High-throughput linear amplification for single-cell template strand sequencing with sci-L3-Strand-seq

Peter Chovanec[1], Trevor Ridgley[1], Yi Yin[1,*]

1 Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA
* Lead contact; email: yiyin@mednet.ucla.edu

Summary

Despite the advances in single cell whole genome sequencing, the ability to detect structural variations from mitotic crossovers remains challenging, while sister chromatid exchanges remain undetectable. Here we describe sci-L3-Strand-seq, a combinatorial indexing method with linear amplification for DNA template strand sequencing that cost-effectively scales to millions of single cells. We provide a computational framework to fully leverage the throughput, as well as the relatively sparse but multifaceted genotype information within each cell that includes strandedness, digital counting of copy numbers, and haplotype-aware chromosome segmentation, to systematically distinguish seven possible types of mitotic crossover outcomes and resulting genome instability. We showcase the power of sci-L3-Strand-seq by quantifying the rates of error-free and mutational crossovers in thousands of cells, enabling us to explore enrichment patterns of genomic and epigenomic features. The throughput of sci-L3-Strand-seq also gave us the ability to measure subtle phenotypes, opening the door for future large mutational screens. Furthermore, mapping clonal lineages provided insights into the temporal order of certain genome instability events, showcasing the potential to dissect cancer evolution. Altogether, we show the wide applicability of sci-L3-Strand-seq to the study of DNA repair, structural variations, and genome instability in disease.

Introduction

DNA double stranded breaks (DSB) pose a major risk to genome integrity. During the S and G2 phase of the cell cycle, the use of homologous DNA from sister chromatids or homologs facilitates DSB repair through homologous recombination (HR). After resection of the DSB ends, HR can initiate repair through single-ended invasion via synthesis-dependent strand annealing (SDSA) or break-induced replication (BIR), or in combination with second-end capture produce a Holliday junction (HJ) intermediate that can undergo dissolution or resolution. Dissolution of a double HJ leads to a non-crossover (NCO), while resolution can produce both NCO and crossover (CO) outcomes [1,2]. Despite the error-free nature of HR, COs can lead to large-scale genome rearrangements that pose a risk to the stability of the genome.

We have previously categorized genomic rearrangements produced from mitotic COs into seven classes that can be broadly divided into events arising from allelic HR and non-allelic HR (NAHR). Meanwhile,

unresolved HR intermediates could also affect proper chromosome segregation [3]. Briefly, allelic HR can generate error-free sister chromatid exchanges (SCE, class 1) (Fig1A and FigS1B), or large-scale copy-neutral loss-of-heterozygosity (cnLOH, class2) (Fig1B) if parental homolog was used as repair template. With the proximity of the newly replicated sister, the bias of using the sister chromatid is over 100-fold compared to using the parental homolog harboring single-nucleotide variations (SNVs) [4]. The high prevalence of repeats in mammalian genomes provides an alternative source of homology that can result in rearrangements between non-allelic sequences. NAHR between intra-chromatid repeats can give rise to copy number variations (CNV, class 3) (Fig1C and FigS1C) and inversions (class 4) (Fig1D); while those between inter-chromosomal repeats can result in translocations (class 5) (Fig1E). Lastly, unresolved HR intermediates could give rise to aneuploidy (class 6) (Fig1F and FigS1D) and uniparental disomy (UPD, class 7) (Fig1G). The exact relationship between crossover and whole-chromosome alterations has not been fully-established. Chua and Jinks-Robertson [5] found increased chromosome loss following a mitotic crossover event in yeast, possibly due to increased nondisjunction. UPD may be formed either through an aneuploid intermediate resulting from nondisjunction, or through meiosis-like homolog separation [6,7] in a single division. Altogether, the ability to detect a wide variety of repair outcomes genome-wide without bias is key to understanding the full complexity of mitotic crossover.

To enable the study of these genomic rearrangements, we recently developed a set of cost-effective methodologies that combine single-cell combinatorial indexing (sci) with reduced amplification bias using linear amplification (L), and 3-level barcoding for increased throughput called sci-L3 [8].The principal method, sci-L3, enables single-cell whole genome sequencing (WGS) for thousands of cells and addresses several limitations of methods widely used for the study of mitotic CO outcomes. The few studies examining mitotic CO outcomes genome-wide rely on bulk whole genome sequencing (WGS) of subcloned individual cells [9–12]. The need to manually clone daughter cells is labor intensive, intractable for scaling to large sets of mutants, and unfeasible for *in vivo* studies. Sci-L3-WGS can directly map reciprocal cnLOH and UPD events in single cells, including *in vivo* that otherwise requires clonal expansion [13], and has the potential to scale to large sets of mutants. Despite the advances gained with sci-L3-WGS, the 99% of mitotic COs that constitute error-free SCEs still cannot be detected at all. Moreover, the detection of inversions and balanced translocation is still reliant on junction-spanning reads that require a high sequencing depth (FigS1A).

The semi-conservative nature of DNA replication facilitates the differential incorporation of the thymidine analog bromodeoxyuridine (BrdU) between sister chromatids. After one cell division with BrdU, the labeling in daughter cells enables the detection of strand exchanges between recombined sister chromatids or homologs [14,15]. These concepts have been applied in a sequencing-based approach called Strand-seq, which allows the genome-wide detection of SCEs at high resolution (Fig1A) [16]. Strand-seq has seen many additional applications such as the improved detection and genotyping of inversions and some of the other structural variations (SVs) (Fig1C-F) [17–19], and the construction of chromosome level haplotype-phased genome assemblies [20,21]. However, Strand-seq is constrained by: 1) a very low throughput that leads to 2) a high cost per cell, 3) and the reliance on bias-prone PCR amplification[16,22]. The development of OP-Strand-seq improved throughput but requires specialized equipment not commonly available[23]. The chemistry in the original and the OP-Strand-seq uses MNase, which prevents chromosome fragmentation pattern-based “digital” information for counting chromosome copy numbers, while transposon-based chromosome fragmentation uniquely enables digital counting [24]. Such digital copy number counts can be further used in a strand-specific and haplotype-aware fashion, which substantially improves analyses of various types of SVs with low coverage.

Design

With the sci-L3-Strand-seq methodology we provide both experimental and computational advances that together establish a unifying framework for mapping all the seven types of mitotic crossover outcomes at scale. Experimentally, we extended the sci-L3 barcoding and linear amplification scheme to Strand-seq, which requires: 1) successful ablation of the nascent strand by employing a novel enzymatic step [25]; and 2) an amplification chemistry that fully preserves strand information. Computationally, we tackled the trade-off between throughput and coverage per cell. Our computational framework 1) at a single cell level fully integrates relatively sparse but multifaceted genotype information by using strandedness, haplotype, digital counting of copy numbers, and their combinations for both chromosome segmentation and systematically annotate different types of mitotic crossover outcomes; and 2) at a population level fully leverages the throughput to dissect subclonal complex rearrangement involving inversions, translocations in duplicated regions. Conceptually, we show the possibility of interpreting the underlying mechanisms for some of the genome instability events given the observed subclonal structure.

Results

Expansion of sci-L3 to Strand-seq

Recombination between identical sister chromatids does not change the DNA sequence and is thus missed in short- and long-read WGS. Strand-seq only sequences the template strands of replication in the next G1 post cell division, enabling mapping of error-free SCE [16,22]. To integrate Strand-seq with the sci-L3 scheme (Fig1H) requires ablation of BrdU labeled nascent strands while maintaining strand directionality in the whole-genome amplification process. We summarize key deviations from the original Strand-seq [22]: 1) nuclei are formaldehyde fixed to enable the split-and-pool barcoding strategy; 2) Tn5 transposome is used instead of MNase for genome fragmentation and first-round barcoding. The single-stranded strand transfer activity of Tn5 transposase preserves template strand directionality, and uniquely enables digital counting of DNA copy number [24] with low per cell coverage; 3) sci-L3 employs sticky-end ligation of the second-round barcode with the T7 promoter in a hairpin loop to improve ligation efficiency and to minimize amplification of non-genomic DNA if the T7 promoter sequence were supplied in duplex. These features further ensures subsequent unidirectional linear amplification with *in-vitro* transcription (IVT) (FigS2); 4) notably, we found that an additional enzymatic step with Uracil DNA Glycosylase (UDG) and endonuclease VIII (EndoVIII) is required to ablate the incorporated BrdU [25].

These novel additions to both the sci-L3 and Strand-seq not only produced Strand-seq libraries at scale, but pose opportunities for computational method development. To tackle the throughput vs. coverage trade-off, we typically build iterative processes that go through cycles of: 1) leveraging population information for a large number of cells (e.g., phasing: assigning heterozygous SNVs to their parental homologs by strandedness of the reads); 2) integrating multifaceted genotype information (strandedness, newly phased haplotype, digital counting of copy numbers, and *their combinations*) at a single cell level, 3) identifying subclonal complex SVs, again at a population level; and 4) revisiting event type annotation at a single cell level given the subclonal information. For example, the limited throughput of Strand-seq typically renders phasing a stand-alone effort [21,26] from calling SVs. Here, we provide examples of an iterative process involving phasing, calling SV, and refining phasing. Additionally, mutational SVs (like deletions and inversions) could also involve strand switches, and are sometimes indistinguishable from SCE. With our increased number of cells analyzed, we identified “twin hotspots” in SCE profiles actually reflecting underlying subclonal inversions. We present our computational toolkit in the order of the seven types of mitotic crossover outcomes below, but highlight where these analyses crosstalk.

We piloted our sci-L3-Strand-seq analysis framework in three cell lines. HAP1 is a human haploid cell line that is extensively used for genome engineering and is an ideal model for future mutational screens [27,28]. We have a clear-cut expectation of observing only Watson (W; positive strand) or Crick (C; negative strand) reads within any given region if originating only from the template strand, and we can accurately assess background levels and hence library quality as a result (see Supplemental Text and Methods). Patski is a mouse female interspecific F1 (C57BL/6J x *M. spretus*) cell line with high quality SNVs published [29,30] and ground truth for benchmarking phasing with large cell number but sparse coverage, as well as for exploring adjustment to phasing methods with pre-existing subclonal LOH events. Lastly, we examined BJ-5ta, a stable human male line derived from neonatal tissue without any karyotypic abnormalities [8,24], but unphased. Commercial and patient-derived human cell lines are typically unphased, yet phasing is required for identifying cnLOH. Collectively, these cell lines are ideal for assessing sci-L3-Strand-seq library quality, benchmarking the ability to phase haplotypes and identify SVs, and using sci-L3-Strand-seq as a unifying framework for crossover outcome analyses.

Sci-L3-Strand-seq captures the replication template strands and identifies SCEs

In HAP1 cells, aside from the known disomic region on chr15[31], sci-L3-Strand-seq generates clear strand specific profiles and strand switches with a median background estimate of 4.9% (meaning 4.9 C reads for every 100 W reads in a W chromosome, FigS3A-B). Importantly, our ability to sequence DNA template strands showed all the enzymatic steps in sci-L3 preserve strand information (Fig2A). Diploidization, however, is commonly observed with HAP1 cells. We devised a cell ploidy filter based on strand-neutral percentage and background estimate (FigS4). Subsequent analyses on HAP1 SCEs are for filtered haploid cells only. For the two diploid cell lines, we also found clear strand states and strand switches, and observed a median background of 3.3% and 4.4% for Patski and BJ-5ta cells respectively (FigS3A-C). Selecting cells with consistent profiles, low background and the expected distributions of strand states for each ploidy eliminated cells with low quality strand information and gave us a total of 1968 HAP1, 2176 BJ-5ta, and 931 Patski single cells (see Supplemental Text, Methods, FigS4A-C, FigS5A-C, TabS1).

Several methods have been developed for identifying strand switches (breakpoints) produced by SCEs and other SVs in Strand-seq datasets[32–34]. However, the sequencing effort between Strand-seq and sci-L3-Strand-seq is substantially different. In Strand-seq, sequencing 700-1,400k read pairs provides 3-5% coverage, while we typically sequence 30-200k read pairs given the substantially increased throughput with sci-L3 of typically 10-100k to potentially 1M single cells. For haploid cells the projected coverage is 1% at 100k read pairs per cell. For diploid cells, sequencing over 300k can achieve 2% coverage (FigS6A). For cost-effectiveness, we therefore performed sub-saturation sequencing that produced a 0.27-0.55% median coverage per cell (FigS6B). Despite this low depth, we were able to reliably identify breakpoints by developing a framework of filters targeting false positive calls[34] (FigS6C, see Methods). In brief, we filter strand switches as a result of misassembly, centromere artifacts, pre-existing SVs including inversions, translocations, and duplications (FigS7, “region filter”), and/or subclonal mutational CO outcomes including CNVs and inversions (TabS2, TabS4). Details for mapping such events and separating SCEs from these other types of breakpoints are discussed in respective sections. On top of manual validation, we examined cells across a range of less than 0.3% to over 1.8% sequencing coverage and found a steady median number of breakpoints throughout with no significant differences (FigS6D-E). We found a high correlation between SCE frequency and chromosome size (R^2 of 0.66-0.94, FigS8A-B) that matched previous observations and further validated our calls [16,22,35]. Overall, we were able to consistently identify true breakpoints and obtain high confidence sets of 21,761 SCEs with median resolutions ranging from 128 to 225 Kb for both haploid and diploid cells (FigS6F).

The ability of sci-L3-Strand-seq to examine thousands of single cells not only enables the mapping of rare spontaneous mitotic COs, but also provides statistical power for the detection of subtle mutant phenotypes. As a proof-of-concept, we used the HAP1 cell line with a CRISPR-Cas9 engineered knockout (KO) of DNA ligase 4 (*LIG4*) (referred to as $HAP1^{lig4-}$) and applied sci-L3-Strand-seq to measure the rate of SCE [36]. With alternative end-joining and non-crossover HR still active, the expectation was to observe only a moderate increase in the number of crossovers. We analyzed 751 $HAP1^{lig4-}$ single cells (TabS1) and found a significant increase in the number of SCEs per cell in the *LIG4* KO (p-value: 0.000376; Mann-Whitney U) with a median increase from 3 to 4 SCEs per cell (Fig2B). Interestingly, HAP1 cells have almost double the SCEs per haploid genome of diploid BJ-5ta and Patski cells, suggesting cell-type difference in the level of SCE. Overall, the ability to observe subtle alteration in the rate of SCEs demonstrates the scalability of sci-L3-Strand-seq and opens the doors for large scale mutational screens.

Human centromeric regions have reduced levels of SCEs

Sci-L3-Strand-seq offers an advantage in detecting CO events by 1) providing strand information for every read, unlike SNVs, and 2) propagating this information across genomic distances. As a result, CO events can be detected even in poorly mappable regions, such as centromeres, by observing opposing strand directionality in flanking unique sequences. Crossover recombination during meiosis is suppressed within centromeres to prevent segregation errors[37]. However, during mitotic proliferation centromeres are prone to ss- and ds-DNA breaks[38]. These breaks could presumably be repaired by crossover, or other repeat-mediated annealing not involving crossovers. Meanwhile, studies of mosaic chromosomal alterations in healthy individuals and in various tumor sequencing revealed large-scale cnLOH as one of the most abundant type of SVs, frequently involving the entire chromosome arm, suggesting increased level of centromere-initiated inter-homolog crossovers [39,40]. We therefore examined whether mitotic centromeres on human non-acrocentric chromosomes have elevated or reduced levels of SCEs. After normalizing by centromere sizes from the T2T-CHM13 genome assembly (FigS9A) [41], we found that in HAP1 cells 5 out of 18 centromeres have significantly reduced rates of SCEs compared to the rest of the chromosome (p-value < 0.001; permutation test) (Fig2C). BJ-5ta cells also displayed significantly lower SCE levels for 8 centromeres (p-value < 0.01; permutation test). Interestingly, HAP1^{lig4-} cells had only 2 centromeres showing reduced level of SCEs (p-value <0.01; permutation test) with a general increase in centromeric SCE compared to HAP1. This includes a significant elevation of SCE level on chr17 and a shared elevation with HAP1 cells on chr19 (p-value < 0.01; permutation test). The exact position of the insertional translocation of chr15 into chr19 [42] was not mapped, but it could have disrupted the centromere, triggering instability and an enrichment of SCEs. Altogether, our data suggest centromeres generally suppress mitotic HR crossovers.

Late replicating domains and several repeat classes show increased rate of SCEs

Focusing on the genome-wide distribution of SCEs we found HAP1, BJ-5ta, and Patski cells had a generally uniform distribution (Fig2D). Nonetheless, we found 4 out 6 hot regions (in red) to be significantly enriched for SCEs in HAP1, 1 in BJ-5ta, and 2 in Patski cells (Fig2D, TabS3 and TabS4). Although not all statistically significant, HAP1^{lig4-} cells have 11 hot regions of enrichment (Fig2B and D), suggesting HR has preferential sites of repair [43].

HR being the favored repair pathway within heterochromatin led us to examine the correlation of SCEs with replication timing domains and AB compartments[44,45]. Late replicating domains and B compartments are highly correlated with each other and composed primarily of heterochromatin with low levels of DNA methylation[46–49]. As expected, we found SCEs were more associated with late replicating domains (Repli-seq) and B compartments (Hi-C) in human cells.

We also wondered if we could determine the type of break that initiated our SCEs based on correlation with END-seq and GLOE-seq datasets[50,51]. Interestingly, END-seq on etoposide (ETO) treated cells showed no correlation with SCEs in either mouse or human (FigS10 and TabS5). Instead, the no treatment controls were significantly correlated with SCEs, suggesting the observed SCEs are the result of spontaneous DSBs rather than Topoisomerase II (TOP2) induced DSBs ($p < 6x10^{-08}$, Spearman correlation test). Moreover, we found no association of SCEs with single stranded nicks (SSN) from GLOE-seq, suggesting spontaneous SSNs are unlikely to be the lesions generating our SCEs (FigS10A and TabS5).

To test whether any other genomic or epigenomic features could be tied to SCE enrichment, we compiled a set of additional 22 human and 39 mouse features that included GC content, sequence divergence, repeat annotations, chromosome domains, and transcription factor binding sites (FigS10). In agreement with our permutation analysis, we found centromeric regions and satellite repeats were negatively correlated with SCEs across both human and mouse cell lines (Fig2C, FigS10A-B and TabS5). In contrast, we found several other repeat classes to be positively correlated with SCEs such as long interspersed nuclear elements (LINEs) and low complexity (LC) regions (FigS10A and TabS5). Interestingly, the strongest positively correlated feature in mouse Patski cells were regions of sequence divergence between the C57BL/6J and *M.Spretus* genomes ($p < 3.4x10^{-07}$; Spearman correlation test; FigS10B). The divergent sequences may prevent inter-homolog (IH) recombination[4], possibly due to the anti-recombination function of mismatch repair [52]. Notably, chromosome fragile sites (CFS) were also positively correlated with SCEs in HAP1 and BJ-5ta cells[53]. Overall, regions with higher rate of initiating lesions, structure forming repeats, and sequence divergence likely further influence the distribution of SCEs[54,55].

Sci-L3-Strand-seq enables accurate phasing of heterozygous variants and detection of LOH

In low-coverage single-cell sequencing, identifying copy neutral events, such as cnLOH and UPD, requires the separation or 'phasing' of heterozygous variants by their maternal or paternal origin. This is because for heterozygous sites, the occurrence of reads that represent both alleles are rare. Phasing is thus required to determine whether long contiguous stretches of variants are homozygous for one haplotype, or alternating between two haplotypes by randomly sampling at heterozygous sites. In regions where the orientation of the inherited DNA template strands differ, Strand-seq captures one parental homolog with Watson and the other with Crick strand-oriented reads ("WC coverage"). This one-to-one correspondence of strand state with haplotype can be used to obtain chromosome level phasing [20,26].

As proof-of-concept that sci-L3-Strand-seq can accurately phase variants, we used Patski cells with known haplotypes (C57BL/6J x *M.spretus*) [30,56] to assess phasing accuracy, and BJ-5ta cells, which have no phasing information available. Reads from WC state regions covered 85.7% (24.5M heterozygous SNVs out of 28.7M) and 92.3% (2.4M heterozygous SNVs out of 2.6M excluding chromosomes X and Y) of the total identified heterozygous sites within Patski and BJ-5ta cells respectively. Out of these, 91.6% (22.5M with WC coverage) were successfully phased for Patski and 90.7% (2.1M SNVs with WC coverage) for BJ-5ta cells (TabS6 and TabS7). Altogether, after applying corrections and filtering (summarized below, and in Methods), we obtained a final average phasing accuracy of 98.6% for Patski cells (Fig3A and FigS12). With human cell lines that are typically unphased, only successful phasing enables LOH calling; we therefore validated phasing by successful calling of LOH events.

Phasing of the WC covered sites with StrandPhaseR initially produced disjoint sets of either one or both SNVs at each site being phased; while the second haplotype can be easily inferred (FigS12A-C) [57]. Unexpectedly, we also found an average 1.3% and 6% of heterozygous sites with the same variant assigned to both haplotypes in Patski and BJ-5ta cells respectively, which we labeled as conflicting sites and excluded from downstream analysis. For Patski, this improved phasing accuracy from 97.4% to 98.1%. With over 900 cells used for phasing, we could leverage the throughput to investigate how subclonal SVs affect phasing accuracy, and in some cases, phase with a different strategy. For example, when pre-existing, subclonal LOH events are present in over 80% of the cells (chr4), we found that manual phasing using the LOH provides proficient and accurate answers. Chromosome 4 has a mixture of both monosomy and UPD (Fig4A) and the UPD events could have WC coverage, potentially interfering

with phasing. Another subclonal LOH on chr12 only affects about 30% of the cells and are exclusively deletions. Deletions do not contribute WC coverage and thus phasing without considering these SVs does not affect phasing performance. Lastly, we addressed inaccurate phasing due to subclonal inversions and centromere-proximal regions, both improving our overall phasing accuracy (see Methods).

Without a ground truth, a reasonable validation of phasing in BJ-5ta cells is the ability to detect long contiguous regions of an extreme haplotype fraction that would otherwise be masked due to random haplotype assignment. We defined extreme as the proportion above 0.9 (primarily haplotype 1) or below 0.1 (primarily haplotype 2) of the total phased haplotypes for regions derived from breakpoint calls and chromosomal arms and considered these as potential regions of LOH (FigS14A and B). For Patski cells we identified a total of 2306 regions with extreme haplotype fractions (TabS8). The segmental deletion type of LOH has a 50% chance of containing identifiable strand switches (FigS18A). Nonetheless, the heritable and irreversible nature of LOH that leads to its clonal expansion enabled us to rescue the other 50% by querying all the cells for the same region to find LOH (Fig3B, see Methods). This led to the rescue of over 400 additional segmental LOH (0.508 versus the expected 0.5; TabS10). Taking the same approach with BJ-5ta cells, we identified a total of 228 regions with extreme haplotype fractions (TabS9), despite BJ-5ta cells having 1/12 as many phased SNVs compared to Patski. Overall, we identified multiple potential regions of LOH that reaffirmed our ability to accurately phase variants.

Digital counting of overlapping reads better informs copy number

As with LOH, the low coverage and amplification noise of single-cell WGS methods makes it challenging to identify CNVs. Promisingly, a digital fragment counting approach implemented for LIANTI was shown to additionally reduce noise over using read depths alone [24]. In essence, collapsing reads into insertion patterns (fragments) based on read ends and Tn5 overhangs gives an approximation that is closer to the absolute copy number. We took this concept further by first devising a rigorous deduplication scheme that accounts for possible IVT artifacts and additionally makes use of the sci-L3-Strand-seq strand information (FigS11, see Methods). Subsequently, we counted the number of overlaps between insertional patterns that are indicative of copy number gain or loss (Fig3C). These digital copy number counts are powerful when combined with additional layers of information including haplotypes and strand directionality. For example, a WC chromosome or region could have W-specific overlapping reads suggesting at least two W copies, but only one C copy, which immediately suggests a copy number gain event (2 W and 1 C). Similarly, haplotype-specific overlapping reads also suggest copy number gain or UPD. We thus build machine learning tools to integrate such information for SV calling.

Strand switch- and clonal structure-based segmentation with machine learning classifiers facilitate accurate annotation of SVs

Chromosome segmentation for SV analysis in sparse single-cell sequencing is challenging. Conceptually, one can use a binned chromosome window [33], a binned number of SNV sites [8], or Hidden Markov Models [13] for segmentation. Mechanistically, mutational SVs that change copy number and/or strand directionality can couple with strand switches [18]. Therefore, not all strand switches reflect an SCE event. For example, a deletion on a WC chromosome (50% in theory) retains only one of the two strands. We thus use strand switches to segment the genome for calling other types of mutational SVs. As mentioned above, subclonal SVs that do not couple with strand switches (e.g., a deletion on a WW chromosome) can be rescued. We use strand switches-defined segments for initial LOH scan, and perform clonal rescues from

the LOH scan, and then use both breakpoint-defined and rescued segments as candidate regions for SV annotation.

Using digital counts together with phased haplotype fractions, strand information, and various other metrics capturing read depth and background levels (FigS15A-C) we trained a set of support vector machine (SVM) classifiers for the annotation of cnLOH (segmental and UPD), CNV (gain or loss, segmental and aneuploidy), and “drop-to-zero” regions (DTZ; see Methods) (FigS16 and FigS17). The training consisted of manually annotated datasets with 1) 106 HAP1 and 62 HAP1^{lig4-} cells for haploid gain events, 2) 24 HAP1 and 13 HAP1^{lig4-} cells for haploid DTZ events, and 3) 75 BJ-5ta and 64 Patski cells for all four mutational event classes in diploid cells (see Supplemental Text and Methods). Using a 70:30 training to test data split we found all 12 SVM classifiers achieved high accuracies ranging from 0.971 to 1 (FigS16 and FigS17, TabS12). Examining the importance of each feature revealed the biggest drivers of accurate loss and cnLOH classification were the phased haplotype fraction and strand state features, while for gain it was the digital counts of overlapping reads (FigS16B, FigS17B and D). We found that stratifying the overlap counts with strand information facilitated the identification of unbalanced CNV gains and importantly helped us distinguish CNV loss from cnLOH. Incorporating haplotype information into the overlap counting also helped resolve additional CNV gains, losses, and complex events that we discuss below (haplotype-aware segmentation). Overall, our digital counting of overlaps substantially improved our ability to identify copy number changes with sub-saturation sequencing and together with other metrics enabled us to train accurate machine learning models for the annotation of various SVs.

Haplotype-aware segmentation addresses lack of strand switch in cnLOH, enabling their identification, and improves other SV calls

Our rescue scheme, leveraging the high-throughput nature of sci-L3-Strand-seq, can map clonally expanded deletions lacking strand-based segmentation (expected for 50% of deletions), but also highlight the limitations of using strand switches for SV annotation (FigS18A and TabS10). We thus reviewed what type of SVs can be systematically missed with strand-switch-based and rescued segmentation. Most notably, cnLOH are generational and typically do not contain strand switches. They are copy-neutral in nature (unlike deletions) and thus lack loss of one strand that would enable initial discovery for a candidate region to scan for rescue (FigS18B). Manual examination of cnLOH events that were rescued revealed them to be complex events such as cnLOH embedded within a whole chromosome gain and therefore a deletion in nature. Motivated by the lack of strand switches with cnLOH, we devised a haplotype-aware approach for segmentation (FigS18A-C; see supplemental text).

Copy-neutral chromosomes without an LOH strand switch can contain 1) segmental cnLOH, 2) a linkage switch without a cnLOH produced by inter-homolog crossovers where the two recombined chromatids co-segregated, or 3) just normal heterozygous chromosomes without an inter-homolog recombination (FigS18C). In the first two scenarios, reads that contain phased variants can be split into Haplotype1 and Haplotype2 to reveal strand switches within the individual haplotypes. Firstly, given a WC region with a segmental cnLOH, the separation of the reads by haplotype forms C to WC and W to DTZ regions that creates identifiable point of segmentation (FigS18C top panel). For WW or CC regions with a cnLOH, the separation produces regions with a W or C to DTZ and a W to WW or a C to CC transition that are both hard to detect due to our low sequencing coverage (FigS18C middle panel). Nevertheless, if the segmentation was successful for clonal cnLOH in WC chromosomes, the cnLOH within same-strand chromosomes can be rescued. Secondly, inter-homolog crossovers that produce a linkage switch (uncoupled with BrdU labeling) without a cnLOH also produce identifiable W to C and C to W transitions with the separation of reads by haplotype, enabling segmentation (FigS18C bottom panel). While our analysis is focused on the detection of cnLOH, linkage switches without cnLOH are diagnostically

important as they can have phenotypic consequences in cases where haplotype-specific enhancers are brought to and hijacked by oncogenes [58]. It is also of interest to understand whether recombined chromatids in an inter-homolog crossover event have any bias in subsequent chromosome segregation [5]. Notably, these events elude most bulk and single-cell WGS analysis. We manually examined newly annotated events and found the haplotype-aware segmentation enabled the identification of new cnLOH events, which contain the expected segmental DTZ in one haplotype and matching W and C state reads in the other haplotype for an otherwise fully WC strand-state chromosome (FigS19A and TabS11).

Using the haplotype-aware approach on Patski and BJ-5ta cells enabled not only the identification of cnLOH, but also de-novo segmental loss, and copy number gain events that originally lacked segmentation with strand-switch calling (FigS19). With low sequencing coverage the drop in read count for loss events can be difficult to identify (2 copies vs 1 copy). However, splitting reads in these regions by haplotype with the contrast in read counts between each independent haplotype makes differentiating the drop much easier (1 copy vs 0 copies). This is because regions termed 'drop to zero' (DTZ – 0 copies) containing only background reads can be thresholded on the background levels established from W or C-only regions (see Methods), essentially converting the problem into DTZ identification. This enabled us to annotate additional clonal loss events that were just outside of our haplotype fraction thresholds for rescue (FigS19B right panel).

For most of the new gain annotations that we manually validated, we found them to be trisomies with missed SCE calls (5 out of 7 manually examined events in Patski cells) (FigS19C first panel). In these scenarios, a missed strand-switch results in the two regions with opposing information to average out and create the appearance of a balanced copy number (e.g. 2-1 and 1-2 into 3-3). The haplotype-aware approach finds WC regions for a single haplotype, meaning at least two copies of one haplotype are present, signifying a copy number gain (FigS19C second panel). For non-WC regions, haplotype-specific overlapping reads indicate gains (FigS19C third panel). Notably, WC strand states present within both haplotypes can identify balanced gains (4-copy) and ultimately help disentangle complex events (alternate haplotype loss in a tetrasomy, FigS19C fourth panel) and high copy-number events (FigS19D). Beyond event annotation, the haplotype-aware segmentation approach also highlights inversions in the mouse reference genome that appear as reciprocal "SCE" breakpoints between the two haplotypes for WC chromosomes (FigS19E). These could also be actual homozygous inversions. Altogether, we show the haplotype-aware segmentation framework as described here facilitates the identification and interpretation of various mitotic crossover outcomes from low coverage single cells.

Annotation of SVs reveals clonal population structures

After further stratifying our initial mutational SV annotations (SCE in the last cycle of BrdU labeling are not clonal) into whole chromosome or segmental events, we performed hierarchical clustering with the optimal number of clusters to get an overview of cells with at least one event (Fig4A and FigS20A-C, see Methods). In Patski cells we found 16 clonal clusters out of 177 total clusters with more than 10 cells that accounted for 58% (519/894) of all annotated cells (Fig4A). Since almost all the cells have a pre-existing chr4 LOH, 96.03% of Patski cells contain at least one annotated SV event compared to only 13.51% of BJ-5ta cells (FigS20E). Interestingly, we found a single annotated SV difference between most clonal populations in Patski, suggesting a stepwise addition of events where cells with only a single SV would represent the founder population. This led us to identify the following early founding events: monosomy on chr4, UPD on chr4, trisomy on chr15, trisomy on chr19, and a segmental loss on chr2 (Fig4A – highlighted on the side); which we later use to reconstruct a hierarchy of events that reveals a bias in SV occurrence. With the wealth of events in Patski cells we repeated the features correlation analysis and found the underlying (epi)genomic context did not allow us to distinguish between error-free (SCE) and

mutational repair (the borders of segmental SVs; termed “adjacent breakpoints”) (FigS10B). In BJ-5ta cells, despite the sparsity of SVs, we observed 3 clonal populations with more than 10 cells (3/58) accounting for 28% (83/294) of all annotated cells (FigS21A). The most prominent clone contained an interstitial loss on chr4 (44 in 294 cells). Manual examination of coincident breakpoint hotspots (double breakpoints) without an initial SV annotation revealed a further 9% (4/44) of BJ-5ta cells with the same SV (Fig3B and FigS21). Our ability to identify most of these clonal events gave us confidence in our SV annotations.

Despite the smaller clonal populations in haploid cells, we found 14 out of 60 HAP1 clonal clusters and 7 out of 48 HAP1^{lig4-} with more than 5 cells that accounted for 41% (105/255) and 31% (48/157) of all annotated cells respectively (FigS20B and C). We found gains to be almost double that of DTZ events in HAP1 (78 gain to 46 DTZ) and almost three times in HAP1^{lig4-} cells (41 gain to 15 DTZ). The sci-L3 method does not require live cells in general, and could profile cells in the G1 cycle post BrdU labeling with fresh but lethal loss (like DTZ in haploid cells). The “clonal” DTZ present in a few haploid cells likely reflect independent loss occasionally for a whole chromosome. Overall, we observed substantially more cells with at least one annotated SV in HAP1^{lig4-} cells (20.91%) compared to HAP1 (7.98%), excluding the known disomy on chr15 (FigS20E). The substantial increase in mutational SVs, particularly whole chromosome gains, compared to a relatively minor increase in SCE rate in the absence of classical NHEJ is unexpected. One possible explanation could be that a fraction of the observed SCEs are a result of NHEJ repair of isochromatid breaks instead of mitotic crossover.

Detection of Patski cells with a near-haploid genotype

Less restrictive clustering of Patski cells (29 total clusters) combined a set of 76 cells containing an almost ubiquitous level of segmental and chromosomal loss (Fig4A). Cells with similar characteristics were not found in any of the other examined cell lines. Certain cancers have been shown to aberrantly express meiotic genes that could lead to widespread chromosomal loss [59,60]. However, with only two haplotype switches per cell at most (FigS22A), meiosis-like division cannot explain the observed near-haploid genotype. Karyotype alterations in a single punctuated burst have also been observed in yeast [61] and human cancers [62]. A study in colorectal cancer organoids showed multipolar spindle defects as the causal mechanism [63]. These defects can give rise to cells with reciprocal karyotypes. Interestingly, a second cluster consisting of 29 Patski cells displays similar levels of segmental and chromosomal gain events (FigS22B), again suggesting a single punctuated burst event as the cause.

Clonal lineage suggests monosomy as necessary intermediate for UPD

Having observed chr4 LOH in Patski cell with sci-L3-WGS[8], we now confidently resolve the events into monosomy and UPD with sci-L3-Strand-seq (Fig4A) with the implementation of digital copy number counts (applicable to sci-L3-WGS as well), and most conclusively by observing both W and C reads for a single haplotype. UPD can arise from two sequential nondisjunction events with a trisomy followed by the loss of the non-duplicate chromosome or with a monosomy followed by its duplication. An alternative process leading to UPD in a single cell division was observed in yeast, termed reciprocal UPD (RUD), where chromosome homologs instead of sister chromatids segregate into reciprocal daughter cells in a meiosis I manner[6]. The presence of aneuploidy intermediates makes RUD an unlikely source of UPD in our diploid cells. Most of the clonal clusters consist of a mixture of cells with both chr4 monosomy and UPD that create mirror populations centered around the status of chr4 (Fig4A and FigS22B). Collapsing cells with identical events revealed chr4 monosomy had a higher diversity of events compared to chr4

UPD, suggesting monosomy preceded the appearance of UPD (FigS22C). We also observed two cells with a trisomy of chr4, in which the gained copy matches the lost haplotype in the monosomy, suggesting a reciprocal chromosomal nondisjunction produced the trisomy and the monosomy, and monosomy was the primary intermediate of UPD (the trisomy had the “wrong” duplicated haplotype). A cluster of cells that retain both parental copies of chr4 has a unique interstitial loss on chr2, and lack the highly prevalent chr3 and/or chr5 trisomy events, suggesting that they are more likely early founders in the population.

In BJ-5ta, both aneuploidy and UPD are particularly rare (Fig4D, FigS20A, TabS4). Previous analysis by single-cell WGS in BJ cells treated with UV radiation revealed almost no large-scale genome rearrangement [24], while UV of the same dose increases mitotic crossovers by 8,000-fold in yeast [64]. This suggests BJ cells could have an extremely stable genome that agrees with our observation. Alternatively, cells with aneuploidy could have a strong selective disadvantage in BJ-5ta cells. As our single-cell assay does not require cells to be viable, we would expect high levels of single non-clonal events if selection were the main reason for the observed low levels of aneuploidy. Nevertheless, the higher fitness of UPD compared to monosomy, due to UPD being copy-neutral, makes negative selection less likely, particularly if the UPD is formed in a reciprocal, meiosis-like manner of chromosome segregation. Despite their low rate, we observed UPD only when a monosomy for the same chromosome was also present in the population, again suggesting UPD needs to go through a monosomy intermediate.

Clonal lineage suggests recurrent SVs

Having identified several potential founder populations in Patski cells, we wanted to characterize the lineage structure of the various clonal populations using diffusion maps, a non-linear dimensionality reduction technique (FigS23A) [65,66]. We found four main branch points that can be broadly summarize by: 1) a segmental loss on chr2 (DC1), 2) chr4 LOH with combinations of chr12 monosomy, and chr3 and/or chr5 trisomy (DC2), 3) chr4 LOH with a combination of chr3 and/or chr19 trisomy (DC2), and 4) multiple trisomies encompassing up to 8 chromosomes (DC3 and 4) (Fig4B and FigS23B; see supplemental text). As mentioned above, the unique features of the chr2 loss cells, such as chr4 heterozygosity, suggest an early branching point from other prevailing clonal populations. The opposite haplotype of chr1 and chr15 trisomy in these cells, contrasting other clonal populations, together with a chr15 founder population, points to the same events arising multiple times independently (Fig4A, FigS22D). Similarly, the presence of both monosomy and UPD on chr4 with respect to other SVs, forming mirror populations, suggests several independent duplications of monosomy resulting in UPD occurred in the course of accumulating other SVs (FigS22B and D). The presence of chr4 LOH with only one of the additional trisomies on chr3, chr5, and chr19, or chr12 monosomy suggests these represent the earliest branching events of this lineage (Fig4B). Within these early branches the same set of events seem to have recurred, creating clonal populations with multiple combinations of events on both chr3 and chr12, or chr5 and chr12, or chr3 and chr19, or chr3, chr5 and chr12 (Fig4B and FigS22D). These, together with the multiple trisomy cells and the founder event clones reaffirm a striking predisposition of Patski cells for acquiring SVs within a preferred subset of chromosomes, a bias previously observed in yeast with perturbed checkpoint [67], and often seen in cancers [62,68,69].

SVs alter the rate of SCEs

An increased rate of SCEs has been observed in numerous cancers and can be a sensitive indicator of genomic stress and instability [70–75]. The significantly higher rate of SCEs in Patski cells compared to BJ-5ta ($p=3.76x10^{-04}$; Mann-Whitney U) (Fig2B), along with the 96.03% of cells having at least one

annotated event (FigS20E), raises the question of whether specific SVs coincide with increased rates of SCEs, or whether cells with unstable genome have higher instances of both. Focusing on dominant whole chromosomal events, we found cells with a chr4 UPD or a chr12 monosomy had a significantly higher number of SCEs (Fig4C). Since monosomy on chr12 is only present within cells that also have a chr4 monosomy or UPD, we broke down cells into five categories to disentangle the contribution of the different events to the rate of SCE (FigS24A, see Supplemental Text). We found chr4 monosomy decreases SCEs, while chr4 UPD and chr12 monosomy increase SCEs in an additive manner, suggesting the way the two events increase SCE rates are mechanistically independent (FigS24A, see supplemental text). Interestingly, neither the chr2 segmental loss nor trisomies on chr5, 15 and 19 had a significant impact on SCE rates, with only chr3 trisomy cells having a lower SCE rate compared to control cells ($p=2.18x10^{-04}$; Mann-Whitney U; FigS24B and TabS13).

Measuring the rates of SVs

As with SCE, the ability to measure the rate of various SVs arising in cells is invaluable to understanding genome instability in disease and in cellular perturbation experiments. The challenge is in disentangling the many branches of clonal, newly acquired, and recurring events to obtain the true rate of occurrence. Unlike with Patski cells, overcounting in haploid and BJ-5ta cells is less problematic due to the simpler clonal structure and smaller populations. As a compromise between preserving recurrent and removing propagated events we chose to remove the most frequent outlier events (FigS25A). Altogether, we excluded the known disomic region in haploid cells, five events across Patski cells, along with any near-haploid and high proportion of gain cells (FigS25A and B, see methods).

Comparing HAP1 and HAP1^{lig4-} cells, we found the biggest difference came from HAP1^{lig4-} cells having a significantly higher rate of whole chromosome gains compared to HAP1 cells (p-value: 5.32×10^{-12}, 2.5x; chi-squared) (Fig4A, FigS25B and TabS14). A more frequent formation of unresolved crossover intermediates in HAP1^{lig4-} cells that leads to nondisjunction events could explain this higher rate. The reciprocal event of whole chromosome gain in haploid cells is whole chromosome DTZ. However, we did not observe a higher rate of whole chromosome DTZ in the HAP1^{lig4-} cells, likely due to selection or lack of viability. The counterpart in diploid cells would be a trisomy accompanied by a reciprocal monosomy. For BJ-5ta cells we observed a similar number of both event types, supporting a reciprocal origin (Fig4A, Fig1F and TabS14). In contrast, Patski cells contain a significantly higher number of whole-chromosome gains that suggest additional selection (p-value 4.95×10^{-75}; chi-squared) (Fig4D, Fig4A, FigS25C and TabS14). The second most significant difference between HAP1 and HAP1^{lig4-} cells was in segmental DTZs, with HAP1^{lig4-} once again having the higher rate of occurrence ($p=5.18\times10^{-4}$, 1.7x; chi-squared) (Fig4A and FigS25B). Possible explanations for segmental DTZ could be break induced replication (BIR). Since sci-L3-Strand-seq only sequences the template strands, the conservative replication of BIR [76,77], where both strands are nascent, would lead to complete degradation of the DNA and manifest as a DTZ event in haploid cells and a segmental loss in diploid cells. Micronuclei fractionated from bulk nuclei could contribute to both whole-chromosome or segmental DTZ (see Supplemental Text).

Site of a HAP1^{lig4-} translocation shows an increased occurrence of SVs

After categorizing SVs into arm, terminal, interstitial or whole chromosome, we found that in both haploid cell lines, whole arm events were the most frequent (Fig4B, FigS25C-D). The relative proportions of the four categories were maintained between HAP1 and HAP1^{lig4-} cells, with the notable exception of terminal

DTZ that does not span the whole arm (Fig4B, FigS25C and D). Genome-wide distribution revealed HAP1^{lig4-} cells have a substantial number of terminal DTZ events on the short arm of chrX (chrXp), which are absent from HAP1 cells (FigS20B and C). The region on ChrXp corresponds to a “SCE” hotspot that we later found to be translocated with the acrocentric chr13 (FigS7B and F, see **Strand-state correlation identifies known and novel translocations**). Interestingly, we found this region to be susceptible to three patterns of loss. In Class 1, only the translocated portion of chrX is lost, leaving chr13 intact (8 cells). In Class2, the entire chr13 is lost along with the translocated portion of chrX (2 cells). We manually inspected these two cells and found segments containing reads on the centromere proximal portion of the acrocentric chr13 that explain how the translocated chrX fragment was retained (FigS25E). In Class 3, a break on chr13 leads to the loss of the majority of chr13 to the right of the translocation, thus leaving the translocated chrXp intact (2 cells) (FigS20C and FigS25E). The third pattern involving almost the entire right arm of chr13 suggests the chrXp translocation is very close to the chr13 centromere, or in the sub-telomeric region of the acrocentric left arm (FigS25E). We also observed similar patterns with gain events that overall points to the chrXp-chr13 translocation forming a fragile region prone to further genome instability (FigS20C).

Opposite strand reads identify reference or population inversions

Inversions are challenging to identify and, as a result, represent an understudied class of copy-neutral SV. The advent of Strand-seq is helping to address this gap [16,17,19], but even with Strand-seq, an inversion must be observed across multiple cells to be confidently identified, as it cannot be distinguished from a double SCE in a single cell. The large numbers of cells captured with sci-L3-Strand-seq makes identification easier, as the likelihood of observing subclonal inversions becomes greater. We identify inversions from the opposite strand of otherwise fully WW or CC chromosomes where: 1) a heterozygous inversion would show up as a region of WC, or 2) for homozygous inversions and haploid cells as WW in an otherwise CC chromosome, and vice versa (Fig5A and FigS26A, see supplemental text and methods). Overall, we identified 19 global (shared by HAP1 and BJ-5ta) inversions in human cells, along with 9 inversions unique to BJ-5ta cells, 2 unique to HAP1^{lig4-} and 5 shared by both HAP1 and HAP1^{lig4-} cells (Fig5B), with an overall median size of 300Kb (FigS26B). We found 65.7% of our inversion calls overlapped with previously described inversions [17], including 18 of our global inversions, 4 BJ-5ta unique inversions and 1 HAP1 and HAP1^{lig4-} shared inversions (Fig5B). For Patski cells, we identified a total of 38 inversions with a median size of 200Kb (Fig5B and FigS26B), with 1 overlapping with a limited study of 9 previously mapped inversions in C2 mouse embryonic stem cells. We also found the misoriented contig on chr14 in mm9 was still present in mm10 (FigS7D) [16].

Identifying large and sub-clonal inversions from breakpoint hotspots

Identifying inversions from the opposite strand of otherwise fully WW or CC chromosomes has three caveats: 1) opposite strand reads are either background or genuine inversions. Small inversions or inversions in a small percentage of cells regardless of their size may not meet our threshold to eliminate background reads, and could be missed; 2) using only WW or CC chromosomes is limiting; and 3) operationally we only use chromosomes without any breakpoints. We filter immediately adjacent strand switches (2 Mb in haploid and 10 Mb in diploid). Dual strand switches as a result of inversions smaller than the filter size are first removed, subjecting the chromosome to inversion analysis. However, larger inversions are omitted due to the presence of mapped SCEs. We therefore also called inversions by examining adjacent double SCE hotspots (Fig5C-D). SCEs are not clonal in nature and should not

co-occur in the same cell even at hotspots. Out of the four candidate regions identified across all four cell lines, we found two bona fide inversions. The first inversion is located on chr3 in only 9 HAP1^{lig4-} cells (1.2% of the population) measuring approximately 30.9Mb (Fig5C). We did not find an equivalent inversion in the parental HAP1 cells, suggesting it was newly formed within the knockout cells. The second inversion located on chr18 was found in 39 Patski cells (4% of the population) measuring approximately 9.7Mb (Fig5D). Examining the strand states of the region shows the inversion is heterozygous. Our results show that sci-L3-Strand-seq can identify both global and sub-clonal inversions.

Strand-state correlation identifies known and novel translocations

HAP1 cells have been previously shown to contain an insertional translocation of chr15 into chr19 and a balanced translocation between chr9 and chr22 (Philadelphia chromosome) with a BCR-ABL fusion [31,42]. Visualizing chr9 and chr22 in all HAP1 cells shows matching strand states of the fused segments with a detectable strand switch in the expected 50% of the cells (Fig6A). As with inversions, a translocation breakpoint needs to be matched across multiple cells to distinguish it from a SCEs. Based on this, we reasoned that from hundreds of single cells we could generate genome-wide strand-state correlation maps that would highlight the location of translocations. We calculated the Pearson correlation between 1Mb sized bins in 439 HAP1 cells that each contained more than 20 reads per Mb and plotted the resulting correlation matrix (Fig6B, see Methods). We found strong positive correlation between the expected chr15 disomic region and chr19 as well as between the terminal points of chr9 and chr22 (FigS27A and Fig6C), while the negative correlation highlighted some of the inversions (FigS27D-E). With 62 high-coverage HAP1^{lig4-} cells, we observed both known translocations alongside a new balanced translocation that arose between chr13 and chrXp (Fig6D and FigS27B). Additionally, we explored using cells with lower coverage at less than 20 reads per Mb and with a lowered resolution using 5Mb bins. From 689 cells we obtained three clear translocations with strong strand state correlations off the diagonal (FigS27C). Aside from translocations, the correlation heatmap also captured the inverted duplication event on the terminal end of chrX in HAP1^{lig4-} cells (Fig6D and FigS7F). We did not observe but expect chromosome-level positive correlation for telomere-telomere fusion of non-homologous chromosomes, and a decreased representation of WC chromosomes (50% expected) if homologs are fused. Neither BJ-5ta nor Patski cells contained any detectable translocations. For a subset of BJ-5ta cells the correlation heatmap likewise captured the inverted duplication at the start of chrX (FigS27D-F and FigS7E). Altogether, we show that sci-L3-Strand-seq enables us to detect both unbalanced and balanced translocations genome-wide.

Resolving complex events provides new insights into the HAP1 disomic region

Having identified two inversions on chr15 in HAP1 cells, we unexpectedly found that one was located within the known disomic region and was never identified in previous studies (Fig7A) [27,42]. The duplication of chr15 was produced by an insertional translocation into chr19 (Fig6) and it must have originated from the lost parental chr15 prior to haploidization since it contains heterozygous SNVs [78]. From this, we know the strand state of the disomic region is a composite of chr15 and chr19 (Fig7B). Therefore, since chr19 shows a matching strand state with the gained copy, but opposite to the inversion, we know the inversion must be heterozygous and located on chr19 (Fig7B and C, FigS28). We can also confidently rule out the other two possibilities of 1) a heterozygous inversion on the chr15 copy, or 2) a homozygous inversion on both the chr15 and chr19 copies (see Supplemental Text). Resolving complex rearrangement involving

duplication, translocation and inversion in the same region and providing the likely order of events highlights the power of sci-L3-Strand-seq for dissecting complex SVs.

Discussion

Here we describe the extension of sci-L3 methods to Strand-seq that enables the detection of error-free SCE and mutational crossover outcomes in thousands of single cells (Fig1). The development of sci-L3-Strand-seq with complementary computational approaches offers a low-cost comprehensive assay for a multitude of applications including the study of HR repair, SVs and genome instability outcomes, along with high-throughput mutational screens.

With a 2-day experiment (FigS2), sci-L3-Strand-seq produces single cell whole genome libraries with strand information without any specialized equipment at a cost of $0.15 per cell for 10,000 cells and appreciably less with greater cell numbers; on par with the previously published sci-L3-WGS [8]. The scalability, and the resulting sub-saturation sequencing and low coverage of single cells motivated development of new computational approaches. Below we discuss the development of new genome segmentation methods, leveraging iteratively information at both the single-cell and population level, and new observations of genome instability enabled by these new developments.

Given the mechanistic nature of mitotic crossover, the resulting genome rearrangement is typically large, involving millions of bases. If we know the junction at a transition from heterozygous to LOH region, random allelic dropout at heterozygous sites due to low coverage is no longer an issue: hundreds of single-read covered SNVs to the left of the transition (heterozygous region) will collectively have a 0.5 haplotype fraction, and 0 or 1 for the LOH region on the right. The key to address sparsity in coverage is therefore proper chromosome segmentation. To our knowledge, we developed the first haplotype-specific strand switch-based approach for chromosome segmentation, which uniquely enabled detecting cnLOH and is conceptually novel from segmentation by HMM or fixed bin size [13,18,32,33]. NAHR between non-allelic repeats is typically more genetically tractable with various reporter assays [79–82] and genome-wide with population studies [83,84], although for the latter, it is difficult to attribute the NAHR event to meiotic or mitotic origins. Allelic HR involving non-repetitive sequence homology from the sister chromatid or homolog has been difficult to map genome-wide. The original Strand-seq addresses the gap in mapping SCE, and here, haplotype-aware chromosome segmentation further allows us to dissect inter-homolog crossover with recombined chromatids co-segregating into the same daughter cell without causing cnLOH (FigS18C) or with recombined chromatids segregating apart into different daughter cells leading to cnLOH. We envision that the analysis framework pioneered here will eventually lead to knowledge on HR partner choice of: 1) homology on the sister chromatid, 2) allelic homology on the chromosome homolog, and 3) non-allelic repeats.

In addition to haplotype-aware chromosome segmentation, at the single cell level, we also provide SVM methods to integrate multifaceted genotype information, including strandedness, haplotype fraction using phasing inherent to the dataset, digital counting of copy numbers, and their combinations. The digital copy number counts were uniquely enabled by employing tagmentation-based genome fragmentation, overcoming unevenness in IVT-based linear or PCR-based exponential whole genome amplification in single-cell sequencing. These single-cell level analyses, together with haplotype-aware segmentation, were built into an iterative process to leverage population information from a large number of cells.

Here we summarize the five scenarios where our analyses go through single-cell level and population level cycles. First, we initially treat all strand switch calls as potential “SCE” (single-cell level), and examine single or double “SCE” hotspots for SV such as deletions, inversions and translocations (population level). This not only allows us to sensitively call subclonal SV sometimes present in only 1% of the cells, but also helps filtering potential SCE calls (single-cell) that are in fact boundaries of SVs. Second, we start with all the cells for initial phasing of SNVs (population), and then call LOH events with strand-switch segments and imperfect phasing (single-cell). We also survey all the cells for regions with LOH calls and rescue the 50% of LOH expected to not have a strand-switch (population). Identifying the set of cells with LOH then collectively feedback to refine phasing (population level), orthogonal to strand-based phasing. Improved phasing accuracy allows better LOH calls and rescues (single-cell). Third, both LOH called with strand switch-defined segments (single-cell) and rescued LOH segments (population) were used as input of SVM for SV annotation. This strategy was devised to account for SV not associated with strand switches, but it also enabled better event calls in cells with particularly low coverage. Sci-L3 produces single cell libraries with 10-fold coverage range. Most ultra high throughput single cell sequencing methods have similar or higher variability of per cell coverage. Cells with higher coverage can thus serve as bait to inform segmentation. Fourth, similarly, cnLOH identified by haplotype-aware segmentation can rescue cnLOH without haplotype-specific strand switches. Fifth, we build clonal maps (population) from all annotated SVs (single-cell), and the clonal structure of specific events (population) help inform manual validation for training SVM (single-cell). In sum, we tackle the throughput vs. coverage trade-off by building iterative analysis processes that go through cycles of single-cell and population level information. We discuss eleven new bioinformatics tools in Supplemental Text developed and employed in this process, several of which are also applicable to sci-L3-WGS.

Combining improved analyses of all the seven types of mitotic crossover outcomes and the genome-wide and scalable nature of sci-L3-Strand-seq, we summarize new observations and insights gained from the examined mouse and human cell lines. We mapped an unprecedented number of 21,761 SCEs from a total of 5,826 single cells, which allowed us to examine subtle mutant phenotypes. Upon a knockout of the classical NHEJ pathway by deleting *LIG4*, we observed a moderate but significant increase in SCEs. Despite its error-free nature, an increased rate of SCEs can be a sensitive indicator of genomic stress, instability, and possibly act as an early biomarker of tissue specific susceptibility to cancer[70–75]. In addition to SCE, we also observed elevated levels of SVs in our *LIG4* knockout cells, including a somewhat surprising 2.5-fold increase in whole chromosome gain, suggesting the increase in HR repair may lead to an increase in gross chromosome rearrangement.

Sci-L3-Strand-seq enables examining crossovers emanating from a large-cluster of repeats like the centromere as long as there are enough unique sequences flanking the repeats. We examined the rates of SCEs within centromeres of human non-acrocentric chromosomes and observed a decreased rate of SCEs. The preference for HR in centromeres is central to their maintenance and rapid divergent evolution[85–87]. Repeat amplifications and deletions in centromeres can be produced with single-strand annealing (SSA) and gene conversion between unequal repeats. With our observed decreased rate of SCE, it is possible crossovers are actively suppressed in favor of SSA and gene conversions[88]. The size and composition of human centromeres can greatly vary between individuals or cell lines[89]. Therefore, the CHM13 centromere lengths used may not be an ideal representation of our cell lines.

Interestingly, in contrast to centromeres we found late replicating domains and several repeat classes, namely LINEs and LC regions, to be enriched for SCEs. These could be explained by either a higher rate of initiating lesions, such as non-B structure forming repeats, or a higher propensity of certain regions to utilize HR over other repair pathways[44,45,54,55]. Alternatively, it has been recently proposed that SCEs can arise independently of HR from stalled replication forks through POLQ-mediated end joining[90]. The study

utilized the haploid KBM-7 cell line (HAP1 parental line) and PARP inhibitor (Olaparib) treatment to induce replication fork stalling that produced under-replicated DNA in mitosis. In agreement with our data, they observed an increased rate of SCEs within CFS, suggesting our cells may sometimes be utilizing an HR independent repair pathway in difficult to replicate regions. We specifically chose to study spontaneous SCEs without damage-inducing agents. The correlation between SCE abundance with END-seq (profiles DSB) but not GLOE-seq (profiles SSN) suggests that DSB is the main source of spontaneous SCEs. Uniquely to inter-specific F1 hybrid mouse cells, we found a positive association with sequence divergence, suggesting that the sister chromatid is the favored repair template compared to diverged donor, likely due to the anti-recombination role of mismatch repair [52].

Aside from error-free SCEs, sci-L3-Strand-seq enabled us to examine different types and relative abundance of mutational SVs. With haplotype-aware segmentation that enabled cnLOH calls we found that segmental deletions are significantly more abundant compared to cnLOH (2x in BJ-5ta and 19x in Patski), contrary to analysis of mosaic chromosome alterations in the hematopoietic lineages [40] but in agreement with neuronal disorders [91]. We also mapped a subclonal inversions involving as few as 9 cells (1% of the population), and resolved a complex rearrangement involving a duplication, a translocation and a previously unidentified heterozygous inversion present only in the translocated copy in the HAP1 cell line.

Mutational SVs allow us to build clonal lineages and dissect the sequential order of events. By examining subclonal structures, we found that: 1) monosomy seems to be a necessary intermediate for UPD in diploid cells; 2) DTZ in diploid arise in cells with pre-existing loss on the corresponding chromosomes; and 3) recurrent SVs from clonal map suggest certain chromosomes are hot for repeated duplication. In meiosis, crossovers engage homologs to provide necessary tension for chromosomes to be pulled into different poles in Meiosis I. Similar phenomena have been observed in the formation of reciprocal uniparental disomy (RUD)[6]. However, we observed several lines of evidence that monosomy preceded UPD in the clonal histories of both Patski and BJ cell lines. Similarly, the observation that DTZ in cells with matching loss events in the population suggests DTZ in diploid cells requires a loss intermediate and thus is a two-step event. As for recurrent chromosome gain, McCulley and Petes[67] previously showed that yeast have significantly increased trisomy rate specifically for chromosomes II, VIII, X and XII in the checkpoint-defective diploid strains. Similar genome destabilization and/or selection mechanisms causing chromosome specificity could be shared in mammalian cells. Copy number changes are an almost universal feature of cancers, impacting fitness and cancer progression[92,93]. Patterns of whole chromosome and arm gains have been shown to be tumor specific and non-random, with recurrent events being linked to malignant transformation and growth [69,94,95]. These recurrent events arising at different time points may also suggest that statistical modeling of tumor heterogeneity needs to assume polyclonal origins rather than using a strong prior belief that tumor is monoclonal [96].

The ability to simultaneously examine error-free SCE and mutational SVs could allow us to look at features differentiating these crossover outcomes, and their crosstalk. With the limited number of SVs observed in this study, we did not identify (epi)genomic features that clearly separate error-free SCE and mutational SVs. While they may share the same initiating lesions, repair partner choice of identical sister, homolog harboring heterozygosity, and non-allelic repeats remains to be explored in a larger scale study. We were able to link certain SVs with a variation in the level of SCEs, suggesting specific CNVs cause genomic instability. Similarly, in the example of chr13-chrX in HAP1^{lig4-}, a translocation became a source of genome instability, resulting in a higher rate of SVs. In line with human cancer studies, these events may represent early aneuploidies that drive malignant transformation[95,97], suggesting sci-L3-Strand-seq could be a powerful assay for detecting early cancerous states.

In addition to the Patski cells with gradual diversification, we found a surprising number of near-haploid cells. Several instances of haploidization have been previously described, including yeast cell

haploidization in the absence of the homologous recombination protein RAD52[98] and division without DNA replication observed in the skin cells of zebrafish[99]. However, most notably, near-haploid karyotypes have been observed in leukemias and solid tumors[100,101], with the HAP1 parental cell line (KBM-7) itself being derived from a chronic myeloid leukemia (CML) patient[102]. We also observed cells with near-triploid karyotypes, suggesting they may represent reciprocal events of the near-haploid cells. Such punctuated bursts of karyotype alterations in a single cell division have been characterized in patient-derived tumor organoids of colorectal cancer[63]. They showed such events can arise from multipolar spindle defects, where genome-duplication and mitotic failure can lead to tripolar division that produces three daughter cells with extensive karyotype alterations.

Limitations

The limitations of sci-L3-Strand-seq include the low sequencing coverage that restricts the resolution of events[8]. As a result, pinning down chromosomal elements as the source of events may be hard, especially for error-free SCE. Events such as microCNV as well as small and complex rearrangements in chromothripsis may not be fully detected[18]. These resolution issues can be partially resolved by increasing the number of cells sequenced, if “mini-bulk” of cells with shared events can be identified by examining subclonal structures. While the resolution of SCE is difficult to refine for single events, piling up more SCEs genome-wide may eventually lead to better understanding of underlying genomic elements involved. Sci-L3-Strand-seq provides the technical throughput for exploring clonal lineages. However, it remains difficult to resolve the temporal order of all the events with certainty. This limitation implicitly affects accurately estimating the rate of mitotic crossover outcomes. For example, SCE rate is mapped for the single cell division for BrdU labeling, but the detection of inversions and translocations requires multiple single cells (nine cells in an example subclonal inversion we were able to map), preventing the identification of newly arisen events without minimal clonal expansion. Additionally, our assay uses nuclei without distinguishing their source of live vs. dead cells, although this has the advantage to look at raw DNA repair outcomes, staining for nuclei from live cells may be explored if needed for certain biological applications. Lastly, sci-L3-Strand-seq still requires BrdU labeling and thus is not applicable to fixed samples, rendering it less accessible for clinical samples. It is also less generalizable compared to sci-L3-WGS for model organisms without established BrdU labeling methods. Our development of haplotype-aware segmentation uniquely detects cnLOH. However, both cnLOH or co-segregated inter-homolog crossovers are rare in mammalian systems, except in early embryogenesis[103]. Our methods have a greater potential in systems with smaller nuclei like yeast, for meiosis with homolog pairing, and organisms with somatic homolog pairing like dipteran insects including fruit fly and mosquitos.

Conclusion

Sci-L3-Strand-seq substantially expands on the sci-L3 methods, enabling the detection of SCEs, inversions, balanced translocations, and chromosome level haplotype phasing useful for analyzing LOH and CNV. It additionally improves upon Strand-seq methods with Tn5-based digital counting and its linear amplification, high-throughput, and low-cost implementation. We develop computational methods that directly tackle the throughput vs. coverage trade-off, and iteratively leverage single-cell and population level information for comprehensive analyses.We demonstrated sci-L3-Strand-seq provides a quantitative view of crossover outcomes, with insights into clonal population dynamics. The generalizability of sci-L3 enables future adaptation of sci-L3-Strand-seq to other modalities such as RNA, enabling the examination of SCEs in populations of mixed cell types to address cell-type specific error-free and mutational HR [25].

We anticipate sci-L3-Strand-seq will have a wide range of applications ranging from the study of basic molecular mechanisms to obtaining insights into complex diseases.

Resource availability

Lead contact and materials availability
Further information should be directed to and will be fulfilled by the Lead Contact, Yi Yin (yiyin@mednet.ucla.edu).

Data and code availability
Raw sequencing data and processed data have been deposited at GEO and are publicly available as of the date of publication. Additional raw data were obtained from GSE281238. All original code and supporting data has been deposited at Zenodo and are publicly available at 10.5281/zenodo.15036453 as of the date of publication.

Acknowledgments

This work was supported by National Institute of General Medical Sciences (5R35GM142511 to Y.Y) and Damon Runyon Cancer Research Foundation (DFS-43-20 to Y.Y). T.R. was supported by the National Institute of Health Training Grant in Genomic Analysis and Interpretation (T32HG002536).

Author Contributions

Conceptualization: P.C., T.R., Y.Y.; Data curation: P.C, T.R., Y.Y.; Software: P.C., T.R.; Formal analysis: P.C., T.R., Y.Y.; Funding acquisition: Y.Y., T.R.; Investigation: P.C., T.R., Y.Y.; Methodology: P.C., T.R., Y.Y.; Supervision: Y.Y.; Writing - original draft: P.C., T.R., Y.Y.; Writing - review & editing: P.C., T.R., Y.Y.

Declaration of Interests

The authors declare no competing interests.

Supplemental information

Document S1. Text and Figures S1–S28

Table S1: Related to FigS3 and FigS4. Breakdown of cell counts across QC filtering steps. Cells output from the sci-L3 pipeline are further filtered for coverage, background, and overall strand state percentage. In addition, several cells were removed by manual inspection of cells with more than 10 breakpoints. Each column shows the number of cells removed by each individual filter. The final cell counts show the cells passing all the QC filters.

Table S2: Related to FigS7. Coordinates and annotation for hg19 and mm10 region filter. The region filter primarily consists of strand switch hotspots that overlap with called inversion and translocations. The disomic region in HAP1 and HAP1lig4- cells was obtained from the manual merging of several hotspot regions. Similarly, the region on chr12 in Patski cells consists of two manually merged hotspots. The hg19 centromeric regions for chr1 and chr16 were blacklisted due to a large number of false breakpoints, in addition to the mm10 reference inversion on chr14. chrY in BJ-5ta cells, similar to chr16, also contained regions with high numbers of false strand switches corresponding to an inversion, an inversion proximal region, along with a region requiring blacklisting.

Table S3: Related to Fig2D. Significance of SCE enrichment for 1Mb bins genome-wide. SCE counts were normalized by the number of bins each breakpoint overlaps with, the total number of cells, and ploidy. SCE enrichment was determined by fitting a gamma distribution and calculating p-values using the cumulative density function, adjusted using the Benjamini-Hochberg approach (see Methods).

Table S4: Related to Fig2D, Fig4 and FigS25. Coordinates of SCE breakpoints and SV annotations. Each SV annotation (col1) is broken down by chromosomal arm, adjacent annotations (col2 - adjacent left; col3 - adjacent right) and classifies the event by location (col4; e.g. arm, terminal, interstitial or whole chromosome). In addition, regions derived from LOH rescue (col5), and cells with high number of gains, cells with high number of loss (near-haploid cells) or cells with spikes are marked (note). The strand states and deduplicated W and C counts are shown for each annotated region. Adjacent annotations include the chromosome left end (LE), chromosome right end (RE), and the centromeres (CEN) (FigS25A). Breakpoints are broken down into SCE and SV annotation adjacent.

Table S5: Related to FigS10. Significance testing of each pairwise correlation between genomic features, SCEs and mutational SV breakpoints. Results of a test for the correlation between paired samples. Pairwise Spearman correlation between SCEs or mutational SV breakpoints and genomic features was calculated and tested for significance (*** < 0.0005, ** < 0.005, * < 0.05). Bonferroni was used for multiple comparison correction.

Table S6: Related to Fig3A. Patski cell phasing accuracy by chromosome.
Detailed breakdowns of phasing accuracy with respect to total heterozygous SNVs called, total heterozygous SNVs with WC coverage, and total phased sites for each of five phasing configurations and their combination: 1) raw phasing with StrandPhaseR, 2) inferred phasing where sites with only 1 phased base have the opposite base inferred for the missing haplotype, 3) phasing with conflicting sites removed (where the same base was assigned by StrandPhaseR for both haplotypes), and 4) phasing with chr4 clonal LOH, 5) inversion correction where the haplotype assignments are reversed inside inverted regions of chr14 and chrX.

Table S7: BJ-5ta cell phasing statistics by chromosome.
Phasing statistics for BJ-5ta cells. Same as for TabS6, except missing accuracy statistics since BJ-5ta does not have a phasing ground truth. Total biallelic SNVs, SNVs with WC coverage, number of cells used for phasing, and phased sites for 3 different configurations: 1) raw phasing with StrandPhaseR, 2) inferred phasing, and 3) conflicting site removal.

Table S8: Patski LOH calling statistics.
General statistics at each major step of Patski LOH calling, including raw LOH calls from breakpointR segmentations, chromosome arms, and rescues. Further statistics include the combined LOH calls for

segmentations and whole arm fractions, where the LOH arm call is used when segmentation is unavailable on that chromosome arm, are also shown. LOH events without merging that were used as inputs to the SVM classifier are summarized in columns E & I, together with the statistics from LOH tables post-merging and post-manual updates.

Table S9: BJ-5ta LOH calling statistics.
General statistics at each major step of BJ-5ta LOH calling, same as in TabS8.

Table S10: Related to Fig3B and FigS13. Counts of LOH events segmented based on strand switches and arms versus rescues. Based on our inability to segment LOH events in WW and CC regions (50%), the expectation is that we will not have the correct segmentation for 50% of segmental events. With the assumption of most events being clonal, we can use existing LOH calls to recover ("rescue") the missing segmentation and enable LOH event annotation.

Table S11: Related to FigS18 and FigS19. Manual examination of additionally annotated events resulting from haplotype-aware segmentation. Examples from both Patski and BJ-5ta cells are shown with original segmentation (TabS4) (1), haplotype-aware segmentation with supplemental table merging (2), raw haplotype-aware segmentation (3), and manually resolved annotation that consists of a combination of raw haplotype-aware segmentation and manual merging into whole chromosome events (4).

Table S12: Related to FigS16 and FigS20. Comparison of HAP1^{lig4-} annotated using SVM model from HAP1^{lig4-} or HAP1 cells. HAP1^{lig4-} data was annotated independently with an SVM model trained using either HAP1 or HAP1^{lig4-} manual annotations. To match annotated events, at least a 20% overlap was required. All additional gain events annotated with the HAP1 SVM model correspond to the known chr15 disomic region. DTZ events not annotated with the HAP1 SVM model are dispersed across numerous chromosomes. Manual inspection of the DTZ events revealed they were false positive and borderline cases, leading to the use of the HAP1 SVM model for both cell lines.

Table S13: Related to Fig4C and FigS24. Summary statistics of SCE counts per groups of mutational SV outcomes. The number of SCEs for cells with whole chromosome clonal events (trisomy, UPD, monosomy) and a combination of thereof (chr3 and chr5 trisomy, chr4 and chr12 monosomy, chr4 UPD and chr12 monosomy) were examined to determine if there is a relationship between mutational crossover outcomes and the rate of SCEs. The "other" category contains all cells not present within the main category. The "chr19 no annotation" category is limited to cells with chr3 trisomy and chr4 monosomy, such that they match the cells with chr19 trisomy in all aspects except chr19 itself.

Table S14: Related to Fig4 and FigS25. Categorical counts of annotated events and significant difference testing. Significant differences between annotated events within and between cell lines were calculated using chi-squared. Pairwise comparison with contingency tables and fold change are included for between cell comparisons. Additional tables include breakdowns of events by their chromosomal location.

Figure legends

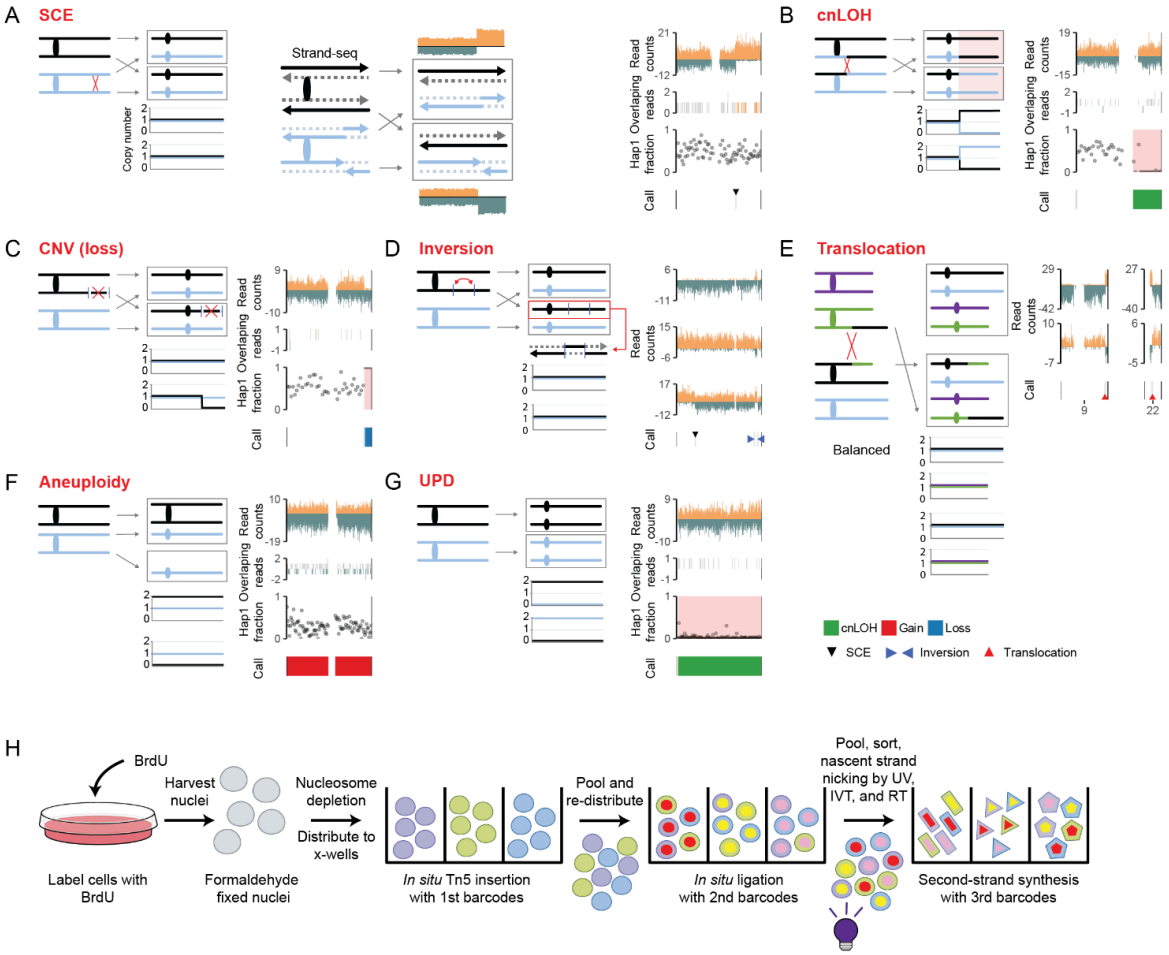

Figure 1: sci-L3-Strand-seq enables the detection of all seven mitotic crossover outcomes.

(A-G) Schematics of the seven types of mitotic crossover outcomes with example sci-L3-Strand-seq data. Homologs with SNVs are shown in black and blue (green and purple for translocation), each containing two identical sister chromatids. Oval indicates centromere. Read counts in the example data show Watson (W) reads colored dark yellow and Crick (C) reads colored dark green. This color code for strand directionality is used throughout this paper unless noted otherwise. Overlapping reads in the example data show the locations of overlaps, with nearby locations offset on the y-axis. Each location belongs to one of three categories based on the type of overlap: dark green represents overlapping C strand reads, dark yellow represents overlapping W strand reads, and gray represents overlapping reads of opposite strands. Each point for the haplotype1 (hap1) fraction is the phased allelic fraction of approximately 50 binned SNVs for Patski and 15 for BJ-5ta. LOH calls are shaded in pink.

(A) Sister chromatid exchanges (SCE). Both the chromosome heterozygosity and copy number remain unchanged after SCE. Only a strand-switch on the replication template strand (solid arrows) can be detected by Strand-seq. Without an SCE, the template strands map continuously to either Watson or

Crick (black homolog). With an SCE, the template strands are exchanged. In the example, SCEs are marked by a black triangle in the “Call” track.
(B) Copy neutral loss of heterozygosity (cnLOH). Inter-homolog (IH) crossover leads to reciprocal cnLOH (shaded in pink) in the daughter cells. Note that with reciprocal events, if both daughter cells are sequenced together with bulk WGS, every position will appear heterozygous given the 1:1 ratio of the two alleles.
(C) Copy number variation (CNV). Insertion or deletions between intra-chromatid sequences with high similarity as an example, resulting from non-allelic homologous recombination (NAHR).
(D) Inversions. Intra-chromatid reversal of sequences end-to-end, for example between repeats, results in no change in chromosome heterozygosity or copy number. Just like with SCE, a double strand-switch on the replication template strand (two inward facing blue triangles) can be detected by Strand-seq. However, confident identification of inversions requires the detection of a shared strand state across multiple cells that negatively correlate with the stand state of the adjoining regions.
(E) Translocations. Exchange of sequence between non-homologous chromosomes. Translocations can be unbalanced, where recombined chromatids segregate apart and result in the loss or gain of chromosomal segments in the daughter cells (not shown); or balanced, where the co-segregation of recombined chromatids results in an equal exchange of chromosomal segments with no change in copy number of heterozygosity. Similar to inversions, the identification of balanced translocations requires the detection of a shared strand state across multiple cells that positively correlates (as opposed to negatively correlating in inversions) between the translocated chromosomes. In black and blue, green and purple are the homologs of two different chromosomes.
(F) Aneuploidy. Aberrant crossovers can cause nondisjunction, resulting in monosomy and trisomy in the reciprocal daughter cells. Example data (trisomy) is shown for only one of the daughter cells.
(G) Uniparental disomy (UPD). Reciprocal UPD is shown with example data for only one of the daughter cells. Other routes leading to UPD include the duplication of a monosomy or the loss of the non-allelic chromosome from a trisomy.
(H) sci-L3-Strand-seq workflow overview. IVT: in-vitro transcription; RT: reverse transcription. See FigS2.

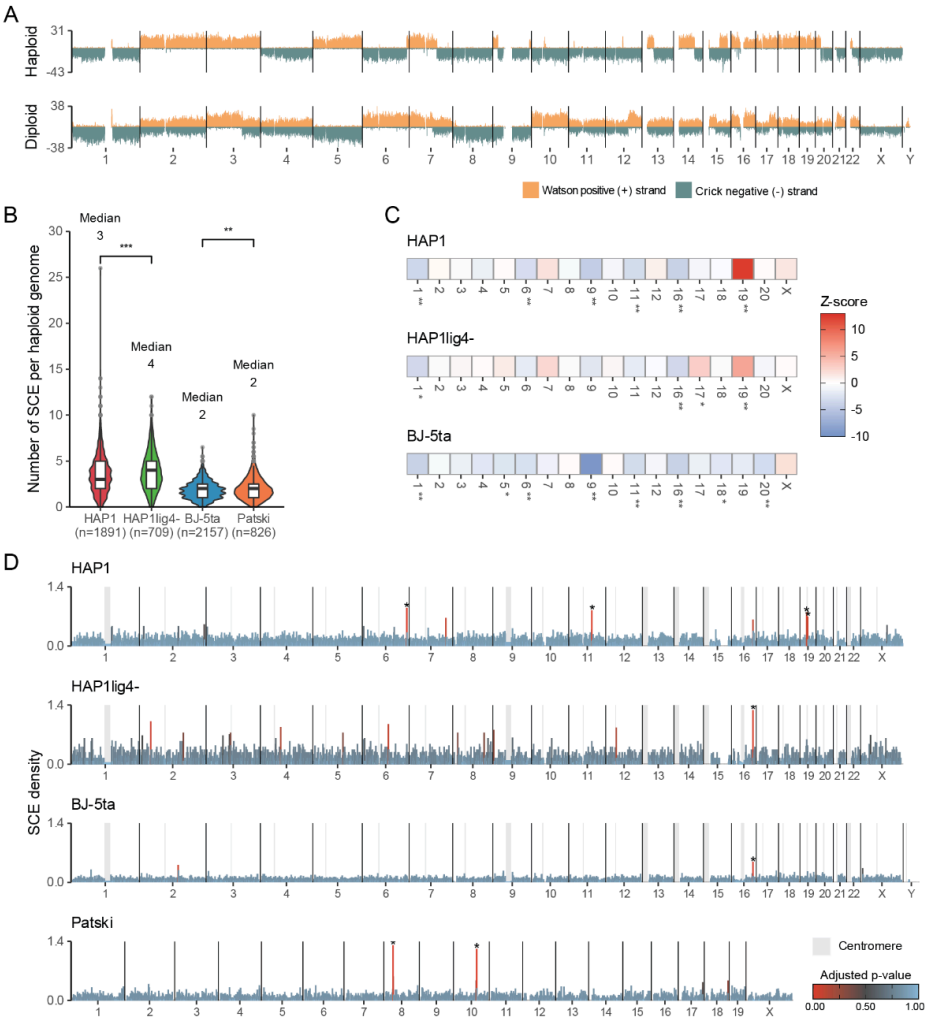

Figure 2: Sister chromatid exchange rate and genomic hotspots.6

(A) sci-L3-Strand-seq example data. Top panel shows a haploid human cell (HAP1; median reads per Mb: 97, about 2.1% coverage), while the bottom panel shows a diploid human (BJ-5ta; median reads per Mb: 99) cell. Reads were counted across 200 Kb bins.

(B) Number of SCE per haploid genome for each cell line. Significant difference level (*** < 0.001, ** <0.01, * <0.05) calculated using Mann-Whitney U, HAP1 vs HAP1^{lig4-} (BH adjusted p-value: 0.00038; MAD for HAP1 1.48, HAP1^{lig4-} 1.48) and BJ-5ta vs Patski (BH adjusted p-value: 0.003; MAD for BJ-5ta 0.74, Patski 1.48). Cells with a high number of loss or gain annotations were excluded.

(C) SCE centromere depletion and enrichment. The permutation test z-scores highlight the centromeres with an SCE enrichment (red) or depletion (blue). Before testing, the centromeres were resized to the T2T-CHM13 width to more accurately determine the expected SCE frequency within the centromeres. In addition, the centromere of chr1 was extended upstream to encompass a proximal inversion that masked centromeric breakpoints. For each non-acrocentric chromosome, the number of SCEs overlapping centromeres was counted after 1000 random permutations of SCE coordinates. The significance levels (*** < 0.0005, ** <0.005, * <0.05 p-value) and z-scores of enrichment or depletion were obtained using the permutation test [104]. Human acrocentric and mouse chromosomes have poor p-arm mappability, preventing examination of SCE across the centromere.

(D) Genome-wide SCE density. The distribution of SCEs are plotted at 1Mb resolution and normalized by the number of bins each SCE overlaps with, the total number of cells, and ploidy of each cell line. Bins with significant enrichment of SCEs are highlighted by an asterisk (BH adjusted p-value < 0.05). The genome-wide SCE enrichment p-values from 1Mb bins are provided in TabS3. Coordinates of SCE breakpoints are provided in TabS4. Centromeres are highlighted in gray. Raw breakpoint count distribution is shown in FigS7. The elevated rate of SCEs in HAP1^{lig4-} cells was accompanied by a higher variance in the genome-wide distribution (variance HAP1: 0.008; HAP1^{lig4-}: 0.018), rendering the hot regions less likely to be statistically significant.

Figure 3

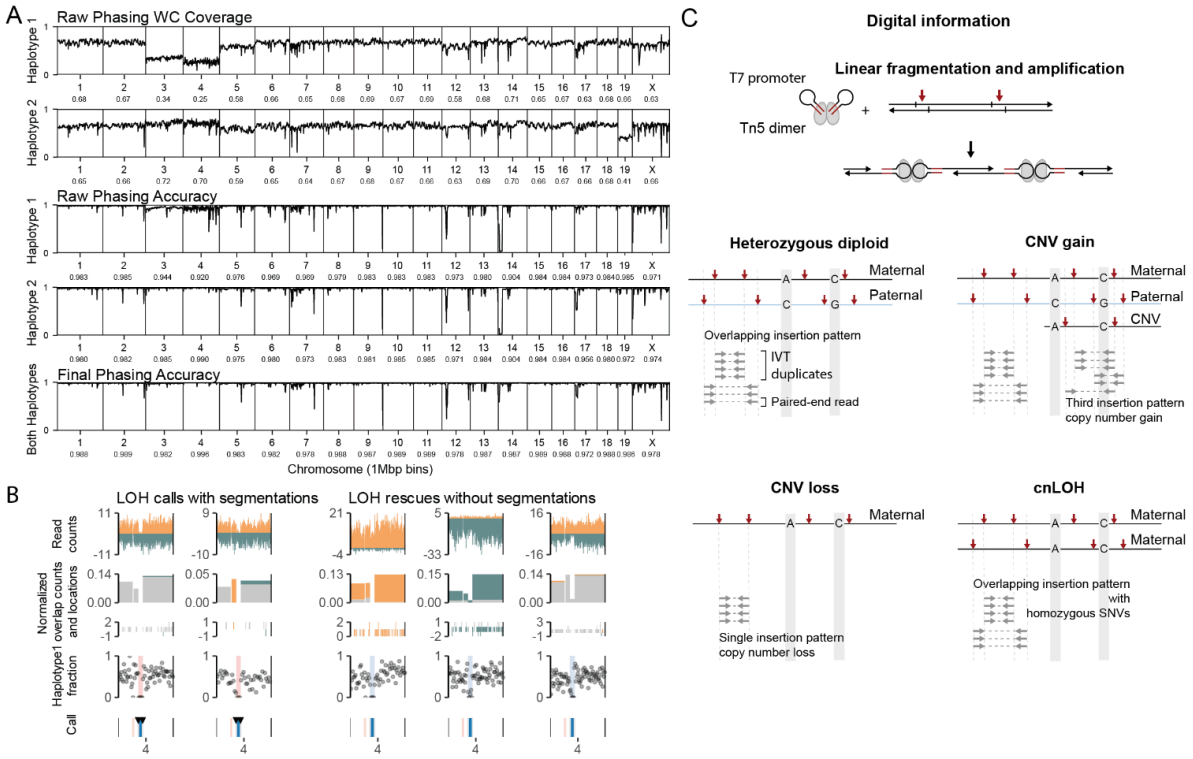

Figure 3. Phasing of SNVs and digital information for calling LOH and CNV.

(A) Phasing accuracy of Patski cells across the mm10 genome. The phasing accuracy was plotted at a 1Mb bin resolution. In general only WC read (around half) will be informative for phasing. The SNV coverage from reads in the WC regions is shown in the top two panels, with the mean chromosome coverage shown under each chromosome label. Phasing by sci-L3-Strand-seq was compared to the existing references to evaluate accuracy. Raw phasing accuracies are slightly different between the two haplotypes (3rd and 4th panel). Bottom panel shows the final phasing accuracy after inference of the opposite haplotype, conflict removal, phasing with chr4 LOH, and inversion correction (FigS12, TabS6-7). Sharp decreases in accuracy can be explained by inversion (FigS13).

(B) LOH calls and rescues. LOH calls (pink) on chr4 in two BJ-5ta cells are shown (left panels) with strand switch-based segmentation (two black triangles within call track). Three additional BJ-5ta cells (right panels) are shown without breakpoint calls for which the LOH calls were rescued (blue) and segmentation based on shared LOH with other cells. Read counts and overlapping reads are plotted as in Fig1A. Red vertical lines within the call track represent centromere bounds. Breakdown of LOH calling and rescue is provided in TabS8-10.

(C) Diagram of digital information within overlapping reads used for chromosome copy number counts. Random genome fragmentation, barcode insertion and T7 promoter ligation (red arrows) is mediated by the Tn5 transposase dimer. IVT amplicon structure is depicted post ligation. The insertion leaves behind a 9 nt single-stranded overhang that is shared between adjacent fragments and ligated to the end of the Tn5 mosaic end duplex sequence (highlighted in red). Due to this shared 9nt sequence, to identify overlapping reads, an overlap is required to be greater than 12 bp (9 nt gap + 3 bp offset for deduplication). With IVT, each unique genomic fragment is linearly amplified that creates blocks of IVT duplicates (gray horizontal dashed lines with inward arrows). A disomic region (top left panel) will contain at most one overlap of two unique insertion patterns (ignoring IVT duplicates), each of which must

originate from one of the parental homologs if heterozygous. In the event of a CNV gain, an overlap with three unique insertion patterns will be observed (top right panel). With a CNV loss, there will be no overlapping insertion patterns observed and only homozygous SNVs will be present (bottom left panel). The presence of overlapping insertion patterns and homozygous SNVs distinguishes cnLOH from CNV loss (bottom right panel).

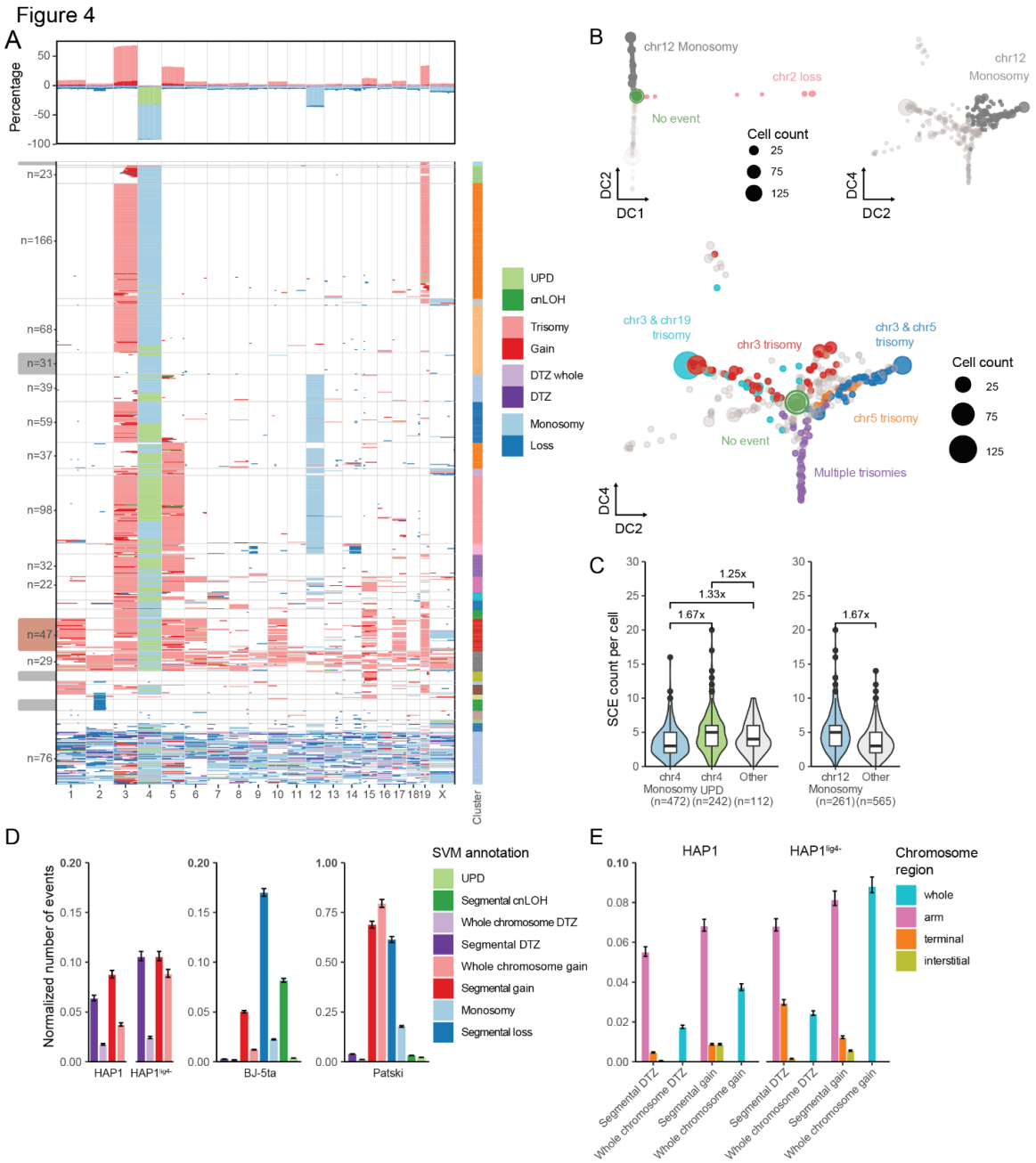

Figure 4. Annotation and quantification of mutational mitotic crossover outcomes.

(A) Heatmap of Patski SV annotation. The heatmap shows only cells with annotated events. Cumulative percentages of cells with an annotation are shown in the top panel. Each row represents a single cell. Cells were ordered and grouped using hierarchical clustering. The chosen 29 clusters (not plotted with the 177 optimal number of clusters here) are shown on the right. Only cell counts (n) of clusters with more than 20 cells are shown. Clusters containing a single annotated event representing founder populations are indicated with a gray bar on the left. The cluster with multiple trisomies is highlighted in dark red. Heatmaps of mutational SVs in BJ-5ta, HAP1 and HAP1^{lig4-} cells are shown in FigS20A-C.

(B) Diffusion maps of mutational SVs in Patski cells. Diffusion components 1 (DC1) and 2 (DC2) are shown in the top panel colored by segmental loss on chr2 (separated out by DC1, x-axis) and monosomy on chromosome 12 (separated out by DC2, y-axis). The bottom panel shows DC2 and DC4 colored by different combinations of trisomy. DC3 separated out two specific clusters, examined in FigS23B. Each point represents a unique set of annotated SVs, while the size of each point represents the number of cells that share the same set of annotated outcomes. High gain and high loss cells were excluded. Multiple trisomy represents cells with trisomy on chromosome 1, 3, 5, 6, 10, 15, 17 and/or 19 (dark red highlight in A).
(C) SCE rates within subsets of Patski SV outcomes. In the right panel, cells were divided into groups based on the presence of chr4 monosomy or chr4 UPD or the absence of both (other category), while the left panel was divided based on the presence or absence of monosomy on chr12. The fold change of the median SCE counts per cell are shown for pairs of annotated events above the relevant plots. Significant differences for the left and right panels were calculated with the Mann-Whitney U test, comparing chr4 Monosomy vs chr4 UPD (p-value: 1.54×10^{-11}); chr4 Monosomy vs Other (p-value: 2×10^{-03}); chr4 UPD vs Other (p-value: 4.6×10^{-02}); chr12 Monosomy vs Other (p-value: 1.94×10^{-13}). Significant differences with additional summary statistics such as median, MAD, and fold change are provided in TabS13.
(D-E) Pre-existing events (see FigS25A) and chromosomes without any event annotation are not plotted. Plotted events were counted per chromosome and normalized by the total number of cells and the ploidy of respective cell line. Sex chromosomes were excluded from BJ-5ta cells. Error bars show the 95% confidence interval.
(D) Normalized counts of mutational SVs. Significant differences were calculated with chi-squared for HAP1^{lig4-} vs HAP1 with segmental DTZ (p-value 5.18×10^{-04}; 1.7x), whole chromosome gain (p-value: 5.32×10^{-12}; 2.5x). Contingency tables of additional event combinations with normalized counts, fold change and chi-squared p-values are provided in TabS14.
(E) Normalized counts of mutational SVs in HAP1 and HAP1^{lig4-} cells split by chromosome location. Significant differences were calculated with chi-squared within HAP1 whole vs terminal DTZ (p-value: 1.20×10^{-06}); HAP1^{lig4-} whole vs terminal DTZ (p-value: 1). Whole categories from (D) are included for reference. Contingency tables of additional event combinations with normalized counts, fold change and chi-squared p-values are provided in TabS14.

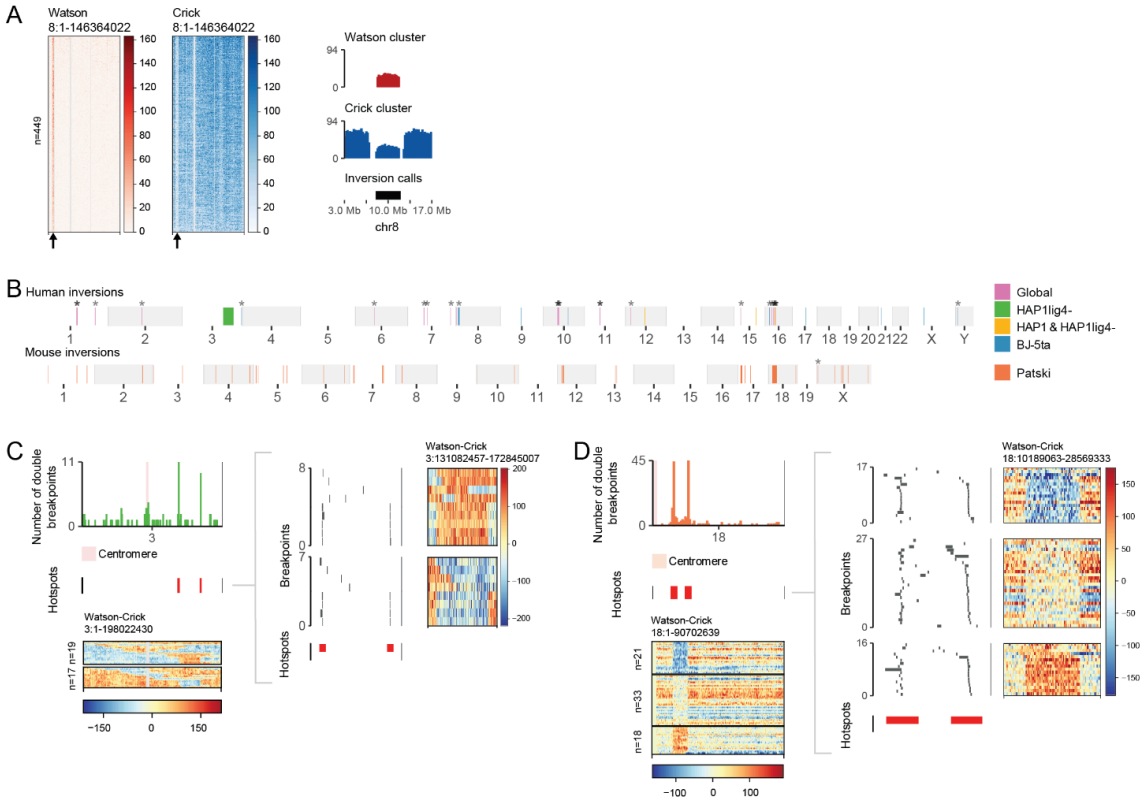

Figure 5. Detecting inversions with sci-L3-Strand-seq.

(A) A universal inversion on chr8 in BJ-5ta. The left panel shows chr8 heatmaps from a subset of 449 single cells with both chromosomes in the C strand orientation. Each row represents a single cell plotted at an approximate resolution of 200Kb. W reads are plotted in red (representing background reads except for the inversion), and C reads are in blue. Masked regions are in gray. The same bin size and color code are used for heatmaps in Fig5 to Fig7 unless noted otherwise. Arrows highlight an inversion at the start of chr8. The right panel shows the pileup tracks of the heatmaps zoomed-in on the inversion, showing the median value for each bin across all plotted cells. The pileup for the W read heatmap of CC chromosomes is shown on the top track, and the C read heatmap of CC chromosomes is shown on the middle track. The bottom track shows our inversion call.

(B) Genome-wide locations of inversion calls. Inversions identified across the three human cell lines are shown in the top panel. Patski inversions are shown on the bottom panel. The size of inversions have been scaled to a minimum of 1Mb to allow visualization. Inversions found in at least two cell lines were classed as global. There are no inversions private to HAP1 cells (all shared with HAP1^{lig4-}). Even chromosomes are shaded gray for visual clarity. Calls overlapping with previously published inversions are highlighted with an asterisk [16,17].

(C) Inversion detected from double strand switch hotspots in HAP1^{lig4-} cells. The sub-clonal inversion measuring 27.98-33.8Mb was observed as a hotspot (red) of coincident strand switches shown in the top left bar plot. Matching whole chromosome heatmaps are shown underneath, where each row represents a single cell with at least two strand switches on chr3. The heatmaps show the inversion as a block of strand state change in a subpopulation of cells. Each roughly 200 kb bin shows the subtraction of W from C read counts. A zoom-in on the breakpoints within the hotspots is shown in the right panel,

revealing 9 cells that contain paired strand switches in this region. The list of subclonal inversions with exact coordinates is provided in TabS2.

(D) Inversion detected from double strand switch hotspots in Patski cells. The subclonal inversion measuring 4.9-14.5Mb was observed as a hotspot (red) of coincident breakpoints shown in the top left bar plot. Whole chromosome heatmaps are shown underneath, same as in (C). The zoom-in on breakpoints within the hotspots is shown in the right panel, revealing 55 cells that contain paired strand switches in this region (TabS2).

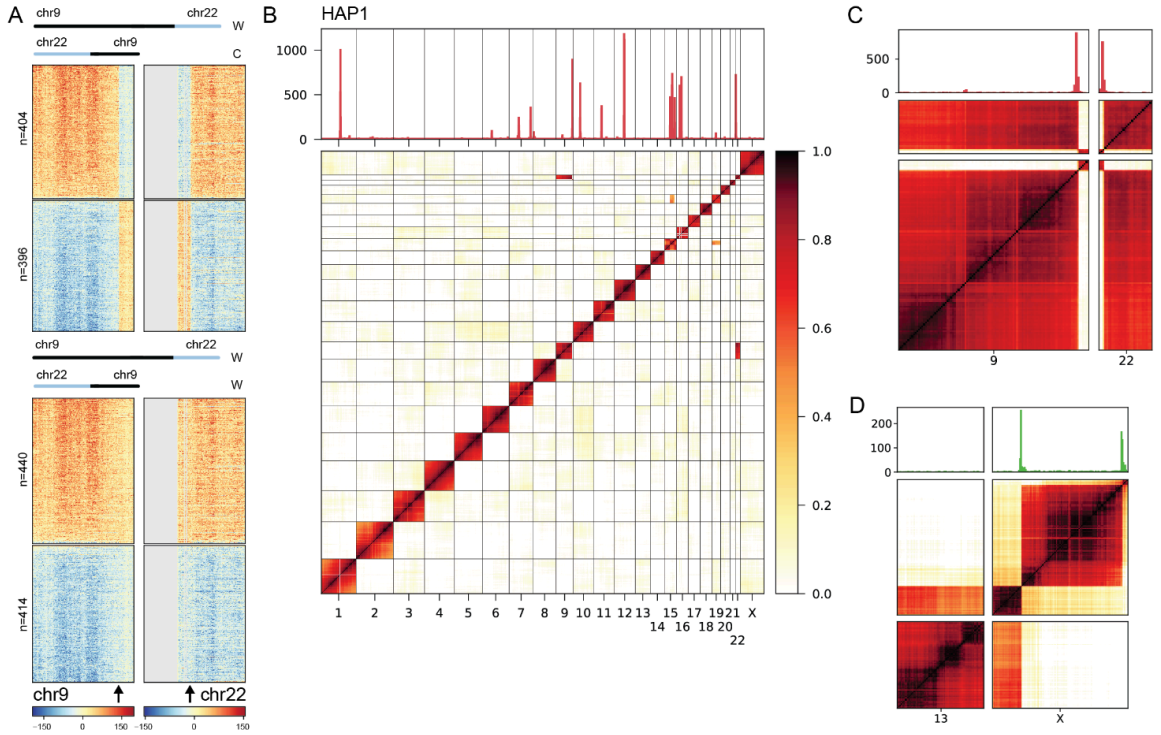

Figure 6. Positive Pearson correlation of strand states reveals translocations.

(A) Strand-state heatmap of a balanced translocation between chr9 and chr22 in the HAP1 cell line. Each row represents a single cell. Each 200kb bin shows the subtraction of W read counts from C. Only the subset of cells without SCE on chr9 are depicted. The translocation appears as a strand switch in ~50% of the cells (black arrows). Cells are ordered based on the strand state at the edges of each chromosome.

(B) Genome-wide heatmap of strand-state Pearson positive correlation in HAP1 cells. Heatmap was plotted at a 1Mb bin resolution using 439 cells. Strand switches without the region filter are plotted in the top panel. Positive correlation *in trans* (off-diagonal) highlights mapped translocations. See also FigS27.

(C) A zoom-in of chr9 and chr22 in HAP1 cells. The strand state of the terminal end of chr9 correlates with the strand state of the terminal start of chr22, highlighting a balanced translocation forming the Philadelphia chromosome. Top panel highlights the breakpoints (without the region filter) that pile-up at the translocation. Scale is the same as in (B).

(D) A zoom-in of chr13 and chrX in HAP1^{lig4-} cells. The terminal start of chrX correlates with the strand state of chr13, highlighting a balanced translocation that was not present in the parental HAP1 cell line. The terminal end of chrX additionally shows a reduced correlation with the rest of the chromosome, highlighting an inverted duplication. Top panel shows strand switches without the region filter. Scale is the same as in (B).

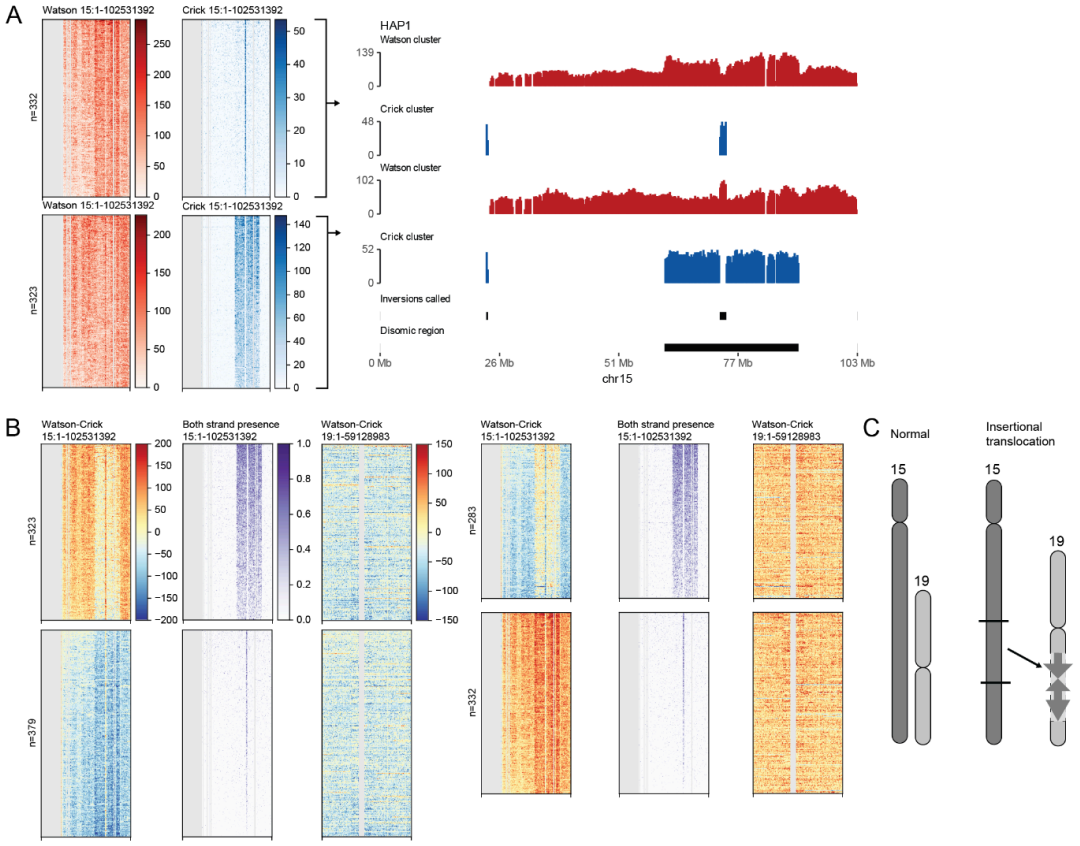

Figure 7. Resolving a complex rearrangement in haploid cells.

(A) A heterozygous inversion within the disomic region on chr15. Each row represents a single cell. Pileup tracks of each heatmap show the median value for each 200kb bin, with an additional track showing inversion calls and the previously published location of the chr15 disomic region (right panel).
(B) The inversion is located within the newly gained copy created by an unbalanced insertional translocation. Strand state heatmaps of chr15 and of the translocation chromosome**.** Each row represents the same single cell across all heatmaps. Cells were divided into two sets, each consisting of three heatmaps, based on the strand state of chr19. Cells with Crick chr19 are on the left, while cells with Watson chr19 are on the right. The subtraction of W from C reads is shown on the first (chr15) and third (chr19) columns. The second column shows regions where both W and C reads are present. Details in FigS28.
(C) Diagram showing the final rearranged structure on chr19. The original structure of chr15 (dark gray) and chr19 (light gray) is shown on the left, while the insertional translocation into chr19 containing an inversion is shown on the right.

Methods

Cell culture

BJ-5ta (CRL-4001, ATCC) and Patski (gift from Disteche lab) cells were cultured in DMEM medium supplemented with 10% FBS and 1x Pen-Strep. HAP1 cells were cultured in IMDM medium supplemented 10% FBS and 1x Pen-Strep. All cells were cultured at 37°C, 5% CO2. For sci-L3-Strand-seq, cells were cultured with BrdU at 40 µM final concentration for 24 hours prior to fixation. HAP1 knockout cells for LIG4- (referred to as HAP1^{lig4-}) are as described in [36] (gift from Shendure lab).

Sci-L3-Strand-seq library generation

Fixation and nucleosome depletion was performed as previously described in 'Methods and molecular design of sci-L3-WGS and sci-L3-target-seq', subsections 'Single Cell Preparation and Nucleosome Depletion' and 'Tagmentation (first-round barcodes) and ligation (second-round barcodes)' [8]. Notably, cells were trypsinized and fixed with 37% formaldehyde (final 1.5% concentration) in 1x PBS at a cell density of 1 million/mL for 10 minutes at room temperature with gentle tube inversion. For first round barcoding,nuclei were distributed into individual wells. Any remaining nuclei were stained with DAPI at a final concentration of 5 ng/µL and used for FACS as previously described [25]. After the ligation reaction was stopped by the addition of the stop solution (lysis buffer (LB: 60 mM Tris-Ac pH 8.3, 2 mM EDTA pH 8.0, 15 mM DTT) with 0.1% Triton-X100 (LBT)), nuclei were pooled and stained with Hoechst-33258 to a final concentration of 5 ng/µL and the quenched population was sorted (100-300 nuclei per well) into a 96-well plate containing 3 µL of LB. Right after gap extension with the Bst WarmStart 2.0 polymerase at 68°C for 5 minutes and its inactivation with 80°C for 10 minutes, Hoechst-33258 was added at a final concentration of 10ng/µL and incubated at room temperature for 10 minutes in the dark. A 1 or 5 minute exposure in a Bio-Rad Gel Doc with a 365nm UV bulb, or a dose of 27-4000 mJ/cm2 was administered with a UVP crosslinker (CL-3000L) as previously described [25]. At this stage, each well had a volume of around 7.7 µL, to which 0.3uL of USER enzyme (a mix of uracil DNA glycosylase and endonuclease VIII, referred to as UDG+EndoVIII in the text, NEB) was added and incubated at 37°C for 15 minutes. Next, 0.9 µL of its inhibitor UGI (NEB) was added with a further incubation at 37°C for 10 minutes. Finally, the T7 in vitro transcription system was assembled by adding 2 µL H2O, 2 µL T7 Pol mix and 10 µL rNMP mix (NEB, HiScribe T7 Quick High Yield RNA Synthesis Kit) and incubated at 37°C for 10-16 hours. The remaining section, 'RNA purification, RT and SSS', was performed exactly as previously described [8].

Read processing and alignment

Sci-L3 read barcodes were identified and processed as previously described [8]. Briefly, the main steps of de-multiplexing consist of:

1. Swapping read 1 (R1) and read 2 (R2) based on RT primer location such that all combinatorial barcodes are in R1. RT primers (GGGATGCAGCTCGCTCCTG) are identified based on Levenshtein distance of less than or equal to 3 from the reference sequence.
2. The 3rd round barcode (SSS, 6nt, no error allowed) and UMI (4nt) from R1 are attached to the read name. Reads are split into individual fastq files based on barcode sequence, enabling all subsequent steps to be performed in parallel. Reads without an identifiable barcode are discarded.

3. The remaining barcodes are extracted by first trimming the Tn5 Mosaic End double-stranded (MEDS) sequence from R1 (5' adaptor: AGATGTGTATAAGAGACAG; maximum error rate: 0.2, minimum overlap: 19) and R2 (5' adaptor: AGATGTGTATAAGAGACAG; 3' adaptor: CTGTCTCTTATACACATCT; maximum error rate: 0.2, minimum overlap: 13) using cutadapt v3.0 [105] and parsing the MEDS adjacent sequences containing the 1st (Tn5, 8nt, 1 error allowed) and 2nd (ligation, 7nt, 1 error allowed) round barcodes (bc2 7nt - 4nt spacer - bc1 8nt - MEDS). Errors are calculated using Levenshtein distance. Successfully identified barcodes are written into the read name, else both R1 and R2 are discarded.

Subsequently, paired-end reads were aligned to a hs37d5 and mm10 hybrid genome reference using BWA v0.7.15-r1140 [106] and converted into BAM format with only primary alignments using Samtools v1.10 [107]. Using the 1st and 2nd barcodes in the read names, the 3rd round barcode BAM files were additionally split into individual single cell BAMs. Barcode combinations with less than 10000 reads were discarded. Single cell hybrid genome BAMs were additionally split by mouse or human contigs, retaining only standard chromosomes. Each BAM was filtered for proper read pairs, a maximum insert size of 2000, and excessively soft-clipped reads based on a maximum soft-clipping ratio of 0.5 (defined as the number of soft-clipped bases over the number of matched bases). We found that excessively soft-clipped reads disproportionately contribute to higher background estimate levels (FigS3A). Finally, to avoid collisions, any split and filtered BAM that had less than 30% of total reads of the correct species or less than 10000 reads was discarded.

A general processing workflow implemented with snakemake [108] is available at https://github.com/recombinationlab/sciL3Pipe.

Complexity estimation

Single cell library complexity was estimated using preseqR v4.0.0 [109,110] by extrapolating the percentage of unique fragments being observed with an increasing sequencing effort. Per cell insert size was calculated as the mean of all cell fragment sizes. Percent coverage was calculated from the extrapolated duplication histogram by multiplying the unique fragment number by the mean insert size and normalizing by the mappable genome size. PreseqR was set to predict the expected number of species represented at least once in a given sample (r=1).

Breakpoint calling and filtering

Strand breakpoints were called using breakpointR v1.10.0 [34]. The window size for calculating ΔWs was set to 1Mb for haploid cells and 5Mb for diploid cells. The minimum number of reads required between two breaks for strand state assignment was set to 10 and 50 for haploid and diploid cells respectively. The smaller window size and lower minimum required reads were chosen for haploid cells due to the ability to filter false breakpoint calls based on strand states (strand state filter). Blacklisted regions for hg19 and mm10 (version 2) [111] were excluded from all analysis. For mm10, we additionally included a high signal region we consistently found on chr4-89540413:89544084.

breakpointR run command for haploid cells:

```
breakpointr(inputfolder = <>, outputfolder = <>, windowsize = 1000000,
binMethod = "size", pairedEndReads = TRUE, pair2frgm = FALSE, min.mapq = 20,
filtAlt = TRUE, peakTh = 0.33, trim = 10, background = 0.05, minReads = 10,
maskRegions= <>)
```

breakpointR run command for diploid cells:

```
breakpointr(inputfolder = <>, outputfolder = <>, windowsize = 5000000,
binMethod = "size", pairedEndReads = TRUE, pair2frgm = FALSE, min.mapq = 20,
filtAlt = TRUE, peakTh = 0.33, trim = 10, background = 0.05, minReads = 50,
maskRegions=<>)
```

Haploid breakpoint filters

To filter incorrect haploid breakpoints we:

1. Breakpoints overlapping with our genomic region tables (see “Breakpoint region filter and SCE filter”) were removed (region filter; regions shown in FigS7). The region filter supersedes all other filters and clustered event rescues.
2. Remove breaks that are within (<=) 2Mb of each other (clustered breakpoint filter). The 2Mb cutoff was chosen based on the bimodal distribution of distances between nearest breakpoints (FigS6C). Breakpoints that have been removed by the region filter are not included when finding clustered events.
3. Only retain breaks with a strand state consistent with haploid cells (e.g. retain WW-CC and CC-WW, but remove WW-WC or WC-CC breaks) (strand state filter).
4. For pairs of breaks overlapping centromere coordinates or within 3Mb of the centromere, if the strand states of the outer segments for each break are identical, both breaks are removed (e.g. break 1 strand state: **WW**-CC, break 2 strand state: CC-**WW**) (centromere filter).
5. Rescue clustered events that have been filtered by examining the strand states of the outer segments of a cluster (e.g. **WW**-CC CC-WW, WW-**CC**), if the strand states differ, a new breakpoint is created by taking the outermost coordinates of the cluster. A rescued break supersedes the clusters event filter, centromere filter and the strand state filter.

Diploid breakpoint filters

To filter diploid breakpoints we followed the same procedure as for haploid cells except the strand state filter was not used and for the clustered events filter and the centromere filter we used a distance of (<=) 10Mb (FigS6C).

Breakpoint region filter and SCE filter

After an initial round of filtering (without a region filter) for both haploid and diploid cells, we identified regions with high density of breakpoints using the hotspotter function from the breakpointR v1.10.0 package [34]. Hotspots for single and multi-breakpoints per chromosome were called separately and combined. For human cell lines, we additionally combined overlapping hotspots into single intervals and summarized the number of breakpoints and number of cell lines for each hotspot. Using these two numbers, we normalized the total number of breakpoints within a hotspot by the total number of cells and weighted each value by the number of cell lines across which a hotspot was found, giving us an “importance score” for each hotspot. An “importance score” above 0.1 was used to categorize hotspots into global and cell line specific regions. Ultimately, we found that regions with high breakpoint counts corresponded to either inversions (see “Inversion identification”), translocations (see “Translocation annotation”), duplications (FigS7E-F), or errors in the genome reference as previously described [16,17]. Manual inspection of hotspots without a structural variation (SV) or a reference error revealed these hotspots had no obvious direct strand switches despite breakpoints being called. We found these hotspots were adjacent to inversions, suggesting that inversions may be causing breakpointR to call false

positive breakpoints, possibly due to a missed call directly at the site of the inversion. We therefore categorized these as “inversion proximal” in our region filter.

In addition to inversion proximal hotspots, we also found the human centromeric regions of chr1 and chr16 contained hotspots with a high frequency of false positive breakpoint calls. However, from chromosome-wide heatmaps of all cells we found that by applying a region filter, we were also removing a substantial number of true positive breakpoints. By clustering cells by the strand state on the end of each chromosome we could pick out cells with centromere breakpoints that were filtered out. To salvage these true positive breakpoints we implemented an extra centromere rescue procedure. For each cell with a breakpoint call within the centromeric regions of chr1 and chr16 (before region filtering), we examined the strand state of both chromosomal arms independently. If the composite strand state of the two arms indicated a strand switch (haploid: arm1 - WW, arm2 - CC | arm1 - CC, arm2 - WW; diploid: WW,CC | CC,WW | WW,WC | WC,WW | WC,WW | CC,WC | WC,CC) the centromere breakpoints were rescued.

The two sets of breakpoint calls (with and without the region filter) ultimately enabled us to classify breakpoints into categories corresponding to their origin (e.g. inversion, translocations, SCE, etc.). After each round of filtering, the strand state of each breakpoint and segment was re-evaluated with `background=0.1` and `minReads=10` settings.

Segmental structural variation can additionally result in a strand switch and be incorrectly classified as an SCE. To account for breakpoints resulting from SVs, we used the SV annotation from our SVM classifier (see “Annotation with Support Vector Machines classifiers) to define adjacent breakpoints.

Haplotype-aware breakpoint calling

For haplotype-aware breakpoint calling, we first split aligned reads into separate BAM files based on the presence of haplotype 1 or haplotype 2 phased SNVs. Reads containing a mixture of both haplotype SNVs were excluded. For each cell, both haplotype BAMs were used as inputs into breakpointR with haploid settings described above. BreakpointR was not designed to detect drops to zero (see “Drop to zero annotation (DTZ)”) that are common features in cells split by haplotype. To overcome the limitation in calling breakpoints for DTZ regions, we devised two orthogonal approaches.

First, we reasoned that by artificially introducing background noise to each haplotype specific BAM, breakpoints could be called since the DTZ regions would contain sufficient reads to have a WC state assigned. To create a pool from which reads could be randomly sampled, we merged the BAMs of 50 cells with a background estimate of 0, i.e. every chromosome had both Watson and Crick reads. Based on the read depth of each cell we created dynamic bins (size dependent on coverage) across which reads would be randomly sampled. Random reads corresponding to 10% of each cell's total reads were introduced as background noise. We found that with each random initialization, we recovered a disjoint set of breakpoint calls, leading us to perform five random initialization and taking the union of all breakpoints identified.

Second, we took advantage of the HAP1 SVM for identifying segments with DTZ (see “Drop to zero annotation (DTZ)”). For Patski cells we binned the entire genome into 4 Mb bins and calculated RCC and RCBC scores (see “Variables for Support vector machines (SVM) classifiers”) for each bin, which were passed to the SVM to enable identification of DTZ bins. The sparsity of SNVs in the human genome led us to use 10 Mb bins for BJ-5ta cells. To consolidate the bins identified as DTZ, we first removed any calls for bins with a mappable width of zero. Single bin gaps between consecutive DTZ bins were relabeled to

DTZ and any single lone calls were removed. A final merging across two bin gaps was performed and the edges of each DTZ segment was used as a new breakpoint.

The final set of haplotype-aware breakpoints is a combination of the original haplotype-unaware breakpoints, breakpoints from haplotype split reads with five different background noise random initializations, and the edge bins of DTZ calls. For segmentation identified by more than one approach, we prioritized original haplotype-unaware calls, followed by calls from haplotype split reads, and finally DTZ calls.

Sample selection and cell filters

To select high quality cells, we performed filtering using the background estimate, strand state percentage and coverage. The background estimate for each cell was calculated using breakpointR v1.10.0 [34]. Cells with a background estimate above (>) 0 and below (<=) 0.08 (FigS3B), and with more than (>) 2 median reads per megabase were retained. We grouped regions with a WC strand state and/or not biased enough to be assigned to WW or CC states together and defined these as having a “strand-neutral” state. Examining the strand-neutral percentage across diploid cells showed the expected centered distribution at 50%. We additionally observed a peak above 75% resulting from unsuccessful libraries (FigS4A). We therefore excluded any cells with a strand-neutral percentage above (=>) 75% from downstream analysis. Haploid cells displayed a tri-modal distribution of strand-neutral percentages, with the primary distribution representing good quality haploid libraries starting just below 25% (FigS4A). Since haploid cells should not have any regions with both W and C reads, we chose a stringent threshold that excluded any cell with a strand-neutral percentage of more than (=>) 15, leaving some margin to account for a known disomic region [42].

Cell barcode collisions in human and mouse cell mixing experiments were identified from the preferential alignment of read pairs to the species genome of origin. However, identifying collisions from alignments is not possible with experiments that mixed two cell lines of the same species. In such cases, we additionally utilized the presence of the Y chromosome to distinguish male and female cell lines, and examined SNVs private to each cell line to exclude misassigned cells (FigS3C).

To identify private SNVs, we first aggregated all cell line specific reads into a single BAM file (pseudo-bulk) and used BCFtools [107] to call SNVs. For BJ cells, an available bulk WGS dataset (SRA: SRP102259; [24]) was used instead of our pseudo-bulk. Records private to each cell line were output using the `isec` command. Multi-allelic SNV sites were split into biallelic records. For haploid cells, only homozygous SNVs were retained. For each single cell, a ratio of the number of reads containing a private SNV over the total number of reads covering private SNV positions was calculated. The ratios were scaled to fall between 0 and 1 and any cell above 0.5, meaning it contained more SNVs private to another cell line than its assigned line, was excluded from downstream analysis (FigS3C).

BCFtools SNV call run command:

```
bcftools mpileup -O z --skip-indels --ignore-RG --redo-BAQ --min-BQ 13
--per-sample-mF -a 'AD,ADF,ADR,DP,SP,SCR' -f <genome.fa> <combined.bam> |
bcftools call --multiallelic-caller --variants-only -O z -o <out.vcf.gz>
```

We also manually examined cells with more than (>) 10 breakpoints and excluded any cells from downstream analysis that contained a high number of false positive breakpoints, leading to the removal of 2 BJ-5ta and 3 Patski cells.

Chromosome heatmaps

To visualize the strand status across thousands of single cells we used an extension of the bioconductor package profileplyr v1.8.1 [112] and Deeptools v3.5.0 [113], implemented in sciStrandR, to plot heatmaps. Each row of the heatmap represents a single cell and each column a genomic interval. All heatmaps were plotted at a 200 Kb maximum resolution with counts per million (cpm) normalization, defined as: $cpm = \frac{strand\ fragment\ count}{(total\ fragment\ count/1000000)}$. Bins overlapping blacklisted regions [111] and centromeres were masked. Count matrices were created for both Watson (W; positive) and Crick (C;negative) strands from R1 reads extended by 800bp. Additional matrices were produced from operations on the W and C count matrices. Heatmaps capturing the difference between strands counts were produced by subtracting the W from the C count matrix and setting empty bins to 0. Being able to plot the combined heatmaps with only a single scale makes it difficult to distinguish between WC states and low coverage regions as both will have values close to 0. To distinguish WC state regions, an additional matrix was constructed that highlighted the presence of both W and C reads. The matrix was constructed by dividing the parallel minimum by the parallel maximum of the W and C read counts for each bin.

A final matrix containing the number of breakpoints within each bin was created to help sort and categorize the heatmap matices: (1) Cells containing a single breakpoint within the plotting interval were split into the first two categories based on the strand orientation of the first 3% of bins. For cells with low coverage in the starting bins (sum of CPM less than 5), the first 10% of bins were examined instead. (2) Any cells with more than a single breakpoint were split into the next two categories. All cells with a breakpoint were additionally sorted by breakpoint position within each category.

SCE distribution and correlation with genomic features

SCE enrichment and depletion within centromeres

The enrichment or depletion of SCEs from the centromeres of non-acrocentric chromosomes was evaluated using a permutation test from the regioneR v1.30.0 Bioconductor package [104]. Before testing, the centromeres were resized to the widths within the CHM13 v2 reference [41,114]. Additionally, the centromere of the human chr1 was extended upstream to encompass a proximal inversion that masked centromeric breakpoints. For each non-acrocentric chromosome, the number of SCE overlaps with centromeres was counted after 1000 random permutations of SCE coordinates. An SCE was counted as an overlap only if it overlapped by more than (>) 40% of its size. Human acrocentric and mouse chromosomes were not evaluated due to poor mappability of the short p arms.

SCE hotspots

SCEs were counted across 1Mb bins and normalized by the total number of bins each SCE overlapped. Each bins counts were further normalized by the total number of cells and the ploidy of the cell line (HAP1 and HAP1^{lig4-}: 1; BJ-5ta and Patski: 2). The normalized tracks were fitted with several common univariate probability distributions and based on Q-Q plots the gamma distribution was selected. P-values were calculated using the cumulative density function and further adjusted using the Benjamini-Hochberg approach as implemented in the core R stats package.

SCE correlation with genomic and epigenomic features

To examine the correlation of SCEs with various genomic and epigenomic features we downloaded processed datasets from previous studies, along with gene and various repeat class annotations. Repeat annotations were downloaded from the UCSC Genome Browser RepeatMasker (https://genome.ucsc.edu/cgi-bin/hgTables). The resulting collection of features were processed into either count data (gene and repeat annotations), signal tracks (genetic features, e.g. GC percentage) or normalized signal tracks (epigenetic marks). Datasets obtained varied in their window sizes and were further processes to obtain an averaged signal across 1Mb sized windows. SCEs were reduced to only the midpoint coordinate when counting overlap with 1Mb windows. Spearman correlations were computed on the resulting matrix and p-values were further adjusted using the Bonferroni method.

For hg19, the following datasets were used: (1) HAP1 AB compartments, lifted over from hg38 to hg19 [115]; (2) HAP1 and BJ Repli-seq replication domains [116]; (3) HCT116 Gloe-seq [51]; (4) NALM6 END-seq after etoposide treatment and no treatment control (NT) [50]; (5) Chromosome fragile sites, lifted over from hg38 to hg19 [53]; (6) HepG2 ChIP-seq for RAD21 and SMC3 [117].

For mm10, the feature matrix previously generated in [8] was subset and resized from 100Kb to 1Mb windows. The following datasets were used in the final matrix: (1) sequence divergence and (2) CNV gains between the strains [30]; (3) END-seq after etoposide treatment and no treatment control and (4) CTCF and RAD21 ChIP-seq in C57BL/6 and Spretus activated B cells, along with (5) TOP2A and TOP2B ChIP-seq in MEFs [50]; (6) Patski allelic ATAC-seq data and PolIIS5p ChIP-seq [118]; (7) mESC Repli-seq replication domains [119].

Phasing

Variant Calling

For Patski cells, a pseudo-bulk VCF was created from single cell BAM files using 1569 cells from the following libraries: nextseq190419 (yi293_GAACCG n=2, yi293_AGGACG n=55, total=57 cells), nextseq190830 (n=979), nextseq190906 (yi311_AATTGA n=38, yi311_ACTTAA n=8, yi311_CGGTGG n=5, yi311_GAGTTT n=4, yi311_GGGCCG n=29, yi311_GGGTGC n=24, yi311_GTCTAT n=12, yi311_TAAAGA n=6, yi311_TAAGCG n=29, yi311_TAATGC n=5, yi311_TATTCT n=8, yi311_TTTCGC n=21, total = 189 cells), and nextseq190907 (n=344). BAM files were merged, sorted, and indexed with Samtools [107]. Human and mouse reads were split using a custom Python script. Variants were called using BCFtools `mpileup` and `call` commands against mm10 [107]. The merged and filtered Patski VCF file was intersected with the published SPRET/EiJ SNVs [56] using the bcftools `isec` command and then filtered for biallelic SNVs to obtain the final set of SNVs for phasing. There were 33,105,073 total SNVs of which 28,673,432 had the heterozygous (0/1) genotype informative for phasing.

For BJ-5ta cells, SNVs were called with BCFtools `mpileup` and `call` against hg19 using previously published bulk WGS data [24], resulting in 4,280,029 total SNVs, with 4,073,787 SNVs for chromosomes 1-22 (sex chromosomes were excluded from phasing) of which 2,569,252 had the heterozygous (0/1) genotype informative for phasing. The 2,847 multiallelic SNVs were not used for phasing.

Phasing with StrandPhaseR

Phasing was performed using StrandPhaseR (v1.0.0; R 4.1.0) [57]. Phasing was performed using 949 Patski and 2180 BJ-5ta cells. WC regions were extracted using a modified version of the extract function in breakpointR [34], included in sciStrandR, and partitioned by chromosome for phasing (chr1-19 and X for Patski and chr1-22 for BJ-5ta). All cells that pass QC and are present in the WC regions file of a particular chromosome were used for phasing, except in two cases: 1) the phasing crashed prior to completion, in which case fewer cells were used (see “Errors Encountered During Phasing”), 2) the chromosome contained many LOH events, in which case a subset of heterozygous cells were used for phasing (see “Additional manual phasing for Patski”).

Inferred Phasing

StrandPhaseR generates phased haplotypes where substantial fractions of loci between haplotypes 1 and 2 are disjoint (818K of 1,931K phased Patski sites and 107K of 174K phased BJ sites had both alleles assigned to haplotypes). SNV loci that were assigned for only one of the haplotypes were inferred for the other using the reference. In some cases, the same reference allele was assigned to both haplotypes, despite us only using heterozygous SNVs. In these instances, the ambiguously phased SNVs were filtered out as “conflicts”.

Correction of phasing over inversions

Previously published inversions were lifted over from mm9 to mm10 using the UCSC Liftover Tool [16,120]. The large inversions on chr14 and on chrX were also identified by our inversion calling. We therefore chose to use only these two inversions for correction, whereby alleles within any inverted regions are exchanged between the haplotypes 1 and 2.

Additional manual phasing for Patski

We found that Patski chr4 phasing accuracy was significantly lower than for other chromosomes. After an early pass of LOH calling where 761 out of 934 Patski cells were found to have Haplotype 1 LOH on chromosome 4, it became clear that most of the Patski cells were uninformative for phasing chr4. Initially, a subset of 10 heterozygous Patski cells were used for phasing of just 22K and 63K sites for haplotypes 1 and 2, respectively, but better SNV coverage was desired. As the majority of cells contain LOH, they are effectively pre-phased without the need for Watson-Crick strand information. Cells with the whole-chromosome haplotype 1 LOH fraction >= 0.9 were aggregated and sites deemed to be homozygous (allelic depth of the most frequent allele / allelic depth of all observed alleles >=0.75) were manually phased as the first haplotype. We required that at least 2 cells have a read at each SNV site and assigned the most frequent allele across all LOH cells as the Haplotype 1 allele. Because the Patski VCF was filtered for biallelic SNVs, the second unobserved allele at each position could be assigned to Haplotype 2. Unlike StrandPhaseR that requires Watson-Crick information to phase heterozygous sites, all SNVs in chr4 LOH cells with a coverage of 2 or more were phased. Similarly, we collected 281 Patski cells with hap1 LOH on chr12 (only 13 had hap2 LOH and were excluded) and attempted the same manual phasing. However, the accuracy failed to improve over StrandPhaseR and was therefore not used.

Phasing Accuracy Calculation in Patski

Phasing accuracy is calculated for both haplotypes using the Patski ground truth alleles with maternal SNVs from the REF/B6 and paternal SNVs from the ALT/SPRET. Because haplotype 1 and 2

assignments are made arbitrarily for each chromosome (in other words, each chromosome is phased into 1 of the 2 haplotypes and not the entire genome), we do not know ahead of time which haplotype corresponds to the REF or ALT ground truth. Accuracy was calculated by binarizing bases matching the REF allele from the Patski VCF to 0 and those matching the ALT allele as 1, then adding these values to their respective haplotype vectors. This results in one haplotype having a very large sum (almost exclusively 1s) and the other haplotype having a very low sum (almost exclusively 0s). Therefore, the haplotype with a large sum is determined to correspond with the ALT ground truth while the haplotype with a small sum corresponds to the REF ground truth. As these assignments are arbitrary and vary by chromosome, the max of each haplotype vector sum is taken (i.e., suppose 9 ALT alleles and 100 phased sites, then taking the maximum of 9 or 100-9=91 gives 91 and 50 ALT in 100 phased sites would reflect a random guess between the 2 haplotypes). Finally, the max is divided by the relevant number of heterozygous SNVs, het SNVs with WC coverage, or phased sites (using same toy example, 0.09 accuracy with respect to ALT or 0.91 accuracy with respect to REF alleles is reported as 0.91 phasing accuracy for that particular haplotype and 0.5 accuracy would represent the worst possible phasing performance). Invalid bases are counted as inaccuracies in the cases of raw and inferred phasing accuracy calculations; they are accounted for in the accuracy denominator after the conflict removal step. Plots were created using 1Mbp binned accuracies, while the phasing accuracies in supplemental tables and plotting labels are unbinned.

Structural variation annotation

Inversion annotation

Using the heatmap count matrix categorization (see “Chromosome heatmaps”), we picked cells with no breakpoints and with whole chromosome WW or CC states. This chosen subset of each count matrix was collapsed into a bedgraph format by counting the percentage of cells at a particular position with a value greater than (>) 0 (percentage column operation). Subsequently, peaks above 0.4 (reads present in at least 40% of the cells with whole chromosome WW or CC states, roughly 20% in the population) were called as inversions with the exception of Patski, where due to a low number of cells with whole chromosome WW or CC states we increased the peak threshold to 0.5 (reads need to be present in at least 50% of the cells). Using the heatmap count matrices limits the inversion resolution to the 200 kb bins of the matrix, but it allows us to detect larger inversions (e.g. human chr8 in BJ-5ta cells), inverted duplications (e.g. chrX in BJ-5ta and HAP1^{lig4-} cells), and problematic regions (e.g. mouse chr14 in Patski cells). Inversion calling was run for WW and CC chromosomes with and without the region filter to ensure the filtering did not influence the calls.

Breakpoints resulting from inversions should appear as double breakpoints. Because of our clustered breakpoint filter, the double breakpoints from inversions less than 2 Mb in haploid and 10 Mb in diploid cells are removed, allowing these chromosomes to be used for inversion calling (only chromosomes with no breakpoints are used). However, larger inversions retaining their double breakpoints would not be used. Additionally, because of the chosen cutoff of 40% in human cells and 50% in mouse cells, double breakpoints originating from sub-clonal inversions would also not be identified. To overcome these two possible instances of missed inversion calls, we additionally examined breakpoint hotspots for chromosomes with more than one breakpoint. We identified a low frequency inversion in HAP1^{lig4-} cells on chr3 that we included in our final inversion call set. We also found a problematic region on chr12 in BJ-5ta cells that caused false double breakpoint calls. The same analysis also revealed double breakpoints on chr4 in BJ-5ta cells that matched a region of segmental loss that had a missed annotation in several cells. For downstream analysis we filtered both the chr12 and chr4 double breakpoints.

Translocation annotation

Translocated chromosomes, in the absence of a sister chromatid exchange, share the same template strand states. We used the correlation of strand states between chromosomes across N number of cells to identify potential translocations. A measure of template strand states was calculated for 1 Mb bins, using the W/C ratio as defined in breakpointR v1.10.0 [34]:

$$W/C\ ratio\ =\ \frac{W-C}{W+C}$$

where W is the number of Watson fragments, and C is the number of Crick fragments within each bin. Fragments in blacklisted regions for hg19 and mm10 (version 2) [111] were excluded. A ratio was calculated only if a bin contained more than (>=) 10 fragments. Bins for which a ratio was not calculated in more than half of all cells were masked. To maximize the bins with sufficient read depth, we limited the analysis to cells with between (>) 20 and (<) 80 median reads per Mb. Any missing values were set to 0 and the resulting W/C ratio matrix was used to calculate a Pearson correlation of the strand states between all bins. For translocation, only the positive Pearson correlation was used, as the negative correlation highlights inversions. It is also possible to use cells with less than (<) 20 median reads per Mb to identify translocations, but at a reduced resolution (5Mb bins size tested).

LOH calling

Pileups were generated from single-cell bams with bcftools `mpileup` using the bulk SNVs described in the “Variant Calling” section. From the bulk VCF, homozygous sites were defined as 0.1 < REF allelic depth / (REF allelic depth + ALT allelic depth) < 0.9 while heterozygous sites were defined by a REF allelic depth >= 0.9 or <= 0.1. After phasing, sites with a total allelic depth > 0 are binarized as (1,0) or (0,1) for homozygous sites and (1,1) for heterozygous sites.

LOH was scanned across chromosomes 1-19, X for 931 Patski cells (female) and 1-22 for 2178 BJ-5ta cells (male). In humans, chromosomes 13, 14, 15, 21 and 22 are acrocentric and only contain phased haplotypes on the q-arm. Strand state region files were generated - retaining regions of all sizes except for the mitochondria - using a modified version of the extract function in breakpointR [34], included in sciStrandR. The strand state regions, together with chromosome arms, were used for chromosome segmentation of each single cell. For chromosome arm segmentation, any segments spanning centromeres were split into 2 segments with the left segment terminating at locus centromere start and the right segment beginning at locus centromere end+1. Acrocentric chromosomes (all Patski and 5 in BJ-5ta) begin at centromere end+1.

For each single cell, SNVs from the joint phasing file were intersected with each cell’s pileup file, then mapped to each of the chromosome segments for calculating the haplotype fractions. Segment fractions are calculated as the sum of binarized tuples containing haplotype 1 (1,0) or (1,1) divided by the sum of binarized tuples containing either haplotype 1 or 2 (1,0) (0,1) (1,1) within the defined segment; hence, each segment fraction is defined as the “haplotype 1 fraction”. Any segments that contain >= 20 phased SNVs and that span >= 10 Mb and exhibit a haplotype 1 fraction >= 0.9 or <= 0.1 are called as LOH candidates. Additionally, haplotype 1 fractions are calculated for each of the chromosome arms (a single arm per chromosome for Patski, two arms per chromosome for BJ-5ta). The LOH calling priority is: 1) chromosome segments that met the LOH candidate criteria above, 2) if a chromosome also meets the arm criteria, use the segmental LOH calls, 3) if a chromosome does not have segmental calls, but does have full arm LOH call, then use the arm call. The LOH calling itself is done in two stages, 1) the

segments from strand state regions are examined via a “segment scan”, and 2) chromosomal arms are examined via an “arm scan”.

There were 3 cases where non-adjacent LOH segments were manually merged. In the first case, an LOH region was too short to be called but meets the fraction threshold (14 cells). In case 2, a short interstitial heterozygous region was between two LOH regions of the same type (both Hap1 or both Hap2). Here, both LOH regions are merged to include the heterozygous region, after which the haplotype 1 fraction is re-calculated (2 cells). Case 3 is similar to case 2 except that the interstitial heterozygous regions contain an annotated inversion and may be longer than 10Mb. Adjacent chromosome segments with extreme haplotype 1 fractions of the same type (Hap1 <= 0.1 or Hap2 >= 0.9) are merged into one segment in the LOH event table (63 cells). In all cases, the merge column of the LOH table is marked ‘M’ versus ‘Y’.

Haplotype 1 fractions and evidence thresholds

We tested three different haplotype 1 faction thresholds: 0.1 & 0.9 [0.9], 0.05 & 0.95 [0.95] and 0.85 & 0.15 [0.85]. With Patski cells, there were 6 fewer cells with LOH calls at the 0.95 haplotype threshold vs 0.9. These calls mostly appear to be artifacts resulting from noisy segmentation and aneuploidy (Fig3D). Additionally, there were 7 fewer Patski cells with LOH calls when the haplotype 1 fraction threshold was reduced again from 0.9 to 0.85. This time, however, many strong LOH candidates were not picked up, even in cells that contained many LOH events across chromosomes. For BJ-5ta cells, chromosomes 2, 10, and 11 exhibited the greatest range in raw LOH call counts, depending on the thresholds chosen. Some example cells demonstrate that many LOH calls are missed at the 0.95 threshold, while questionable calls are introduced at the 0.85 threshold. We also implemented an “evidence” threshold that required a minimum number of phased SNVs to be observed within a segment for it to be considered for LOH calling. We found many short “rescue” regions (see “LOH rescue”) in BJ-5ta cells that were called using evidence of 20 and 30, but were not seen when the evidence threshold was increased to 40 and 50. We began to see true positive LOH missed at the 50 evidence threshold. Based on these tests, we chose to use the 0.1 & 0.9 [0.9] thresholds, requiring all LOH segments to span at least 10 Mb and contain at least 20 phased SNVs within the candidate LOH region for calling, and at least 40 phased SNVs for BJ rescues.

LOH rescue

First, overlapping candidate LOH regions from the segment and arm scans are combined. These represent all the cells with a candidate LOH in a particular region. Subsequently, chromosomes without existing LOH calls are scanned, generating candidates of possible rescue segments. Rescues do not span a full arm, otherwise they would be detected by the arm scan. Each cell and chromosome pair can have numerous LOH rescue candidates, so the best rescue segments are first chosen by picking the longest non-overlapping segments for each cell. If LOH segments of the same class (both having fraction >= 0.9 or both <= 0.1) are overlapping, then the overlapping regions are automatically merged and the number of phased SNVs in the rescue segment - evidence - is recalculated, along with haplotype 1 fraction and segment size. As an example, if cell A has an LOH rescue candidates at chr1:10,000,000-50,000,000 (shared with cell B) and at chr1:40,000,000-60,000,000 (shared with cell C), then the LOH rescued for cell A is chr1:10,000,000-60,000,000. In rare cases where opposite haplotype 1 fractions overlap - less than 0.1 across chr2:10,000,000-chr2:40,000,000 (Hap1) and more than 0.9 across chr2:35,000,000-chr2:90,000,000 (Hap2) - the longer haplotype is fully retained with the shorter one being truncated so it is non-overlapping - less than 0.1 chr2:10,000,000-chr2:34,999,999 and more than 0.9 chr2:35,000,000-chr2:90,000,000.

Data preparation for annotation

Deduplication

Fragments - read pairs merged into a single interval - were deduplicated using only read 1 (R1) coordinates after MEDS sequence trimming (FigS11). First, each R1 start position was extended upstream by up to 3 bp (offset), creating placeholder fragments that together with the original fragments were used to examine any start position overlaps. Fragments - original or their placeholder - with an identical start position to an initial fragment - the first fragment in the stack of overlapping fragments - were marked as duplicates. Finally, the R1 end positions after unmasking soft-clipped bases were examined and any fragments with identical end positions to the initial fragment were marked as duplicates. The union of both start and end position duplicates were considered IVT or PCR duplicates and removed from downstream analysis (FigS11).

We based the deduplication on only R1 due to the secondary enzymatic activity of T7 RNA polymerase called RNA-dependent RNA polymerase activity (RdRP), where T7 RNAP is capable of initiating transcription from RNA primed RNA templates resulting in RNA-RNA duplexes [121–123]. In the event of self-priming or cross-priming of two transcripts, the 3' end of the resulting transcript duplex would be truncated by RNases treatment resulting in a truncated R2 despite originating from an identical gDNA fragment [8,24]. We additionally observed soft-clipping of R1 start and end positions after trimming of MEDS. When soft-clipped bases were unmasked, we found these fragments had identical end coordinates. Consequently, to circumvent any potential issues with MEDS trimming, we incorporated the secondary deduplication step based on the end of R1 after unmasking soft-clipped bases.

Segmentation

Chromosomes of diploid cells were segmented for annotation in the following three ways:

A. Strand state region files containing all strand orientations (CC, WC, WW, Undefined) were parsed for each chromosome and single cell [34]. The region files contain segments as defined by the filtered - without the region filter (see "Breakpoint calling and filtering" - breakpoint boundaries. Regions that completely overlap centromere annotations are split into two segments with the upstream segment ending at centromere start -1 and the downstream segment beginning at centromere end +1.
B. Each chromosome without breakpoints is additionally segmented by its constituent arms based on centromere coordinate. Acrocentric mouse chromosomes are segmented from the downstream centromere end+1 to chromosome end. In BJ-5ta, p-arms are segmented from 1 to centromere start-1 and q-arms from centromere end+1 to chromosome end.
C. Rescue segments (see "LOH rescue") are defined by breakpoint boundaries from other cells that contain candidate LOH calls. Rescue segments are tested in every cell except for those already containing an LOH candidate on the same chromosome. Because a cell-chromosome pair can contain many rescue regions, the longest non-overlapping regions that exhibit haplotype fractions passing a threshold become the rescue segments. In the case that rescue segments overlap, they are merged. If merged segments overlap a centromere then they are split into two rescue segments with the upstream segment ending at centromere start-1 and the downstream segment beginning at centromere end+1.

Segments shorter than (<) 10 Mb or with less than 100 unique fragments were excluded from gain annotation. For drop to zero and loss annotation, only segments shorter than (<) 10 Mb were filtered.

For haploid cells, segmentation was performed as described for diploid cells in A, with the exception of still using the breakpoint region filter. Instead of keeping breaks within our marked regions as we did for diploid cells, we instead chose to manually add segmentation for the disomic region on chr15 (60471098-60471099 and 90518644-90518645, translocations on chr9 (132330486-134723723) and chr22 (22806017-24576823) and additionally on chrX (31719597-33113806) for HAP1^{lig4-} cells. The duplication on chrX in HAP1^{lig4-} cells was not added as it would have resulted in a 7Mb segment that would have been filtered out (10Mb segment filter). Manually adding breaks for known events, rather than retaining called breakpoints over these regions, yielded better annotation of the chr15 disomic region and the translocations in haploid cells without a noticeable impact on annotations across the rest of the genome (see “Post-processing and analysis of annotated events”).

Identification of high spike cells

We observed a subset of cells containing sharp peaks and extensive bin-to-bin variation in read counts. To avoid introducing noise into our SVM annotation classifier, we developed a “spiky score” to remove such cells from annotation. Fragment counts independent of strand were first counted across 1 Mb bins for each cell. A running median across 11 Mb was used to smooth the fragment counts and enabled sharp changes in fragment counts to stand out (FigS5A). The number of peaks (bins) with a smoothed absolute value above (>) 5 were counted for each cell. This threshold was chosen based on the general distribution of peaks across all cells. Because the number of spikes within each cell was dependent on sequencing depth, we additionally applied a loess normalization to the peak counts (FigS5A). Any cells with a normalized peak count - spike score - of more than (>=) 5 were labeled as “spiky” cells.

Variables for Support vector machines (SVM) classifiers

Digital overlap counts (DOC)

The Tn5 transposase leaves a 9bp overhang after cutting. To distinguish adjacent from overlapping fragments, the pairwise overlap of deduplicated fragments was calculated and any pair with more than (>) a 12bp overlap was counted as an overlapping fragment region. In addition, the strand orientation of each overlapping fragment pair was categorized into Watson (DWOC), Crick (DCOC) - same strand overlaps - or opposite strand overlaps (DOOC). Finally, the digital overlap count for a particular genomic segment was normalized by mappability and sequencing depth (FigS17C):

$$DOC_{normalized} = log_2((\frac{Overlap\ count}{segment\ mappable\ size}) \div Median\ reads\ per\ Mb) \times 10^7 + 1$$

Relative chromosome coverage (RCC) score

The RCC score for a genomic segment is calculated as follows:

$$RCC = (\frac{fragments\ in\ segment}{total\ fragments\ in\ cell}) \div (\frac{segment\ mappable\ size}{mappable\ genome\ size})$$

Genome mappability is defined as the proportion of a segment or genome that does not contain ambiguous bases (N) over the total number of bases. Due to differences in coverage between chromosomes, we additionally centered each segment’s RCC score by the median RCC score across all cells, such that a score of 2 could be generally interpreted as aneuploid (FigS17B):

$$RCC_{centered} = \frac{RCC_i}{median(RCC)}$$

Background estimate and relative chromosome background coverage (RCBC) score

RCBC allows us to distinguish true signal from background noise and define regions with no signal as DTZ. Background estimates were calculated using breakpointR v1.10.0 [34]. The background estimate is calculated from segments with a combination of WW or CC template strand states for which the opposite strand reads represent the background. The background estimate for a cell is the mean of the background from WW segments ($\frac{\Sigma C+1}{\Sigma W+1}$) and CC segments ($\frac{\Sigma W+1}{\Sigma C+1}$), where C and W represent the total count of fragments with the respective strand state. From cells with high coverage that contained segmental and chromosomal deletions, we noticed that the background is strand independent and displays an even distribution across the genome (FigS17A). We used these observations to derive the RCBC score that captures the relative contribution of background to the true template strand signal:

$$WW_{background} = \left(\frac{Crick\ fragment\ count}{total\ fragments\ in\ cell}\right) \div \left(\frac{WW\ segment\ mappable\ size}{mappable\ genome\ size}\right)$$

$$CC_{background} = \left(\frac{Watson\ fragment\ count}{total\ fragments\ in\ cell}\right) \div \left(\frac{CC\ segment\ mappable\ size}{mappable\ genome\ size}\right)$$

$$RCBC = mean(ww_{background} + CC_{background})$$

To account for differences in coverage between chromosomes, we additionally centered each RCBC segment score by the median RCC segment score:

$$RCBC_{centered} = \frac{RCBC_i}{median(RCC)}$$

Final strand states

In haploid cells a key feature for segmental gains annotation is the WC strand state (FigS16). In instances where a breakpoint was missed, a false WC strand state can be assigned that negatively impacts annotation. A similar case can arise in diploid cells, where a segmental or whole chromosomal loss can result in a false strand state assignment. To circumvent such edge cases, we additionally examined strand states of each segment at a 5 Mb resolution, assigning states for each bin as defined in breakpointR v1.10.0 [34]. Any segment with a W/C ratio of more than (>) 0.8 or less than (<) -0.8 (90% of fragments) would be assigned a WW and CC state respectively. A segment with a W/C ratio between (>) -0.2 and (<) 0.2 is considered a WC state. Everything else is marked as an unassigned state (set as ? by breakpointR). If the strand state of the entire segment does not match the dominant state from the 5 Mb bins (more than (>=) 50% of bins), the segment strand state is reassigned to a composite of the top two most frequent bin states - the “final strand state”.

Annotation with Support Vector Machines (SVM) classifiers

All SVM classifiers were run using the R package e1071 v1.7-9 [124] with the radial basis kernel and default settings. For classifier prototyping we used the Orange data mining toolbox v3.30 [125]. For each cell line, a manual annotation dataset was curated. To evaluate the SVM classifiers we randomly sampled and split the manual annotation into a 70:30 training and test datasets. Confusion matrices were calculated using the R package caret v6.0-90 [126]. To obtain a representative rate of (mis)classifications, we performed 70:30 random sampling of the manually annotated datasets 20 times and used the mean value of all the confusion matrices. We use F1 score to evaluate SVM performance. The importance of each feature was assessed by randomly permuting the values of each feature 20 times, one feature at a time, and measuring the resulting decrease in F1 score. For haploid cells, the same 70:30 test and training dataset split was used for assessing feature importance. In diploid cells the number of features used in each classifier was much higher (4 in haploid versus 9-11 in diploid), meaning the importance of certain features may not be apparent in particular random samplings of the manually annotated dataset. As a result, for diploid cells we used the entire manually annotated dataset to assess each feature’s

importance. For the final annotation of all cells we trained each classifier on entire manually annotated datasets. Despite having a separate manual annotation for HAP1^{lig4-} cells for performance evaluation, the final annotation of both HAP1 and HAP1^{lig4-} cells used the HAP1 manual annotation.

Drop to zero annotation (DTZ)

Segments - especially in high coverage cells - contain background fragments that can prevent calling complete deletions, or “drops to zero”, based solely on empty bins (FigS15A). To identify DTZ segments, we trained an SVM classifier using a manually curated training dataset with RCC centered and RCBC centered scores as the input variables (FigS16B, FigS17B and D). The DTZ annotation setup is equivalent between haploid and diploid cells.

Loss of heterozygosity (LOH) annotation

Loss (monosomy) and cnLOH (copy neutral LOH) annotations only apply to diploid cells. The equivalent of loss in haploid cells is DTZ. In diploid cells we found the most important features for identifying LOH were haplotype fraction and final strand state. We therefore chose to annotate both loss and cnLOH with a single SVM (FigS17). All segments previously classified as DTZ were left out of LOH classification. Additional features, such as digital overlap counts, were important for distinguishing between loss and cnLOH, and between LOH and gain. We ultimately used the following input variables to train an SVM classifier for LOH annotation: RCC centered, RCBC centered, mappable width, coverage percentage, final strand states, ROC normalized, RWOC normalized, RCOC normalized, haplotype 1 fraction (FigS17B and D).

When manually annotating cnLOH, we found a substantial bias of events toward a single haplotype (Hap1 or Hap2) (FigS14). The resulting underrepresentation of one of the haplotypes in the training datasets caused the classifier to completely overlook cnLOH events with haplotype 1 fraction either above 0.9 or below 0.1 when applied to the full dataset. To overcome this haplotype bias, we inverted the haplotype 1 fraction below 0.5 (|1-haplotype fraction|) so that both haplotypes were now represented by a single fraction going from 0.5 to 1 (instead of 0 to 1).

Gain annotation

To simplify classification, we combined any increase in copy number of a segment or a whole chromosome under a single umbrella term, gain. In haploid cells, we used the final strand states, RCC centered, ROC normalized and coverage percentage as the input variables for classifying gain. A small number of haploid gain annotations contained a composite final strand state - WW-CC or CC-WW (FigS16C and “Final strand states”). These states arose from missed breakpoint calls that resulted in a false positive gain call. From the composite strand state we know these classifications are false and therefore we corrected them during post processing.

For diploid gains annotation, we mostly used the same input variables as for LOH (FigS17B and D). We first removed all segments classified with LOH before performing gain classification. We additionally modified the same strand overlap scores. Normalized same strand overlaps (RWOC and RCOC normalized) are informative for identifying gains in WC state segments. In WC segments they have the potential to identify chromosome imbalance arising from only a single chromosome duplication (trisomy). However, with WW and CC segments we found instances where the normalized DWOC and DCOC score proved detrimental and the singular normalized DOC score was sufficient for an accurate classification. We therefore set the normalized DWOC and DCOC score to 0 for WW and CC state segments respectively.

Finally, we added the “strand-neutral” state percentage (see “Sample selection and cell filters”) as additional variables for gain annotation in diploid cells. The strand-neutral state percentage were important features for identifying BJ-5ta cells with a high percentage of gain events that likely correspond to 4n cells. Despite also including these two features in the Patski cell classifier, their importance for gain annotation was negligible (FigS17B). The exclusive presence of high percentage gain cells in BJ-5ta suggests a cell line specific feature. Removing the “high gain” cells before evaluating feature importance put the features important of the BJ-5ta classifier in line with the Patski classifier (FigS17D). The “high gain” BJ-5ta cells display a range of strand-neutral state percentages that are also present in cells without any gain classifications, showing that these two features are not amenable to a simple cutoff that could have been performed during initial sample selection (FigS20D).

Post-processing and analysis of annotated events

For HAP1 and HAP1^{lig4-} cells we ran the SVM annotation using segmentation with and without the region filter (see “Segmentation”). Even though we obtained better event calls with the region filter and manually added regions, in HAP1 cells there were four instances where the coordinates better reflected the annotated event without the region filter, and were therefore switched out. In HAP1^{lig4-} there were only two such instances that were also switched out. In addition, there were four false positive gain event calls in HAP1^{lig4-} cells that were not present without the region filter, and were removed from the final annotated set.

We can be confident that any segments in haploid cells with a WW-CC or CC-WW state are false for gain as these composite states arise from missed breakpoint calls in single copy segments. Therefore, we manually corrected any such misclassifications of gain. Similarly in diploid cells, any segments without an LOH annotation, but with a WW-CC or CC-WW state were reclassified as loss. The two exceptions to this rule were for the composite state segments on chr14 in Patski cells and chr1 in BJ-5ta cells. In both cases, the composite state was the result of an inversion. Therefore, any composite state segments on chr1 in BJ-5ta or chr14 in Patski and with a haplotype fraction between (>) 0.2 and (<) 0.8 were not reclassified as loss. Due to the relative low frequency of composite state segments in diploid cells we had to exclude CC-WW or both CC-WW and WW-CC states from the classifier due to their absence in the training data.

We next consolidated the segment annotation by merging adjacent segments with identical labels. To retain a contiguous segmentation for merging, segments shorter than (<) 10 Mb and segments longer than 10Mb but with less than 100 unique fragments were readded (see “Segmentation”) and labeled as short and low coverage segments respectively. Chromosomes with a single unique annotation label (ignoring short and low coverage segments) were considered whole annotations, with everything else falling under segmental annotation. For whole chromosome annotations, all individual segments were merged into a single interval. The haplotype 1 fraction difference between adjacent segments was additionally examined in diploid cells to avoid merging segments with a haplotype switch. Merging was allowed only if the difference did not exceed 0.2. If the difference was greater, all whole chromosome segments were relabeled as segmental and further examined if segmental merging was required. For segmental annotation merging, the following rules were applied:

- All consecutive segments with identical annotations are merged together.
- Annotated segments flanked by short segments (<10Mb) will be expanded to include a single short segment on either side. However, if a short segment is situated between conflicting annotations, no merging will be performed.

- Consecutive segments with identical annotations separated by a single short segment (<10Mb) will all be merged together with the short segment.
- For segments in diploid cells that have met the above criteria, the haplotype 1 fraction of each annotated segment is additionally examined to avoid merging segments with a haplotype switch. If the haplotype 1 fraction difference of the two segments does not exceed 0.2, the segments are merged.

Following merging, annotated events were examined for a set of location features such as the right side of the centromere (RCEN) and the right end of the chromosome (RE) that enabled us to further classify events into arm, interstitial, terminal and whole (FigS25D). In the example of RCEN and RE, the annotation would be labeled as an arm event.

Clustering and diffusion maps of annotated events

To construct the distance matrix, each cell was first partitioned into 1Mb bins and each bin that overlapped an annotated event was assigned a value 1 to 8, depending on the event category. To improve the clustering resolution we additionally added a weight for each set of annotation categories. For haploid cells, DTZ and whole chromosome DTZ was given a weight of 0, while gain and disomy was given a weight of 40. For diploid cells, DTZ and whole chromosome DTZ was given a weight of 0, loss and monosomy was given a weight of 20, cnLOH and UPD was given a score of 40, and gain and trisomy was given a weight of 60. The weights were chosen based on the copy number of each event category. Alternatively, to prioritize cluster separation by chromosome 4 monosomy vs UPD - allowing us to highlight the mirrored nature of clonal populations whose sole branch point was the status of chromosome 4 - we altered the weights for DTZ and whole chromosome DTZ to 40, and cnLOH and UPD to 0. The weighted count matrix for each cell line was subsequently used to obtain the Canberra distance matrix using the factoextra package in R [127]. We chose Canberra distance as it was the only distance metric in combination with ward.D2 hierarchical clustering that separated “high gain” from near-haploid cells (FigS22B). An optimal number of clusters was chosen using the silhouette method implemented in factoextra [127]. With the maximum number of clusters to consider set to 200, the optimal number of clusters for HAP1 cells was 60, while for HAP1^{lig4-} cells it was 48, and for BJ-5ta cells it was 58. For Patski cells the optimal number of clusters was 177. However, with 177 clusters, most clusters contained only a single cell. This led us to split Patski cells into only 29 clusters that still preserved the pattern of clonal divergence. Lastly, we chose to manually reorder the 29 clusters to better highlight the gradual accumulation of events (Fig3D).

Diffusion maps were made using the destiny R package v3.8.1 [66]. Assigning the highest weights to gains and trisomy raised concerns that the diffusion map clusters would be driven by the weights, rather than categories of events. We tested numerous distance metrics and chose to use Canberra, the same distance used for hierarchical clustering, due to its tolerance of different weight combinations (FigS23A). With Canberra, we observed minimal impact on the structure of the diffusion maps when the weights were inverted to 60 for DTZ and whole chromosome DTZ and 0 for gain and trisomy. Other weights, such as the default euclidean distance used by destiny, turned out to be substantially influenced by different weights (FigS23A).

For Patski and BJ-5ta cells, the percentage of the mappable genome covered by gain or trisomy was calculated for each cell within a cluster and the median of all cells was used to determine which clusters contained cells with unusually high percentage of gain - “high gain” - annotations. Similarly, the percentage of loss or monosomy was used to determine “high loss” or near-haploid clusters. A threshold of greater than (>) 45 percent median coverage was chosen to identify “high gain” and “high loss”

clusters. Despite the speckled nature of gain annotations in Patski cells classed as “high gain” - compared to the contiguous gains in BJ-5ta cells - we chose to exclude “high gain” cells from downstream analysis for both cell lines.

Comparing annotated event rates within and between cell lines

We used chi-squared tests to determine significant enrichment of annotated events (SVs) within and between cell lines. In addition, we applied chi-squared tests to ascertain the significance of chromosomal location for each event class. For BJ-5ta cells, the sex chromosomes were excluded. For contingency table generation, events were counted on a per chromosome arm basis, accounting for cell ploidy. For whole chromosome events, counting by chromosome arm produces a count of two - except for acrocentric chromosomes - and counting segmental events by chromosome arm produces a count of one.

	HAP1	HAP1^{lig4-}	BJ-5ta	Patski
Cell count	1968	751	2157	829
Ploidy	1	1	2	2
Chromosomes per cell	23	23	44*	40
Chromosome arms per cell	41	41	78*	40
Total chromosome count	45,264	17,273	94,908*	33,160
Total arm count	80,688	30,791	168,246*	33,160

*not counting sex chromosomes

References

1. Mitchel, K., Zhang, H., Welz-Voegele, C., and Jinks-Robertson, S. (2010). Molecular structures of crossover and noncrossover intermediates during gap repair in yeast: implications for recombination. Mol Cell *38*, 211–222.

2. Yin, Y., Dominska, M., Yim, E., and Petes, T.D. (2017). High-resolution mapping of heteroduplex DNA formed during UV-induced and spontaneous mitotic recombination events in yeast. Elife *6*. https://doi.org/10.7554/eLife.28069.

3. Chovanec, P., and Yin, Y. (2021). A mapping platform for mitotic crossover by single-cell multi-omics. Methods Enzymol. *661*, 183–204.

4. Moynahan, M.E., and Jasin, M. (2010). Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nat. Rev. Mol. Cell Biol. *11*, 196–207.

5. Chua, P., and Jinks-Robertson, S. (1991). Segregation of recombinant chromatids following mitotic crossing over in yeast. Genetics *129*, 359–369.

6. Andersen, S.L., and Petes, T.D. (2012). Reciprocal uniparental disomy in yeast. Proc Natl Acad Sci U S A *109*, 9947–9952.

7. Hong, S., Sung, Y., Yu, M., Lee, M., Kleckner, N., and Kim, K.P. (2013). The logic and mechanism of homologous recombination partner choice. Mol Cell *51*, 440–453.

8. Yin, Y., Jiang, Y., Lam, K.-W.G., Berletch, J.B., Disteche, C.M., Noble, W.S., Steemers, F.J., Camerini-Otero, R.D., Adey, A.C., and Shendure, J. (2019). High-Throughput Single-Cell Sequencing with Linear Amplification. Mol. Cell *76*, 676–690.e10.

9. Loeillet, S., Herzog, M., Puddu, F., Legoix, P., Baulande, S., Jackson, S.P., and Nicolas, A.G. (2020). Trajectory and uniqueness of mutational signatures in yeast mutators. Proc. Natl. Acad. Sci. U. S. A. *117*, 24947–24956.

10. St Charles, J., Hazkani-Covo, E., Yin, Y., Andersen, S.L., Dietrich, F.S., Greenwell, P.W., Malc, E., Mieczkowski, P., and Petes, T.D. (2012). High-resolution genome-wide analysis of irradiated (UV and γ-rays) diploid yeast cells reveals a high frequency of genomic loss of heterozygosity (LOH) events. Genetics *190*, 1267–1284.

11. Sampaio, N.M.V., Ajith, V.P., Watson, R.A., Heasley, L.R., Chakraborty, P., Rodrigues-Prause, A., Malc, E.P., Mieczkowski, P.A., Nishant, K.T., and Argueso, J.L. (2020). Characterization of systemic genomic instability in budding yeast. Proc. Natl. Acad. Sci. U. S. A. *117*, 28221–28231.

12. Sui, Y., Qi, L., Wu, J.-K., Wen, X.-P., Tang, X.-X., Ma, Z.-J., Wu, X.-C., Zhang, K., Kokoska, R.J., Zheng, D.-Q., et al. (2020). Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. Proc. Natl. Acad. Sci. U. S. A. *117*, 28191–28200.

13. Loh, P.-R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F., Birmann, B.M., Talkowski, M.E., Bakhoum, S.F., McCarroll, S.A., et al. (2018). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature *559*, 350–355.

14. Latt, S.A. (1973). Microfluorometric detection of deoxyribonucleic acid replication in human metaphase chromosomes. Proc. Natl. Acad. Sci. U. S. A. *70*, 3395–3399.

15. Perry, P., and Wolff, S. (1974). New Giemsa method for the differential staining of sister chromatids. Nature *251*, 156–158.

16. Falconer, E., Hills, M., Naumann, U., Poon, S.S.S., Chavez, E.A., Sanders, A.D., Zhao, Y., Hirst, M.,

and Lansdorp, P.M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods *9*, 1107–1112.

17. Sanders, A.D., Hills, M., Porubský, D., Guryev, V., Falconer, E., and Lansdorp, P.M. (2016). Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Res. *26*, 1575–1587.

18. Sanders, A.D., Meiers, S., Ghareghani, M., Porubsky, D., Jeong, H., van Vliet, M.A.C.C., Rausch, T., Richter-Pechańska, P., Kunz, J.B., Jenni, S., et al. (2020). Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. Nat. Biotechnol. *38*, 343–354.

19. Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F.A., Harvey, W.T., et al. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell *185*, 1986–2005.e26.

20. Porubský, D., Sanders, A.D., van Wietmarschen, N., Falconer, E., Hills, M., Spierings, D.C.J., Bevova, M.R., Guryev, V., and Lansdorp, P.M. (2016). Direct chromosome-length haplotyping by single-cell sequencing. Genome Res. *26*, 1565–1574.

21. Porubsky, D., Ebert, P., Audano, P.A., Vollger, M.R., Harvey, W.T., Marijon, P., Ebler, J., Munson, K.M., Sorensen, M., Sulovari, A., et al. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat. Biotechnol. *39*, 302–308.

22. Sanders, A.D., Falconer, E., Hills, M., Spierings, D.C.J., and Lansdorp, P.M. (2017). Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat. Protoc. *12*, 1151–1176.

23. Hanlon, V.C.T., Chan, D.D., Hamadeh, Z., Wang, Y., Mattsson, C.-A., Spierings, D.C.J., Coope, R.J.N., and Lansdorp, P.M. (2022). Construction of Strand-seq libraries in open nanoliter arrays. Cell Rep Methods *2*, 100150.

24. Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., and Xie, X.S. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science *356*, 189–194.

25. Chovanec, P., and Yin, Y. (2025). Generalization of the sci-L3 method to achieve high-throughput linear amplification for replication template strand sequencing, genome conformation capture, and the joint profiling of RNA and chromatin accessibility. Nucleic Acids Res *53*. https://doi.org/10.1093/nar/gkaf101.

26. Porubsky, D., Garg, S., Sanders, A.D., Korbel, J.O., Guryev, V., Lansdorp, P.M., and Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. Nat Commun *8*, 1293.

27. Carette, J.E., Raaben, M., Wong, A.C., Herbert, A.S., Obernosterer, G., Mulherkar, N., Kuehne, A.I., Kranzusch, P.J., Griffin, A.M., Ruthel, G., et al. (2011). Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. Nature *477*, 340–343.

28. Llargués-Sistac, G., Bonjoch, L., and Castellvi-Bel, S. (2023). HAP1, a new revolutionary cell model for gene editing using CRISPR-Cas9. Front Cell Dev Biol *11*, 1111488.

29. Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

30. Lilue, J., Doran, A.G., Fiddes, I.T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins, J., Collins, S., Czechanski, A., et al. (2018). Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet *50*, 1574–1583.

31. Andersson, B.S., Beran, M., Pathak, S., Goodacre, A., Barlogie, B., and McCredie, K.B. (1987). Ph-positive chronic myeloid leukemia with near-haploid conversion in vivo and establishment of a continuously growing cell line with similar cytogenetic pattern. Cancer Genet Cytogenet *24*, 335–343.

32. Hills, M., O'Neill, K., Falconer, E., Brinkman, R., and Lansdorp, P.M. (2013). BAIT: Organizing genomes and mapping rearrangements in single cells. Genome Med. *5*, 82.

33. Bakker, B., Taudt, A., Belderbos, M.E., Porubsky, D., Spierings, D.C.J., de Jong, T.V., Halsema, N., Kazemier, H.G., Hoekstra-Wakker, K., Bradley, A., et al. (2016). Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. *17*, 115.

34. Porubsky, D., Sanders, A.D., Taudt, A., Colomé-Tatché, M., Lansdorp, P.M., and Guryev, V. (2020). breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. Bioinformatics *36*, 1260–1261.

35. Hamadeh, Z., Hanlon, V., and Lansdorp, P.M. (2022). Mapping of sister chromatid exchange events and genome alterations in single cells. Methods *204*, 64–72.

36. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.

37. Nambiar, M., and Smith, G.R. (2016). Repression of harmful meiotic recombination in centromeric regions. Semin. Cell Dev. Biol. *54*, 188–197.

38. Saayman, X., Graham, E., Nathan, W.J., Nussenzweig, A., and Esashi, F. (2023). Centromeres as universal hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. Mol. Cell *83*, 523–538.e7.

39. Bao, C., Tourdot, R.W., Brunette, G.J., Stewart, C., Sun, L., Baba, H., Watanabe, M., Agoston, A.T., Jajoo, K., Davison, J.M., et al. (2023). Genomic signatures of past and present chromosomal instability in Barrett's esophagus and early esophageal adenocarcinoma. Nat Commun *14*, 6203.

40. Loh, P.-R., Genovese, G., and McCarroll, S.A. (2020). Monogenic and polygenic inheritance become instruments for clonal selection. Nature *584*, 136–141.

41. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.

42. Essletzbichler, P., Konopka, T., Santoro, F., Chen, D., Gapp, B.V., Kralovics, R., Brummelkamp, T.R., Nijman, S.M.B., and Bürckstümmer, T. (2014). Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. Genome Res. *24*, 2059–2065.

43. Clouaire, T., Rocher, V., Lashgari, A., Arnould, C., Aguirrebengoa, M., Biernacka, A., Skrzypczak, M., Aymard, F., Fongang, B., Dojer, N., et al. (2018). Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures. Mol. Cell *72*, 250–262.e6.

44. Chiolo, I., Minoda, A., Colmenares, S.U., Polyzos, A., Costes, S.V., and Karpen, G.H. (2011). Double-strand breaks in heterochromatin move outside of a dynamic HP1a domain to complete recombinational repair. Cell *144*, 732–744.

45. Tsouroula, K., Furst, A., Rogier, M., Heyer, V., Maglott-Roth, A., Ferrand, A., Reina-San-Martin, B., and Soutoglou, E. (2016). Temporal and Spatial Uncoupling of DNA Double Strand Break Repair Pathways within Mammalian Heterochromatin. Mol. Cell *63*, 293–305.

46. Du, Q., Smith, G.C., Luu, P.L., Ferguson, J.M., Armstrong, N.J., Caldon, C.E., Campbell, E.M., Nair,

S.S., Zotenko, E., Gould, C.M., et al. (2021). DNA methylation is required to maintain both DNA replication timing precision and 3D genome organization integrity. Cell Rep. *36*, 109722.

47. Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

48. Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770.

49. Sidiropoulos, N., Mardin, B.R., Rodríguez-González, F.G., Bochkov, I.D., Garg, S., Stütz, A.M., Korbel, J.O., Aiden, E.L., and Weischenfeldt, J. (2022). Somatic structural variant formation is guided by and influences genome architecture. Genome Res. *32*, 643–655.

50. Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S.S.P., et al. (2017). Genome Organization Drives Chromosome Fragility. Cell *170*, 507–521.e18.

51. Sriramachandran, A.M., Petrosino, G., Méndez-Lago, M., Schäfer, A.J., Batista-Nascimento, L.S., Zilio, N., and Ulrich, H.D. (2020). Genome-wide Nucleotide-Resolution Mapping of DNA Replication Patterns, Single-Strand Breaks, and Lesions by GLOE-Seq. Mol. Cell *78*, 975–985.e7.

52. Welz-Voegele, C., and Jinks-Robertson, S. (2008). Sequence divergence impedes crossover more than noncrossover events during mitotic gap repair in yeast. Genetics *179*. https://doi.org/10.1534/genetics.108.090233.

53. Kumar, R., Nagpal, G., Kumar, V., Usmani, S.S., Agrawal, P., and Raghava, G.P.S. (2019). HumCFS: a database of fragile sites in human chromosomes. BMC Genomics *19*, 985.

54. Brown, R.E., and Freudenreich, C.H. (2021). Structure-forming repeats and their impact on genome stability. Curr. Opin. Genet. Dev. *67*, 41–51.

55. Zeng, C., Onoguchi, M., and Hamada, M. (2021). Association analysis of repetitive elements and R-loop formation across species. Mob. DNA *12*, 3.

56. Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

57. Hanlon, V.C.T., Porubsky, D., and Lansdorp, P.M. (2023). Chromosome-Length Haplotypes with StrandPhaseR and Strand-seq. Methods Mol Biol *2590*, 183–200.

58. Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet *20*, 437–455.

59. Feichtinger, J., and McFarlane, R.J. (2019). Meiotic gene activation in somatic and germ cell tumours. Andrology *7*, 415–427.

60. Lingg, L., Rottenberg, S., and Francica, P. (2022). Meiotic Genes and DNA Double Strand Break Repair in Cancer. Front. Genet. *13*, 831620.

61. Heasley, L.R., Watson, R.A., and Argueso, J.L. (2020). Punctuated Aneuploidization of the Budding Yeast Genome. Genetics *216*, 43–50.

62. Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., Zhang, H., et al. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat. Genet. *48*, 1119–1130.

63. Bollen, Y., Stelloo, E., van Leenen, P., van den Bos, M., Ponsioen, B., Lu, B., van Roosmalen, M.J., Bolhaqueiro, A.C.F., Kimberley, C., Mossner, M., et al. (2021). Reconstructing single-cell karyotype alterations in colorectal cancer identifies punctuated and gradual diversification patterns. Nat Genet *53*, 1187–1195.

64. Yin, Y., and Petes, T.D. (2013). Genome-wide high-resolution mapping of UV-induced mitotic recombination events in Saccharomyces cerevisiae. PLoS Genet *9*, e1003894.

65. Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics *31*, 2989–2998.

66. Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics *32*, 1241–1243.

67. McCulley, J.L., and Petes, T.D. (2010). Chromosome rearrangements and aneuploidy in yeast strains lacking both Tel1p and Mec1p reflect deficiencies in two different mechanisms. Proc Natl Acad Sci U S A *107*, 11465–11470.

68. Watson, E.V., Lee, J.J.-K., Gulhan, D.C., Melloni, G.E.M., Venev, S.V., Magesh, R.Y., Frederick, A., Chiba, K., Wooten, E.C., Naxerova, K., et al. (2024). Chromosome evolution screens recapitulate tissue-specific tumor aneuploidy patterns. Nat Genet *56*, 900–912.

69. Sdeor, E., Okada, H., Saad, R., Ben-Yishay, T., and Ben-David, U. (2024). Aneuploidy as a driver of human cancer. Nat Genet *56*, 2014–2026.

70. Dhillon, V.S., and Dhillon, I.K. (1998). Chromosome aberrations and sister chromatid exchange studies in patients with prostate cancer: possible evidence of chromosome instability. Cancer Genet. Cytogenet. *100*, 143–147.

71. Cefle, K., Ucur, A., Guney, N., Ozturk, S., Palanduz, S., Tas, F., Asoglu, O., Bayrak, A., Muslumanoglu, M., and Aydiner, A. (2006). Increased sister chromatid exchange frequency in young women with breast cancer and in their first-degree relatives. Cancer Genet. Cytogenet. *171*, 65–67.

72. Roy, S.K., Trivedi, A.H., Bakshi, S.R., Patel, R.K., Shukla, P.H., Patel, S.J., Bhatavdekar, J.M., Patel, D.D., and Shah, P.M. (2000). Spontaneous chromosomal instability in breast cancer families. Cancer Genet. Cytogenet. *118*, 52–56.

73. Baltaci, V., Kayikçioğlu, F., Alpas, I., Zeyneloğlu, H., and Haberal, A. (2002). Sister chromatid exchange rate and alkaline comet assay scores in patients with ovarian cancer. Gynecol. Oncol. *84*, 62–66.

74. Dhar, P.K., Devi, S., Rao, T.R., Kumari, U., Joseph, A., Kumar, M.R., Nayak, S., Shreemati, Y., Bhat, S.M., and Bhat, K.R. (1996). Significance of lymphocytic sister chromatid exchange frequencies in ovarian cancer patients. Cancer Genet. Cytogenet. *89*, 105–108.

75. Kim, M.-K., Zitzmann, S., Westermann, F., Arnold, K., Brouwers, S., Schwab, M., and Savelyeva, L. (2004). Increased rates of spontaneous sister chromatid exchange in lymphocytes of BRCA2+/- carriers of familial breast cancer clusters. Cancer Lett. *210*, 85–94.

76. Donnianni, R.A., and Symington, L.S. (2013). Break-induced replication occurs by conservative DNA synthesis. Proc Natl Acad Sci U S A *110*, 13475–13480.

77. Yang, Z., Takai, K.K., Lovejoy, C.A., and de Lange, T. (2020). Break-induced replication promotes fragile telomere formation. Genes Dev *34*, 1392–1405.

78. Bürckstümmer, T., Banning, C., Hainzl, P., Schobesberger, R., Kerzendorfer, C., Pauler, F.M., Chen, D., Them, N., Schischlik, F., Rebsamen, M., et al. (2013). A reversible gene trap collection empowers haploid genetics in human cells. Nat Methods *10*, 965–971.

79. Zhao, Y., Dominska, M., Petrova, A., Bagshaw, H., Kokoska, R.J., and Petes, T.D. (2017). Properties of Mitotic and Meiotic Recombination in the Tandemly-Repeated Gene Cluster in the Yeast. Genetics *206*, 785–800.

80. Zhang, H., Zeidler, A.F.B., Song, W., Puccia, C.M., Malc, E., Greenwell, P.W., Mieczkowski, P.A., Petes, T.D., and Argueso, J.L. (2013). Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae. Genetics *193*, 785–801.

81. Hussmann, J.A., Ling, J., Ravisankar, P., Yan, J., Cirincione, A., Xu, A., Simpson, D., Yang, D., Bothmer, A., Cotta-Ramusino, C., et al. (2021). Mapping the genetic landscape of DNA double-strand break repair. Cell *184*, 5653–5669.e25.

82. Pierce, A.J., Johnson, R.D., Thompson, L.H., and Jasin, M. (1999). XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. Genes Dev *13*, 2633–2638.

83. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

84. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83–89.

85. Balzano, E., and Giunta, S. (2020). Centromeres under Pressure: Evolutionary Innovation in Conflict with Conserved Function. Genes *11*. https://doi.org/10.3390/genes11080912.

86. Barra, V., and Fachinetti, D. (2018). The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. Nat. Commun. *9*, 4340.

87. McFarlane, R.J., and Humphrey, T.C. (2010). A role for recombination in centromere function. Trends Genet. *26*, 209–213.

88. Onaka, A.T., Toyofuku, N., Inoue, T., Okita, A.K., Sagawa, M., Su, J., Shitanda, T., Matsuyama, R., Zafar, F., Takahashi, T.S., et al. (2016). Rad51 and Rad54 promote noncrossover recombination between centromere repeats on the same chromatid to prevent isochromosome formation. Nucleic Acids Res. *44*, 10744–10757.

89. Logsdon, G.A., Rozanski, A.N., Ryabov, F., Potapova, T., Shepelev, V.A., Catacchio, C.R., Porubsky, D., Mao, Y., Yoo, D., Rautiainen, M., et al. (2024). The variation and evolution of complete human centromeres. Nature *629*, 136–145.

90. Heijink, A.M., Stok, C., Porubsky, D., Manolika, E.M., de Kanter, J.K., Kok, Y.P., Everts, M., de Boer, H.R., Audrey, A., Bakker, F.J., et al. (2022). Sister chromatid exchanges induced by perturbed replication can form independently of BRCA1, BRCA2 and RAD51. Nature Communications *13*, 1–16.

91. Sherman, M.A., Rodin, R.E., Genovese, G., Dias, C., Barton, A.R., Mukamel, R.E., Berger, B., Park, P.J., Walsh, C.A., and Loh, P.-R. (2021). Large mosaic copy number variations confer autism risk. Nat Neurosci *24*, 197–203.

92. Weaver, B.A.A., and Cleveland, D.W. (2006). Does aneuploidy cause cancer? Curr. Opin. Cell Biol. *18*, 658–667.

93. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.-Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

94. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu,

H., Liu, J., et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell *33*, 676–689.e3.

95. Girish, V., Lakhani, A.A., Thompson, S.L., Scaduto, C.M., Brown, L.M., Hagenson, R.A., Sausville, E.L., Mendelson, B.E., Kandikuppa, P.K., Lukow, D.A., et al. (2023). Oncogene-like addiction to aneuploidy in human cancers. Science, eadg4521.

96. Espiritu, S.M.G., Liu, L.Y., Rubanova, Y., Bhandari, V., Holgersen, E.M., Szyca, L.M., Fox, N.S., Chua, M.L.K., Yamaguchi, T.N., Heisler, L.E., et al. (2018). The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. Cell *173*, 1003–1013.e15.

97. Gordon, D.J., Resio, B., and Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. Nat. Rev. Genet. *13*, 189–203.

98. Song, W., and Petes, T.D. (2012). Haploidization in Saccharomyces cerevisiae induced by a deficiency in homologous recombination. Genetics *191*, 279–284.

99. Chan, K.Y., Yan, C.-C.S., Roan, H.-Y., Hsu, S.-C., Tseng, T.-L., Hsiao, C.-D., Hsu, C.-P., and Chen, C.-H. (2022). Skin cells undergo asynthetic fission to expand body surfaces in zebrafish. Nature *605*, 119–125.

100. Safavi, S., and Paulsson, K. (2017). Near-haploid and low-hypodiploid acute lymphoblastic leukemia: two distinct subtypes with consistently poor prognosis. Blood *129*, 420–423.

101. Baker, T.G., Alden, J., Dubuc, A.M., Welsh, C.T., Znoyko, I., Cooley, L.D., Farooqi, M.S., Schwartz, S., Li, Y.Y., Cherniack, A.D., et al. (2020). Near haploidization is a genomic hallmark which defines a molecular subgroup of giant cell glioblastoma. Neurooncol Adv *2*, vdaa155.

102. Kotecki, M., Reddy, P.S., and Cochran, B.H. (1999). Isolation and characterization of a near-haploid human cell line. Exp. Cell Res. *252*, 273–280.

103. Wilde, J.J., Aida, T., Del Rosario, R.C.H., Kaiser, T., Qi, P., Wienisch, M., Zhang, Q., Colvin, S., and Feng, G. (2021). Efficient embryonic homozygous gene conversion via RAD51-enhanced interhomolog repair. Cell *184*, 3267–3280.e18.

104. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics *32*, 289–291.

105. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10.

106. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Preprint, https://doi.org/10.1093/bioinformatics/btp324 https://doi.org/10.1093/bioinformatics/btp324.

107. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*. https://doi.org/10.1093/gigascience/giab008.

108. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. F1000Res. *10*, 33.

109. Deng, C., Daley, T., and Smith, A.D. (2015). Applications of species accumulation curves in large-scale biological data analysis. Quant Biol *3*, 135–144.

110. Deng, C., Daley, T., Calabrese, P., Ren, J., and Smith, A.D. (2016). Estimating the number of species

to attain sufficient representation in a random sample. arXiv [stat.ME].

111. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354.

112. Carroll, T., and Barrows, D. (2024). profileplyr: Visualization and annotation of read signal over genomic ranges with profileplyr. Bioconductor. http://bioconductor.org/packages/profileplyr/.

113. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

114. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., et al. (2022). The complete sequence of a human Y chromosome. bioRxiv. https://doi.org/10.1101/2022.12.01.518724.

115. Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693–707.e14.

116. Weddington, N., Stuy, A., Hiratani, I., Ryba, T., Yokochi, T., and Gilbert, D.M. (2008). ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. BMC Bioinformatics *9*, 530.

117. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

118. Bonora, G., Deng, X., Fang, H., Ramani, V., Qiu, R., Berletch, J.B., Filippova, G.N., Duan, Z., Shendure, J., Noble, W.S., et al. (2018). Orientation-dependent Dxz4 contacts shape the 3D structure of the inactive X chromosome. Nat. Commun. *9*, 1445.

119. Marchal, C., Sasaki, T., Vera, D., Wilson, K., Sima, J., Rivera-Mulia, J.C., Trevilla-García, C., Nogues, C., Nafie, E., and Gilbert, D.M. (2018). Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. Nat. Protoc. *13*, 819–839.

120. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res *34*, D590–D598.

121. Arnaud-Barbe, N., Cheynet-Sauvion, V., Oriol, G., Mandrand, B., and Mallet, F. (1998). Transcription of RNA templates by T7 RNA polymerase. Nucleic Acids Res. *26*, 3550–3554.

122. Cazenave, C., and Uhlenbeck, O.C. (1994). RNA template-directed RNA synthesis by T7 RNA polymerase. Proc. Natl. Acad. Sci. U. S. A. *91*, 6972–6976.

123. Zaher, H.S., and Unrau, P.J. (2004). T7 RNA polymerase mediates fast promoter-independent extension of unstable nucleic acid complexes. Biochemistry *43*, 7873–7880.

124. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., and Lin, C.-C. (2024). Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R package e1071 version 1.7-16]. Preprint at Comprehensive R Archive Network (CRAN).

125. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., et al. (2013). Orange: Data Mining Toolbox in Python. J. Mach. Learn. Res. *14*, 2349–2353.

126. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

127. Kassambara, A., and Mundt, F. (2020). Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7]. Preprint at Comprehensive R Archive Network (CRAN).