyosis [80], and 7q11.23 deletions in Williams-Beuren syndrome [81]. Other relevant disorders, such as Fanconi anemia [20], which can cause defective HR, can present with phenotypes such as increased incidence of cancer, myelodysplasia, and congenital anomalies [82]. Therefore, clinically, there is justification for probing SV introduction within the germline as a strategy to appreciate disease manifestation and refine variant interpretation. Although most of the germline variants we analyzed were from healthy individuals, we characterized repair attempts for only a subset of DSB sites present upon sequencing that meet our SV length threshold (50 bp). Therefore, since our study relies on identifiable repair records, we are limited to imperfect repair outcomes. It is reasonable to assume that *CI_{HD}* SVs may be biased toward BIR instead of SDSA, given the location of the cluster's hotspots we identified, and that SDSA is unlikely to leave behind evidence of its usage. To mitigate this selection bias, higher throughput and more specific reporter assays can be applied to confirm the underlying features of repair in both perfect and imperfect repair scenarios, thereby capturing the entire spectrum of homology-based repair pathways and sub-pathways.

Finally, we applied our unsupervised method to classify common, rare, and *de novo* variants in healthy individuals (1KG project) and those with rare diseases (100KG study). Most *de novo* SVs in rare disease cohorts exhibited a higher frequency of *CII* SVs. However, we noticed that *de novo CI* and *CII* SVs in individuals with URDs could be distinguished based on their local levels of H3K27me3 signal compared to SVs from other rare diseases and healthy individuals. This ob-

servation is intriguing, considering that URDs are primarily monogenic and exhibit complex clinical presentations. This suggests that while SVs from patients with URDs may be epigenetically distinct, this finding likely does not play a critical role in disease etiology, but it remains noteworthy. With more samples, the relationship between DSB repair mechanisms and germline SVs could aid in classifying clinical phenotypes in rare diseases.

Overall, our study comprehensively characterized DSB repair mechanisms and their associated features underlying the mechanistic origins of SVs. In contrast to previous studies that primarily relied on large deletions from srWGS, our work utilized multiple population-level SV repositories from various lrWGS and srWGS efforts. Additionally, we considered various epigenomic, DNA shape, and sequence-related characteristics to gain insights into distinct classes of SV mechanisms. This study builds upon earlier computational efforts, which predominantly utilized sequence homology and repeat annotations. Despite several significant findings, we acknowledge certain technical limitations in our study. For instance, we obtained 1975 bp upstream and downstream sequences based on the reference genome when calculating local sequence homology. While effective, this approach overlooks potential haplotype-specific variation. Consequently, locally assembled and consensus sequences could improve the accuracy of these alignments. Furthermore, our analyses concentrated solely on insertions and deletions. We note that we characterized inversions within the HGSC2 and 1KG-ONT datasets. Although given the small sample sizes and homogeneity in their threshold-based label (ie. almost all were labelled as NLH), we could not draw comparisons across SV classes and deploy an unsupervised model. Future computational work should broaden the range of SV types to include complex SVs. Additionally, we employed *de novo* SVs from the 1 KG project to characterize and compare the usage of DSB repair mechanisms in healthy individuals and those with rare diseases. However, it is presumed that the rate of *de novo* variants from the 1 KG project is unexpectedly high [44], at approximately 10 times the expected rate, with approximately 100 *de novo* mutations per child. This indicates that the origin of the *de novo* SVs in this dataset reflects another technical artifact from the LCLs used during sequencing, rather than an accurate representation of mechanisms deployed in the germline or during development. Finally, we used H1-hESCs as a proxy for epigenetic and replication timing annotations of SVs [53]. This assumes that the chromatin profiles of H1 cells are well conserved and exhibit minimal dynamics during the acquisition of SVs. To determine if patterns of DNA accessibility and repair choice have a correlational or causative relationship, future studies should enhance this approach by employing multi-omics data to investigate the epigenomic states of regions proximal to SVs in healthy individuals and those with rare diseases. For example, this could include characterizing 3D genome folding, histone modifications, and DNA accessibility proximal to variants in affected and unaffected siblings using Hi-C and ATAC-Seq assays.

Despite the outlined limitations, our study presents the most thorough computational characterization of DSB repair mechanisms. Our study identifies underlying features that influence the mechanistic origin of SVs in diverse human populations. We anticipate that continued efforts to investigate features that support (or hinder) the introduction of SVs will help us accurately delineate their role in defining genotype-phenotype relationships.

Acknowledgements

Figures created or organized using BioRender. Boev, N. (2025) <https://BioRender.com/13b98ay>. This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care, and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK, and the Medical Research Council have also funded research infrastructure.

Author contributions: N.B. (Conceptualization [supporting], Data curation [lead], Formal Analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Visualization [lead], Writing – original draft [equal], Writing – review & editing [equal], M.G. (Conceptualization [equal], Resources [equal], Writing – original draft [equal], Writing – review & editing [equal], S.K. (Conceptualization [lead], Data curation [supporting], Formal Analysis [supporting], Funding acquisition [lead], Investigation [supporting], Methodology [equal], Project administration [lead], Resources [lead], Supervision [lead], Writing – original draft [equal], Writing – review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

S.K. acknowledges support from the Princess Margaret Cancer Foundation, Canada Research Chair Program, and Terry Fox Research Institute. N.B. acknowledges support from the Canadian Institute of Health Research, Cancer Research Institute (FRN: 201219). M.G. acknowledges support from the National Institute of Health (U24HG007497) and the AL Williams Professorship funds..

Data availability

The HGSVC2 and high-coverage 1000 genome phase 3 SV datasets were downloaded from the IGSR data portal (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/ & http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/).

Additionally, epigenomics and cCRE annotations (<https://screen.encodeproject.org/>) are available through the ENCODE project data portal. 1KG-ONT SVs are hosted on Amazon bucket (https://s3.amazonaws.com/1000g-ont/Gustafson_etal_2024_preprint_SUPPLEMENTAL/20240423_jasmine_intras). The 100KG project data utilized in this study can be accessed by logging into the Genomics England research environment.

Finally, relevant source codes for this study, including scripts for simulation, annotation, model training, and downstream analyses, can be accessed on the project GitHub page (https://github.com/kumarlab-compomics/SV_mechanism) and on Zenodo (<https://doi.org/10.5281/zenodo.17437994>).

References

- Kumar S, Gerstein M. Unified views on variant impact across many diseases. *Trends Genet* 2023;39:442–50. <https://doi.org/10.1016/j.tig.2023.02.002>
- Smedley D, Smith KR, Martin A *et al*. 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N Engl J Med* 2021;385:1868–80.
- Pang AW, MacDonald JR, Pinto D *et al*. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* 2010;11:R52. <https://doi.org/10.1186/gb-2010-11-5-r52>
- Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* 2021;31:2249–57. <https://doi.org/10.1101/gr.275488.121>
- Abel HJ, Larson DE, Regier AA *et al*. Mapping and characterization of structural variation in 17, 795 human genomes. *Nature* 2020;583:83–9. <https://doi.org/10.1038/s41586-020-2371-0>
- Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 2016;17:224–38. <https://doi.org/10.1038/nrg.2015.25>
- Merker JD, Wenger AM, Sneddon T *et al*. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* 2018;20:159–63. <https://doi.org/10.1038/gim.2017.86>
- Trost B, Thiruvahindrapuram B, Chan AJS *et al*. Genomic architecture of autism from comprehensive whole-genome sequence annotation. *Cell* 2022;185:4409–27. <https://doi.org/10.1016/j.cell.2022.10.009>
- Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var* 2024;11:18. <https://doi.org/10.1038/s41439-024-00276-x>
- Mahmoud M, Gobet N, Cruz-Dávalos DI *et al*. Structural variant calling: the long and the short of it. *Genome Biol* 2019;20:246. <https://doi.org/10.1186/s13059-019-1828-7>
- Jeong H, Dishuck PC, Yoo D *et al*. Structural polymorphism and diversity of human segmental duplications. *Nat Genet* 2025;57:390–401. <https://doi.org/10.1038/s41588-024-02051-8>
- Bolognini D, Halgren A, Lou RN *et al*. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* 2024;634:617–25. <https://doi.org/10.1038/s41586-024-07911-1>
- Ebert P, Audano PA, Zhu Q *et al*. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021;372:eabf7117.
- Miller DE, Sulovari A, Wang T *et al*. Targeted long-read sequencing identifies missing disease-causing variation. *Am Hum Genet* 2021;108:1436–49. <https://doi.org/10.1016/j.ajhg.2021.06.006>
- Kieffer SR, Lowndes NF. Immediate-early, early, and late responses to DNA double stranded breaks. *Front Genet* 2022;13:793884. <https://doi.org/10.3389/fgene.2022.793884>
- Ballinger TJ, Bouwman BAM, Mirzazadeh R *et al*. Modeling double strand break susceptibility to interrogate structural variation in cancer. *Genome Biol* 2019;20:28. <https://doi.org/10.1186/s13059-019-1635-1>
- Dileep V, Boix CA, Mathys H *et al*. Neuronal DNA double-strand breaks lead to genome structural variations and 3D genome disruption in neurodegeneration. *Cell* 2023;186:4404–21. <https://doi.org/10.1016/j.cell.2023.08.038>

18. Aguilera A, García-Muse T. Causes of genome instability. *Annu Rev Genet* 2013;47:1–32. <https://doi.org/10.1146/annurev-genet-111212-133232>
19. O'Driscoll M, Gennery AR, Seidel J *et al.* An overview of three new disorders associated with genetic instability: LIG4 syndrome, RS-SCID and ATR-Seckel syndrome. *DNA Repair (Amst)* 2004;3:1227–35.
20. Peake JD, Noguchi E. Fanconi anemia: current insights regarding epidemiology, cancer, and DNA repair. *Hum Genet* 2022;141:1811–36. <https://doi.org/10.1007/s00439-022-02462-9>
21. Sharma R, Lewis S, Wlodarski MW. DNA repair syndromes and cancer: insights into genetics and phenotype patterns. *Front Pediatr* 2020;8:570084. <https://doi.org/10.3389/fped.2020.570084>
22. Sharp AJ, Mefford HC, Li K *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* 2008;40:322–8. <https://doi.org/10.1038/ng.93>
23. Farmer H, McCabe N, Lord CJ *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 2005;434:917–21. <https://doi.org/10.1038/nature03445>
24. Pilić PG, Gay CM, Byers LA *et al.* PARP inhibitors: extending benefit beyond BRCA-mutant cancers. *Clin Cancer Res* 2019;25:3759–71.
25. Abyzov A, Li S, Kim DR *et al.* Analysis of deletion breakpoints from 1, 092 humans reveals details of mutation mechanisms. *Nat Commun* 2015;6:7256. <https://doi.org/10.1038/ncomms8256>
26. Lam HYK, Mu XJ, Stütz AM *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 2010;28:47–55. <https://doi.org/10.1038/nbt.1600>
27. Pang AWC, Migita O, Macdonald JR *et al.* Mechanisms of formation of structural variation in a fully sequenced human genome. *Hum Mutat* 2013;34:345–54. <https://doi.org/10.1002/humu.22240>
28. Ortega P, Gómez-González B, Aguilera A. Heterogeneity of DNA damage incidence and repair in different chromatin contexts. *DNA Repair (Amst)* 2021;107:103210. <https://doi.org/10.1016/j.dnarep.2021.103210>
29. Caron P, Pobega E, Polo SE. DNA double-strand break repair: all roads lead to HeterochROMatin marks. *Front Genet* 2021;12:730696. <https://doi.org/10.3389/fgene.2021.730696>
30. Crosetto N, Mitra A, Silva MJ *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* 2013;10:361–5. <https://doi.org/10.1038/nmeth.2408>
31. Akhtar W, de Jong J, Pindyurin AV *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* 2013;154:914–27. <https://doi.org/10.1016/j.cell.2013.07.018>
32. Schep R, Brinkman EK, Leemans C *et al.* Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. *Mol Cell* 2021;81:2216–30. <https://doi.org/10.1016/j.molcel.2021.03.032>
33. Clouaire T, Rocher V, Lashgari A *et al.* Comprehensive mapping of histone modifications at DNA double-strand breaks deciphers repair pathway chromatin signatures. *Mol Cell* 2018;72:250–62. <https://doi.org/10.1016/j.molcel.2018.08.020>
34. Vitor AC, Huertas P, Legube G *et al.* Studying DNA double-strand break repair: an ever-growing toolbox. *Front Mol Biosci* 2020;7:24.
35. Liu S-C, Feng Y-L, Sun X-N *et al.* Target residence of Cas9-sgRNA influences DNA double-strand break repair pathway choices in CRISPR/Cas9 genome editing. *Genome Biol* 2022;23:165. <https://doi.org/10.1186/s13059-022-02736-5>
36. Pallaseni A, Peets EM, Girling G *et al.* The interplay of DNA repair context with target sequence predictably biases Cas9-generated mutations. *Nat Commun* 2024;15:10271. <https://doi.org/10.1038/s41467-024-54566-7>
37. Hufnagl A, Herr L, Friedrich T *et al.* The link between cell-cycle dependent radiosensitivity and repair pathways: a model based on the local, sister-chromatid conformation dependent switch between NHEJ and HR. *DNA Repair (Amst)* 2015;27:28–39. <https://doi.org/10.1016/j.dnarep.2015.01.002>
38. Fortini P, Ferretti C, Dogliotti E. The response to DNA damage during differentiation: pathways and consequences. *Mutat Res/Fundamental Mol Mech Mutagenesis* 2013;743-744:160–8. <https://doi.org/10.1016/j.mrfmmm.2013.03.004>
39. Beerman I, Seita J, Inlay MA *et al.* Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle. *Cell Stem Cell* 2014;15:37–50. <https://doi.org/10.1016/j.stem.2014.04.016>
40. Gnügge R, Reginato G, Cejka P *et al.* Sequence and chromatin features guide DNA double-strand break resection initiation. *Mol Cell* 2023;83:1237–50.
41. Yin X, Liu M, Tian Y *et al.* Cryo-EM structure of human DNA-PK holoenzyme. *Cell Res* 2017;27:1341–50. <https://doi.org/10.1038/cr.2017.110>
42. Kendek A, Wensveen MR, Janssen A. The sound of silence: how silenced chromatin orchestrates the repair of double-strand breaks. *Genes* 2021;12:1415. <https://doi.org/10.3390/genes12091415>
43. Gustafson JA, Gibson SB, Damaraju N *et al.* High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* 2024;34:2061–73. <https://doi.org/10.1101/gr.279273.124>
44. Byrska-Bishop M, Evani US, Zhao X *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 2022;185:3426–40. <https://doi.org/10.1016/j.cell.2022.08.004>
45. Dumont A, Mendiboure N, Savocco J *et al.* Mechanism of homology search expansion during recombinational DNA break repair in *Saccharomyces cerevisiae*. *Mol Cell* 2024;84:3237–53. <https://doi.org/10.1016/j.molcel.2024.08.003>
46. Aleksandrov R, Hristova R, Stoyanov S *et al.* The chromatin response to double-strand DNA breaks and their repair. *Cells* 2020;9:1853. <https://doi.org/10.3390/cells9081853>
47. Setton J, Hadi K, Choo Z-N *et al.* Long-molecule scars of backup DNA repair in BRCA1- and BRCA2-deficient cancers. *Nature* 2023;621:129–37. <https://doi.org/10.1038/s41586-023-06461-2>
48. Blasiak J. Single-strand annealing in cancer. *Int J Mol Sci* 2021;22:2167. <https://doi.org/10.3390/ijms22042167>
49. Ceccaldi R, Rondinelli B, D'Andrea AD. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol* 2016;26:52–64. <https://doi.org/10.1016/j.tcb.2015.07.009>
50. The National Genomics Research Library v5.1, Genomics England. *Genomics England* 1.
51. Lee S-Y, Wang H, Cho HJ *et al.* The shaping of cancer genomes with the regional impact of mutation processes. *Exp Mol Med* 2022;54:1049–60. <https://doi.org/10.1038/s12276-022-00808-x>
52. Epigenomics Consortium R, Kundaje A, Meuleman W *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30.
53. Thomson JA, Itskovitz-Eldor J, Shapiro SS *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* 1998;282:1145–7. <https://doi.org/10.1126/science.282.5391.1145>
54. Jeffares DC, Jolly C, Hoti M *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* 2017;8:14061. <https://doi.org/10.1038/ncomms14061>
55. Collins RL, Brand H, Karczewski KJ *et al.* A structural variation reference for medical and population genetics. *Nature* 2020;581:444–51. <https://doi.org/10.1038/s41586-020-2287-8>
56. Fowler J, Vinall L, Swallow D. Polymorphism of the human muc genes. *Front Biosci* 2001;6:D1207–15. <https://doi.org/10.2741/Fowler>
57. Nicolas L, Cols M, Choi JE *et al.* Generating and repairing genetically programmed DNA breaks during immunoglobulin class switch recombination. *F1000Res* 2018;7:458. <https://doi.org/10.12688/f1000research.13247.1>
58. Lachenbruch PA, Cohen J. Statistical power analysis for the behavioral sciences. (2nd ed.). *J Am Statist Assoc* 1989;84:1096. <https://doi.org/10.2307/2290095>

59. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates, In *Advances in Knowledge Discovery and Data Mining, Lecture notes in computer science*. Berlin: Springer Berlin Heidelberg, 2013, pp.160–72.
60. Liu Y, Sung S, Kim Y *et al*. ATP-dependent DNA binding, unwinding, and resection by the Mre11/Rad50 complex. *EMBO J* 2016;35:743–58. <https://doi.org/10.15252/embj.201592462>
61. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* 2022;54:725–34. <https://doi.org/10.1038/s41588-022-01065-4>
62. Gunsalus LM, McArthur E, Gjoni K *et al*. Comparing chromatin contact maps at scale: methods and insights. *Biorxiv*, <https://doi.org/10.1101/2023.04.04.535480>. 4 April 2023, preprint: not peer reviewed
63. Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv* 2019;5:eaaw9206. <https://doi.org/10.1126/sciadv.aaw9206>
64. Batté A, Brocas C, Bordelet H *et al*. Recombination at subtelomeres is regulated by physical distance, double-strand break resection and chromatin status. *EMBO J* 2017;36:2609–25.
65. Whittington N, Cunningham D, Le T-K *et al*. Sox21 regulates the progression of neuronal differentiation in a dose-dependent manner. *Dev Biol* 2015;397:237–47. <https://doi.org/10.1016/j.ydbio.2014.11.012>
66. Liao W-W, Asri M, Ebler J *et al*. A draft human pangenome reference. *Nature* 2023;617:312–24. <https://doi.org/10.1038/s41586-023-05896-x>
67. Investigators AoURPG.. Genomic data in the All of Us Research Program. *Nature* 2024;627:340–6.
68. Rodríguez OL, Gibson WS, Parks T *et al*. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front Immunol* 2020;11:2136. <https://doi.org/10.3389/fimmu.2020.02136>
69. Grochowski CM, Bengtsson JD, Du H *et al*. Inverted triplications formed by iterative template switches generate structural variant diversity at genomic disorder loci. *Cell Genomics* 2024;4:100590. <https://doi.org/10.1016/j.xgen.2024.100590>
70. Zhang X, Liu Y, Huang M *et al*. Landscape of double-stranded DNA breaks in postmortem brains from Alzheimer's disease and non-demented individuals. *JAD* 2023;94:519–35. <https://doi.org/10.3233/JAD-230316>
71. Ramesh V, Liu F, Minto MS *et al*. Bidirectional regulation of postmitotic H3K27me3 distributions underlie cerebellar granule neuron maturation dynamics. *eLife* 2023;12:e86273. <https://doi.org/10.7554/eLife.86273>
72. Liu J, Wu X, Zhang H *et al*. Dynamics of RNA polymerase II pausing and bivalent histone H3 methylation during neuronal differentiation in brain development. *Cell Rep* 2017;20:1307–18. <https://doi.org/10.1016/j.celrep.2017.07.046>
73. García Fernández F, Almayrac E, Carré Simon À *et al*. Global chromatin mobility induced by a DSB is dictated by chromosomal conformation and defines the HR outcome. *eLife* 2022;11:e78015.
74. Lee C-S, Wang RW, Chang H-H *et al*. Chromosome position determines the success of double-strand break repair. *Proc Natl Acad Sci USA* 2016;113:E146–54.
75. Roychowdhury T, Abyzov A. Chromatin organization modulates the origin of heritable structural variations in human genome. *Nucleic Acids Res* 2019;47:2766–77. <https://doi.org/10.1093/nar/gkz103>
76. Shokrollahi M, Stanic M, Hundal A *et al*. DNA double-strand break-capturing nuclear envelope tubules drive DNA repair. *Nat Struct Mol Biol* 2024;31:1319–30. <https://doi.org/10.1038/s41594-024-01286-7>
77. Chung DKC, Chan JNY, Strecker J *et al*. Perinuclear tethers license telomeric DSBs for a broad kinesin- and NPC-dependent DNA repair process. *Nat Commun* 2015;6:7742. <https://doi.org/10.1038/ncomms8742>
78. Zhang F, Khajavi M, Connolly AM *et al*. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* 2009;41:849–53. <https://doi.org/10.1038/ng.399>
79. Pentao L, Wise CA, Chinault AC *et al*. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat Genet* 1992;2:292–300. <https://doi.org/10.1038/ng1292-292>
80. Van Esch H, Hollanders K, Badisco L *et al*. Deletion of VCX-A due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis. *Hum Mol Genet* 2005;14:1795–803. <https://doi.org/10.1093/hmg/ddi186>
81. Urbán Z, Helms C, Fekete G *et al*. 7q11.23 deletions in Williams syndrome arise as a consequence of unequal meiotic crossover. *Am J Hum Genet* 1996;59:958–62.
82. Glanz A, Fraser FC. Spectrum of anomalies in Fanconi anaemia. *J Med Genet* 1982;19:412–6. <https://doi.org/10.1136/jmg.19.6.412>
83. Kirsche M, Prabhu G, Sherman R *et al*. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* 2023;20:408–17. <https://doi.org/10.1038/s41592-022-01753-3>
84. Poplin R, Ruano-Rubio V, DePristo MA *et al*. Scaling accurate genetic variant discovery to tens of thousands of samples. *Biorxiv*, <https://doi.org/10.1101/201178>. 24 July 2018, preprint: not peer reviewed
85. Larson DE, Abel HJ, Chiang C *et al*. Svttools: population-scale analysis of structural variation. *Bioinformatics* 2019;35:4782–7. <https://doi.org/10.1093/bioinformatics/btz492>
86. Chen X, Schulz-Trieglaff O, Shaw R *et al*. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;32:1220–2. <https://doi.org/10.1093/bioinformatics/btv710>
87. Alexandrov LB, Nik-Zainal S, Wedge DC *et al*. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 2013;3:246–59. <https://doi.org/10.1016/j.celrep.2012.12.008>
88. Sloan CA, Chan ET, Davidson JM *et al*. ENCODE data at the ENCODE portal. *Nucleic Acids Res* 2016;44:D726–32. <https://doi.org/10.1093/nar/gkv1160>
89. Zhao PA, Sasaki T, Gilbert DM. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol* 2020;21:76. <https://doi.org/10.1186/s13059-020-01983-8>
90. Reiff SB, Schroeder AJ, Kirli K *et al*. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat Commun* 2022;13:2365. <https://doi.org/10.1038/s41467-022-29697-4>
91. Brison O, El-Hilali S, Azar D *et al*. Transcription-mediated organization of the replication initiation program across large genes sets common fragile sites genome-wide. *Nat Commun* 2019;10:5693. <https://doi.org/10.1038/s41467-019-13674-5>
92. Cer RZ, Donohue DE, Mudunuri US *et al*. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* 2013;41:D94–D100. <https://doi.org/10.1093/nar/gks955>
93. Chiu T-P, Comoglio F, Zhou T *et al*. DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* 2016;32:1211–3. <https://doi.org/10.1093/bioinformatics/btv735>
94. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2. <https://doi.org/10.1093/bioinformatics/btq033>
95. Project Consortium ENCODE, Moore JE, Purcaro MJ *et al*. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020;583:699–710. <https://doi.org/10.1038/s41586-020-2493-4>
96. Dixon JR, Jung I, Selvaraj S *et al*. Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;518:331–6. <https://doi.org/10.1038/nature14222>

97. Shah PP, Keough KC, Gjoni K *et al.* An atlas of lamina-associated chromatin across twelve human cell types reveals an intermediate chromatin subtype. *Genome Biol* 2023;24:16.
<https://doi.org/10.1186/s13059-023-02849-5>
98. Kumar R, Nagpal G, Kumar V *et al.* S. HumCFS: a database of fragile sites in human chromosomes. *Bmc Genomics [Electronic Resource]* 2019;19:985.
<https://doi.org/10.1186/s12864-018-5330-5>
99. Milacic M, Beavers D, Conley P *et al.* The reactome pathway knowledgebase 2024. *Nucleic Acids Res* 2024;52:D672–8.
<https://doi.org/10.1093/nar/gkad1025>