Table of Contents

I. Supplementary Discussion

In order to facilitate the use of the structural variants that we identify in the commonly used cancer cell lines profiled as part of this study, we have examined SVs in each cell line and compared these with previously reported recurrent SVs in each cancer type.

Prostate Cancer Cell Lines

It is estimated that greater than 50% of prostate cancers harbor gene fusions for ETS family transcription factors, namely *ERG*, *ETV1*, *ETV4*, and *ETV5* [1]. Our study examined two prostate cancer cell lines, LNCaP and PC-3. Interestingly, neither of these cell lines harbors an ETS family gene fusion. However, LNCaP cells harbor a gene fusion between the *MIPOL1* gene on chromosome 14 and the *DGKB* gene on chromosome 7, which is immediately upstream from *ETV1* [2], and *ETV1* has been shown to be critical for LNCaP cell invasion [3, 4]. Our Hi-C data indicates that the rearrangement resulting in the *MIPOL1-DGKB* gene fusion also results in the creation of a fusion TAD, which includes the downstream *ETV1* gene, suggesting that *ETV1* expression in these cells may be the result of repositioning of the gene into a novel regulatory environment (Supplementary Fig. 24a).

PC-3 cells have been shown previously to express high levels of *ETV4* and lack expression of *ERG*, *ETV1*, and *ETV5* [5]. Further, *ETV4* is essential for anchorage independent growth and cell migration in PC-3 cells [5]. Interestingly, rapid amplification of cDNA ends (RACE) assays of the *ETV4* transcript in PC-3 cells do not identify any 5' fusion partners, indicating the high levels of *ETV4* expression are not the result of a gene fusion [5]. We identify a translocation upstream of the *ETV4* gene fusion its locus on chromosome 17 with a locus on the long arm of chromosome 15 (Supplementary Fig. 24b). This appears to create a novel TAD fusion event as the result of this translocation.

Neuroblastoma Cell Lines

We analyzed data from two neuroblastoma cell lines, SK-N-DZ and SK-N-SH. SK-N-SH carries a translocation between chromosomes 7 and 8 near the *MYC* gene that appears to create a gene fusion (Fig. 5f). SK-N-DZ is a neuroblastoma cell line that expresses high levels of N-MYC and is reported to carry and *MYCN* amplification [6]. The *MYCN* gene is found on chromosome 2. Paradoxically, karyotyping data from the American Type Culture Collection (ATCC) and the European Collection of Authenticated Cell Cultures (ECACC) indicates that SK-N-DZ is a 44XX cell line that lacks both copies of chromosome 2. Interestingly, ATCC karyotyping also identifies several marker chromosomes, one of which contains a homogeneous staining region (HSR). In addition, SNP array CNV typing of the SK-N-DZ cell line identified the presence of chromosome 2 with several copy number alterations [7].

Our Hi-C data indicates that chromosome 2 in SK-N-DZ cells is the single most heavily rearranged chromosome in all of the cell lines we have analyzed in this study, with 46 independent rearrangements identified in chromosome 2 alone (Supplementary Fig. 24c,d). Our interpretation is that chromosome 2 has undergone a complex chromosomal rearrangement in these cells, possibly due to processes such as chromothripsis, and has experienced an amplification of the *MYCN* gene within this heavily rearranged chromosome. This would explain the apparent "loss" of chromosome 2 from karyotyping while still observing chromosome 2 genetic material from SNP array CNV analysis, and potentially explain the presence of a marker chromosome with an HSR.

We have additionally profiled the SK-N-AS neuroblastoma cell line. As mentioned in Fig. 5g, this cell line, like SK-N-SH, carries a translocation near the *MYC* gene that appears to create a TAD fusion event. In addition, recurrent structural variations near the *TERT* gene have recently been identified in neuroblastoma samples [8, 9]. We find a ~50Mb inversion in chromosome 5 in SK-N-AS that maps downstream from the *TERT* gene which appears to create a TAD fusion in the vicinity of the *TERT* gene (Supplementary Fig. 24e). This indicates that the SK-N-AS cell line may serve as a useful model for study the effects of *TERT* rearrangements in neuroblastoma tumors.

Pancreatic Cancer Cell Line

Our study examined one pancreatic cancer cell line (PANC-1), which we found harbors multiple rearrangements near known pancreatic cancer oncogenes, namely *KRAS* and *ERBB2*. Recent whole genome sequencing studies have identified recurrent focal amplifications of the *ERBB2* gene, suggesting it may be a driver of pancreatic cancer [10]. Our analysis of PANC-1 cells identified a ~3Mb deletion on chromosome 17 upstream of the promoter of the *ERBB2* gene, such that the gene forms a neo-TAD with novel interactions with a region normally located more than 3Mb away (Supplementary Fig. 24f).

In addition, we find a translocation between chromosome 11 and 12 that occurs near the *KRAS* gene. *KRAS* is mutated in ~90% of Pancreatic cancers [10]. Analyzing previously generated exome sequencing data shows that Panc-1 carries the G12D potent activating mutation in *KRAS*. There appears to be copy number change of the *KRAS* locus as well, as the G12D allele exists at roughly a 2:1 ratio to the wild type allele. Interesting, there is a further imbalance in the allelic fraction of RNA-seq data, with the G12D allele expressed at a 3:1 ratio at the RNA level. Our Hi-C results identify a complex rearrangement near the *KRAS* gene. Interestingly, the only Hi-C sequence read pair where we find the G12D mutation in the *KRAS* gene on chromosome 12 is a read that aligns between chromosomes 11 and 12, suggesting that the G12D allele may also be the translocated allele.

Breast Cancer Cell Lines

Our study examined two commonly used breast cancer cell lines T47D and MCF7. MCF7 carries a complex rearrangement between chromosomes 17 and 20 with 13 unique rearrangements detected either within or between the two chromosomes (Supplementary Fig. 24g). Interestingly, these rearrangements include regions that are recurrently amplified in breast cancer, such as the 17q23.1 and 20q13.2 loci [11].

T47D contains a translocation between chromosomes 8 and 14 (Supplementary Fig. 24h) where the breakpoint site on chromosome 8 lines in a one of the most frequently amplified regions in breast cancer genomes [11]. This rearrangement appears to create a novel TAD containing the known breast cancer oncogene *ZNF703* [12].

Lymphoma Cell Line

We analyzed previously published Hi-C data from the RL cell line [13]. This is a B-cell lymphoma cell line that has been previously shown to harbor a t(14;18) *IGH-BCL2* rearrangement [13]. We identify this rearrangement using our algorithm. We also find several additional rearrangements in this cell line. One of these rearrangements is a t(3;8) translocation where the chromosome 8 breakpoints maps ~120kb upstream of the *MYC* promoter. The presence of the

BCL2 and *MYC* rearrangements suggests that RL may in fact be a “double hit” lymphoma cell line [14].

Ewing’s Family Tumor

We profiled the SK-N-MC cell line using Hi-C, Bionano optical mapping, WGS, and karyotyping. SK-N-MC was originally described as a Neuroblastoma, but subsequent identification of the *EWSR-FLI1* fusion gene has led to the reclassification of this cell line as a Ewing’s Family Tumor. We identify the *EWSR-FLI1* fusion gene using multiple independent methodologies. Beyond, the *EWSR-FLI1* fusion gene, we also identify multiple intra-chromosomal rearrangements near the *MYC* gene, indicating that the SK-N-MC cell line harbors a *MYC* rearrangement as a second hit mutation.

Melanoma Cell Lines

We profiled two melanoma cell lines, SK-MEL-5 and RPMI-7951. SK-MEL-5 contains a 6Mb deletion on the p-arm of chromosome 9 in a region that spans the *CDKN2A* locus, a region frequently deleted in melanoma [15]. Interestingly, there are no Hi-C reads at all that align to this entire region, indicating that SK-MEL-5 is either haploid for chromosome 9 or that the deletion is bi-allelic. We find several rearrangements in RPMI-7951, but none of them have been previously shown to be associated with melanoma [15].

Lung Cancer Cell Lines

We profiled two lung cancer cell lines as part of our study. *MYC* gene amplifications have been previously shown to be a frequent events in lung cancer samples [16]. In the NCI-H460 cell line, we find a chromosome t(8;12) translocation at the *MYC* locus. Examining the raw coverage plots in the Hi-C data also shows that this region has likely undergone a high-level amplification as well. In A549, we identified a previously described *WDR72-SCAMP2* fusion gene [17]. We find several additional rearrangements, but none are known to be recurrent structural variants in lung cancer [16].

Other cell lines

KBM7, K562 are both myeloid leukemia cell lines contain the *BCR-ABL1* fusion gene. We identify this rearrangement in both cell lines. In addition, we find multiple additional rearrangements, but no recurrent alterations or known structural variants associated with myeloid leukemia.

SJCRH30 is a Rhabdomyosarcoma cell line known to carry the *PAX3-FOXO1* fusion gene. We identify this rearrangement, but we identify no additional structural variants known to be associated with Rhabdomyosarcoma in this cell line.

Several additional cell lines were profiled as part of our study. In most cases we identified multiple SVs in each cell line. However, the SVs we identified did not match any known recurrent SVs in their respective tumor types. These could represent either novel low frequency recurrent SVs or simple passenger mutations. The cell lines in this group include Caki-2, G401, MHH-CALL-4.

II. Supplementary Methods

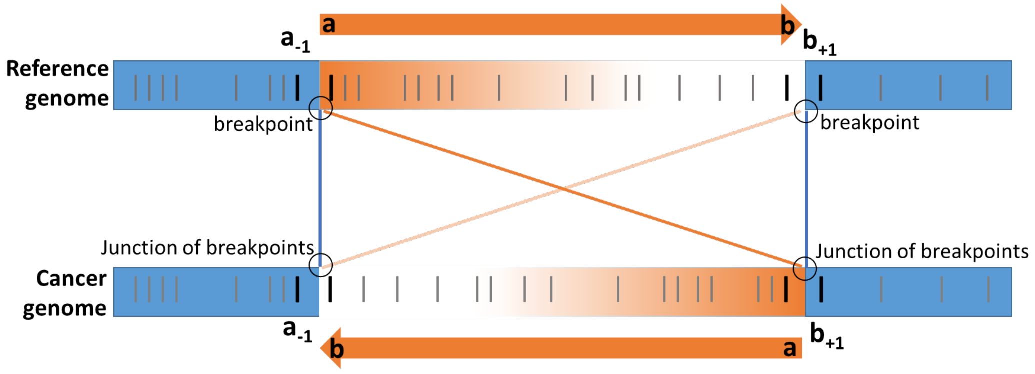
II.a. Filtration and classification of SVs detected by optical mapping

Duplicates of SVs can be generated during SV detection from different contigs mapping to the same region, and such duplicate SV calls are merged into a single SV call. For deletions and insertions, we further remove small indels with a size smaller than 50bp. Many of the deletions we detect overlap with genomic gaps. This is most likely the result of overestimation of gap sizes. In this sense, these are not true deletions but instead assembly errors (or regions with polymorphic gap sizes). We classify deletions as gap errors if the deletion recurrently appears in different cell lines and at least 30% of the deletion overlaps with gaps, and at least 80% of the gap overlaps with the deletion. We remove these "gap errors" from the list of deletions and use them for gap size re-estimation analysis.

We also developed strategies to filter SVs in close proximity to the centromere. In peri-centromeric regions, we noticed that contigs can have ambiguous alignments to multiple regions due to redundant labeling patterns, which result in the appearance of deletions that cross the centromere. We therefore remove recurrent large deletions (80% reciprocal overlap, >1Mb) crossing centromeres. We further stratify deletions larger than 100kb into two categories, one where sequences within the deleted region show reduced mappability and one where the sequences are mappable. We then filter deletions over mappable regions that are not supported by a loss of coverage in WGS data. Deletions that are supported by valley of WGS coverage are annotated as "High confidence" in Supplementary Table 4.

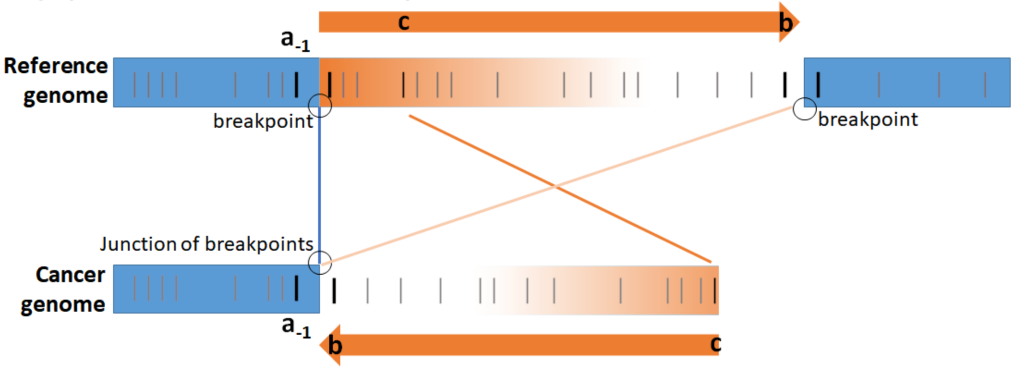
Defining inversions by Irys: A simple inversion involves two breakpoints and each breakpoint is represented by a pair of loci. Figure A below shows an example: the left breakpoint of this inversion occurs between nicking sites a_{-1} and a, and the right breakpoint occurs between nicking sites b and b_{+1}. The orange sequence in the middle is inverted and forms two breakpoint junctions: the left junction between (a_{-1} and b) and the right junction between (a and b_{+1}). We use the distance between sites a and b to approximate the size of this inversion (distance=b-a, Supplementary Table 4). To compare with inversions detected by other methods such as WGS and Hi-C, we used the junction of breakpoints (a_{-1}, b) and (a, b_{+1}) (reported in Supplementary Table 8). Such inversions with both junction of breakpoints resolved and four loci available are called "paired inversions".

Figure A
Irys paired inversion: both breakpoints resolved



Due to technical limitations, Irys may also detect an incomplete inversion in cancer genomes. As shown below, at the left end, Irys detects the junction of breakpoints (a_{-1}, b) in the cancer genome, but at the other end, its contig stops at loci c and cannot reveal the real junction of other breakpoint. Such inversions with only one junction of breakpoint resolved are named “partial inversions” by Bionano Iry (Supplementary table 4). In this scenario, we use the distance between loci b and c (b-c) to calculate the minimal size of this inversion (Supplementary table 4). To compare with WGS and Hi-C, we only use the resolved breakpoint junction (a_{-1}, b). Therefore, the two columns of positions reported Supplementary table 8 only represent the breakpoint junctions and cannot be used to estimate the size of inversion.

Figure B
Irys partial inversion: one breakpoint resolved



According to a recent study from Pendleton et al. [18], some inversions detected against hg19 were no longer detected against hg38, and they turned out to be inverted assembly of contigs in reference genome hg19, which were corrected in hg38. Those inversions have a unique feature that they are flanked by genomic gaps at each side. We scanned through inversions detected by our pipeline against hg38 and also found inversions flanked by gaps, which could represent inverted assembly of genome contig in hg38, or variations across populations that could be in both orientations in human genome. We thus remove such inversions to ensure we focus on genomic rearrangements and not genome assembly anomalies.

We observed that regions of the reference genome that harbor similar sequences distributed across multiple regions appear to harbor recurrent translocations in many samples. This is most likely due to misalignment of optical DNA reads leading to fixed false detection of translocations. A list of recurrent false-positive translocations was hence generated by comparing translocations detected across ten samples (with difference less than 1Mb away for both breakpoints), and the calls matching the list were removed. This list of recurrent translocation did not match any translocations detected by Hi-C or WGS, confirming that these are likely false positives.

II.b. Breakpoint calling based on Hi-C

II.b.i. Overview and rationale

We have developed a computational algorithm for detecting structural variants based on Hi-C data. Previous reports have noted that structural variations cause marked deviation in interaction frequencies in Hi-C datasets as a consequence of the altered linear proximity of regions of the genome due to the rearrangement. Our purpose was to use this signature as a means of finding structural variants *de novo* from Hi-C data. In brief, we use an iterative or progressive approach where we initially identify SVs at a resolution of 1Mb (for inter-chromosomal rearrangements) or 100kb bins (for intra-chromosomal rearrangements). After the initial set of SVs are found using this low resolution binning, we progressively decrease the bin size and re-run the algorithm focused on the regions identified in the previous low resolution SV call set. In other words, we first identify SVs at a resolution of 1Mb, and then we identify SVs within the regions identified at 1Mb resolution using 100kb bins. This process is repeated until the final minimum bin size is reached. This minimum bin size is largely defined based on the restriction enzyme used in the Hi-C experiment. For 6-base cutting enzyme which would be expected to cut every ~4kb, we use a minimum bin size of 10kb. For 4-kb cutting enzymes, which would be expected to cut every ~200-300 bp, we use a minimum bin size of 1kb. Some SVs will not be able to be identified at higher resolution passes of the algorithm. We have seen that this often occurs when the SV breakpoints appear to be near large repetitive regions. For instance, many rearrangements that whose “peak” signal is immediately adjacent to the centromeric region of a chromosome (and therefore the break potentially lies within the centromere), are identified only in low resolution passes of the algorithm, yet these rearrangements can be confirmed by FISH or karyotype analysis. The reason we use this progressive approach is due to algorithmic complexity. We use algorithms for finding SVs that rely on finding maximally summed submatrices in the large interaction matrix (see below for details). The complexity of finding maximally summed submatrices scales with the linear size of the matrix to the third power (*$O(n^3)$).* While the algorithm typically runs for <20 minutes at a resolution of 1Mb, if this was instead run initially at 1kb, this would increase the run time by a billion fold. Therefore, we chose the progressive strategy we have outlined in the methods in order to optimize run time while still allowing for high-resolution identification of breakpoints.

The primary challenge in using Hi-C data to find SVs is to distinguish signal derived from a rearrangement from signal derived for normal deviations in Hi-C interaction frequencies that result from cell biological features of nuclear organization. To this end, we developed a probabilistic model of Hi-C interaction frequencies to account for certain cell type variable or invariant patterns of chromatin interactions. Several cell-type invariant features impact the ability to identify re-arrangements. With regard to inter-chromosomal interactions, this is most clearly observed as preferential interactions between small chromosomes and between the ends of heterologous chromosomes. With regard to intra-chromosomal interaction, the clearest cell type invariant feature that influences our ability to detect re-arrangements are the Topologically Associated Domain (TAD) patterns. In addition, cell type variable patterns in interaction frequency can influence the ability to detect re-arrangements in Hi-C data, most notably the A/B compartment patterns. Lastly, there are a variety of inherent genomic features, including mappability, GC-content, restriction enzyme density, and copy number that must be taken into consideration. The details of how these features are modeled and used to identify re-arrangements are as follows:

II.b.ii. Calculation of expected interaction frequencies

Previous Hi-C studies have demonstrated that there are certain largely cell-type invariant patterns in chromatin interaction frequencies. The three most prominent features that we have identified which can impact the ability to identify structural variants in the genome are the increased association between small chromosomes [19], the increased frequency of interactions between the ends of heterologous chromosomes [20], and the intra-chromosomal TAD patterns of the genome [21, 22]. The first two features, namely the association of small chromosomes and the interactions between heterologous chromosome ends, impact identification of inter-chromosomal re-arrangements, whereas TAD patterns impact the ability to identify intra-chromosomal re-arrangements.

To account for the impact of small chromosome and chromosome end association, we estimated an average inter-chromosomal interaction frequency matrix across nine karyotypically normal cell lines (GM12878, H1 hESC, Mes, MSC, NPC, Troph, IMR90, HUVEC, HMEC). As inter-chromosomal interactions are influenced by A/B compartment patterns, we perform this averaging on an A/B subtracted matrix for each cell line. This is accomplished by taking the normalized interaction matrix, X (described above), and subtracting the matrix D (described above), representing the additive increase or decrease in interaction frequency due to A/B compartment patterns, to generate a new matrix, A, for each cell type:

$$A = X - D$$

Each element $a_{i,j}$ of the matrix A is then normalized by the global average of all elements within A, and then averaged across all 9 cell lines. This generates a new final matrix as follows:

$$\bar{a}_{i,j} = \frac{1}{n}\sum_{1}^{n} a_{i,j}$$

As a result, the value $\bar{a}_{i,j}$ represents the expected fold change in interaction frequency relative to the mean between any two bins i and j as a result of these cell type invariant inter-chromosomal interaction patterns. Of note, this step is performed only using a bin size of 1Mb, as the effect of these features is diminished at higher bin sizes.

We take a similar approach with regards to intra-chromosomal interaction to account for TAD patterns. Specifically, within a given cell type, we calculate the average interaction frequency of all bins i and j separated by a given distance d. Each interaction is then divided by this distance based average to produce a distance normalized interaction frequency:

$$z_{i,j} = \frac{n_{i,j}}{\mu_d}$$

where $z_{i,j}$ is the distanced normalized interaction frequency between bins i and j, $n_{i,j}$ is the normalized interaction frequency, and μ_d is the average interaction frequency between all bins separated by a distance d. The value $z_{i,j}$ is then averaged across the nine normal cell lines:

$$\bar{z}_{i,j} = \frac{1}{n}\sum_{1}^{n} z_{i,j}$$

where $\bar{z}_{i,j}$ represents that expected fold change in interaction frequency between bins i and j separated by distance d relative to the average interaction frequency at distance d. This is performed only for bins at 100kb, as our initial search for re-arrangements starts using 100kb bins and then progresses to smaller bin sizes, and as a result, TAD interactions are largely accounted for in this initial search (see below for details).

II.b.iii. Modeling of Hi-C interaction frequencies

To model Hi-C interaction frequencies, we developed a probabilistic model using a negative binomial distribution parameterized by a mean m, and dispersion parameter, r. These are calculated from the observed normalized interaction frequencies at a given bin size for all inter-chromosomal interactions and for all intra-chromosomal interactions at a given distance, d. For the parameter m, this is simply the mean. For inter-chromosomal interactions, this is given by:

$$m_{inter} = \mu_{inter} = \frac{1}{n}\sum_{i,j=1}^{n} n_{i,j}$$

where $n_{i,j}$ is the normalized interaction frequency between all bins i and j not on the same chromosome. For intra-chromosomal interactions at a given distance d, this is given by:

$$m_d = \mu_d = \frac{1}{n}\sum_{i,j=1}^{n} n_{i,j}$$

where $n_{i,j}$ is the normalized interaction frequency between all bins i and j separated by a given distance d. We similarly compute the variance of the data, σ^2, for inter-chromosomal interactions and all intra-chromosomal interactions at a given distance d. This can then be used to compute the dispersion parameter given by:

$$r_{inter} = \frac{{\mu_{inter}}^2}{{\sigma_{inter}}^2 - \mu_{inter}}$$

for all inter-chromosomal interactions, and:

$$r_d = \frac{{\mu_d}^2}{{\sigma_d}^2 - \mu_d}$$

for all intra-chromosomal interactions separated by a distance d.

While this approach can be used as a very general model of Hi-C interaction frequencies, various cell type variant and invariant features of Hi-C data can influence the interaction frequencies beyond this simple approach, as mentioned above. To account for this, the parameter m is modified for each pair of bins, i and j, to generate a bin specific parameter, $m_{i,j}$. For inter-chromosomal interactions, this is given by the following equation:

$$m_{i,j} = \left(m_{inter} \times e_{i,j} + b_{i,j}\right) \times q_i \times q_j$$

and for intra-chromosomal interactions this is given by:

$$m_{i,j} = m_d \times e_{i,j} \times q_i \times q_j$$

where m_{inter} and m_d are the parameter m for either inter-chromosomal interactions or intra-chromosomal interactions at a distance d, respectively. $e_{i,j}$ is the expected fold change in interaction frequency given the cell type invariant features of genome organization. $b_{i,j}$ is the additive change in interaction frequency due to A/B compartment patterns, and q_i and q_j are the intrinsic coverage biases of bins *i* and *j*. The features $b_{i,j}$ and $e_{i,j}$ are only used when calculating $m_{i,j}$ for inter-chromosomal interactions at a bin size of 1Mb. At higher resolutions for inter-chromosomal interactions, these terms are omitted, leaving:

$$m_{i,j} = m_{inter} \times q_i \times q_j$$

Likewise, $e_{i,j}$ is only used for intra-chromosomal modeling at a bin size of 100kb, and at higher resolution bin sizes this term is omitted, leaving:

$$m_{i,j} = m_d \times q_i \times q_j$$

As a result, we can use these parameters to then calculate the probability of observing $o_{i,j}$, the given number of sequence reads between bins *i* and *j*, using the following equations:

$$P\left(o_{i,j} = k \middle| m_{i,j}, r\right) = \left(\frac{\Gamma(r+k)}{k!\,\Gamma(r)}\right)\left(\frac{m_{i,j}}{r+\ m_{i,j}}\right)^k \left(\frac{r}{r+m_{i,j}}\right)^r$$

In this, the probability P can be considered as P_{inter}, the probability of observing $o_{i,j}$ given that the interaction is an inter-chromosomal interaction, if the parameter $m_{i,j}$ was calculated from m_{inter}. Likewise, P can be considered as P_d, the probability of observing $o_{i,j}$ given that the interaction is intra-chromosomal and separated by distance *d*, if the parameter $m_{i,j}$ was calculated from m_d.

For the purpose of finding re-arrangements, the reason for modeling Hi-C interactions with this approach is to assign a probability of observing the number of reads arising between any two bins *i* and *j* given that *i* and *j* are on different chromosomes or are on the same chromosome and separated by a distance *d*. This can be then compared with the probability of observing the same number of reads given that bins *i* and *j* are in fact re-arranged. If *i* and *j* are re-arranged, then we would expect the number of reads between them to reflect interaction frequencies between bins that are proximal along the linear distance of the chromosome. In this regard, we also need to calculate the probability of observing $o_{i,j}$ given that *i* and *j* are “local” along the linear distance of the genome. While we were previously modeling the probability of observing a specific number of reads given that two regions are intra-chromosomal and separated by a given distance *d*, we can generalize this to a value P_{local}, as a mixture model:

$$P_{local} = \sum_{d=1}^{d_{max}} w_d \times P_d$$

In this case, P_d is the probability of observing $o_{i,j}$ between bins *i* and *j* at distance *d*, w_d is the weight applied to each distance, and d_{max} is the maximum intra-chromosomal distance considered for “local” interactions. The weight w_d is related to the fraction of bins in a given matrix separated by the distance *d*. If we consider *d* as the distance in bins between two loci instead of as the genomic distance, then we can simply calculate w_d as follows:

$$w_d = \frac{2 \times (d_{max} - d)}{{d_{max}}^2 - d}$$

The value of the parameter d_{max} is arbitrary, and can be modified to adjust the sensitivity or specificity of the calls. We typically use a value of d_{max} that is 10 times larger than the bin size.

Having calculated the values for P_{local}, P_{inter}, or P_d, for a given pair of bins, we can now represent these probabilities as an odds ratio. This gives the relatively likelihood of a given interaction resulting from a re-arrangement or from the expected genomic structure. For inter-chromosomal interactions, this is represented as:

$$u_{i,j} = \frac{P_{local}}{P_{inter}}$$

where $u_{i,j}$ is the odds ratio for the interaction between bins *i* and *j*. For intra-chromosomal interactions, this can be represented as:

$$u_{i,j} = \frac{P_{local}}{P_d}$$

In practice, we consider this as a log-odds ratio, such that this value will be positive if the interaction is more likely to result from a re-arrangement and negative if more likely to result from the expected genomic structure. In summary, our probabilistic model allows us to convert the original matrix of observed interaction frequencies into a matrix of log-odds ratios.

II.b.iv. Breakpoint calling (inter-chromosomal)

In the event of a re-arrangement, the observed interaction frequencies of regions in the immediate proximity of the re-arrangement will more closely resemble local intra-chromosomal interactions than inter-chromosomal interactions. As a result, the log-odds ratios of interacting regions in the immediate proximity of the re-arrangement will be positive. Therefore, in order to find re-arrangements, we can search through the matrix of log-odds ratios for sub-matrices with positively summed log-odds ratios. Mathematically, the sum of the log-odds ratios of the elements of a sub-matrix will yield the log-odds ratio of the entire sub-matrix:

$$\log(u_{s\to t,u\to v}) = \sum_{i=s}^{t}\sum_{j=u}^{v}\log(u_{i,j})$$
$$= \log\left(\frac{P_{s,u_{local}} \times P_{s+1,u_{local}} \times P_{s,u+1_{local}} \times \dots \times P_{t,v_{local}}}{P_{s,u_{inter}} \times P_{s+1,u_{inter}} \times P_{s,u+1_{inter}} \times \dots \times P_{t,v_{inter}}}\right)$$

where $u_{k\to l,m\to n}$ is the odds ratio of the rectangular sub-matrix from bin k to l and m to n. This term represents the likelihood of all bins within the sub-matrix deriving from local intra-chromosomal interactions relative to the likelihood of the bins deriving from inter-chromosomal interactions. This can therefore represents the relatively likelihood of the sub-matrix resulting from a translocation or from the expected genomic structure.

Algorithmically, we search for sub-matrices with the maximal sum. We begin by considering the matrix of log-odds ratios at 1Mb between any pair of heterologous chromosomes. Within this matrix, we search for the maximally summed sub-matrix. The coordinates of this sub-matrix are saved, and their values are then set to a number below zero. This process is repeated iteratively so long as the sum of the maximum sum sub-matrix is above a pre-defined threshold. This threshold is determined by converting the odds-ratio of the maximally summed sub-matrix into a p-value. This p-value is defined as one minus the overall probability to the log-odds from the maximally summed sub-matrix, defined by the logistic function:

$$p - value = 1 - P_{translocation} = 1 - \frac{e^{s'}}{1+e^{s'}} = \frac{1}{1+e^{s'}}$$

where $P_{translocation}$ is the probability that the sub-matrix is translocated. The term S' is given by:

$$S' = \log(u_{s\to t,u\to v}) + \log\left(\frac{Prior_{local}}{1 - Prior_{local}}\right)$$

Where $\log(u_{s\to t,u\to v})$ is the odds ratio calculated above for the sub-matrix from bin k to l and m to n, and $Prior_{local}$ is our assumed prior probability of a given region being translocated. Our prior assumption is that the likelihood of any region being involved in a translocation is rare. We therefore assign this prior a value of 10^{-6}. This value is arbitrary, and as we shall see, this term has a relatively small affect on the final p-value.

As the term S' increases, the term $1 + e^{s'}$ can be approximated by $e^{s'}$. Therefore, at large values of S', the log of the p-value approaches negative one times S'. Since we identify the maximally summed sub-matrix and therefore the maximal value of S', this approximation is reasonable in most cases we have considered thus far. However, considering this value alone is not sufficient. There are many possible sub-matrices within a given matrix, and therefore multiple testing must be considered. Specifically, for a matrix with m columns and n rows, the number of possible sub-matrices is given by the following equation:

$$H_{a,b} = \frac{m \times (m-1) \times n \times (n-1)}{4}$$

where $H_{a,b}$ is the number of possible sub-matrices in the matrix between chromosomes a and b. Considering only the matrix between chromosomes 1 and 2 at a resolution of 1Mb, this still contains more than 800 million possible sub-matrices. We compute this term for each matrix between heterologous chromosomes, and summing these values across all chromosomes gives a genome wide value, H_{genome}, for the number of possible sub-matrices. We use a Bonferroni correction for multiple testing correction, such that our predefined threshold for our p-value must be below $0.05/H_{genome}$. As a result, we can define a minimum threshold for $\log{(u_{s\to t,u\to v})}$ as follows:

$$\log{(u_{s\to t,u\to v})_{min}} = -1 \times \log(0.05) + \log\left(H_{genome}\right) + \log{(\frac{1 - Prior_{local}}{Prior_{local}})}$$

In practice, the term $\log\left(H_{genome}\right)$ dominates, as this value is greater than 20 considering only the matrix between chromosome 1 and chromosome 2 alone, and adjusting the value of $Prior_{local}$ has only modest effects on the final threshold value. We then consider any sub-matrix whose log-odds sum is greater than $\log{(u_{s\to t,u\to v})_{min}}$ as being the result of a translocation. As mentioned previously, we search iteratively for maximally summed sub-matrices in the event that there is more than 1 re-arrangement on a given chromosome. However, in the event that we find a re-arrangement on the first iteration, we now expect there to be regions between the two chromosomes in question that result from intra-chromosomal interactions. Therefore, to limit the likelihood of subsequently identifying weaker, non-breakpoint proximal re-arrangements, we increase this minimum threshold by a factor of 2 with each iteration. This was determined empirically as most re-arranged regions have maximally summed log-odds that greatly exceed this threshold, and that subsequent iterations can easily find additional “weak” intra-chromosomal interactions after finding the maximally summed sub-matrix. Lastly, in the event that more than 1 maximally summed sub-matrices are identified for a given pair of chromosomes, we also check if the merged union of the two matrices would also be above this initial threshold. This is because our method of finding the maximal summed sub-matrix will occasionally split a re-arrangement into multiple sub-matrices, particularly if there are sub-sequent mutations such as deletions of on the re-arranged allele.

The above process describes the method used as a first pass to find re-arrangements between chromosomes at 1Mb. After this initial pass, we further analyze the data with increasingly smaller bin sizes. Each iteration uses the coordinates derived from the prior iteration as limits on the search space. In this regard, we search for sub-matrices at 100kb resolution only within regions identified in the initial 1Mb step. This limits the total space that each iteration must search and allows for a computationally tractable solution for identifying re-arrangements even at high resolutions. Furthermore, by beginning at large bin sizes, this allows for the use of even low depth sequencing to find re-arrangements. Computationally, the steps at higher bin sizes are identical except for limiting the search space to the regions defined in a previous iteration. Typically, we decrease the bin size by a factor of 10 for each iteration, from 1Mb, to 100kb, to 10kb, and possibly down to 1kb depending on the restriction enzyme used in the Hi-C experiment.

II.b.v. Breakpoint calling (intra-chromosomal)

The process of finding intra-chromosomal re-arrangements if nearly identical to the process described above for inter-chromosomal re-arrangements with only a few notable differences. First, the sub-matrices use the term P_d instead of P_{inter} to compute the maximal sum sub-matrix:

$$\begin{aligned}\log\,(u_{s\rightarrow t,u\rightarrow v}) &= \sum_{i=s}^{t}\sum_{j=u}^{v}\log\,(u_{i,j})\\ &= \log\,(\frac{P_{s,u_{local}} \times P_{s+1,u_{local}} \times P_{s,u+1_{local}} \times \dots \times P_{t,v_{local}}}{P_{s,u_{d}} \times P_{s+1,u_{d}} \times P_{s,u+1_{d}} \times \dots \times P_{t,v_{d}}})\end{aligned}$$

Furthermore, as the local interaction frequencies can increase dramatically as the distance between bins decreases, the ability to identify small local re-arrangements suffers. In addition, biological variation, such as cell-type specific looping events and changes in intra-TAD interaction frequency, can create false positive signals. Our experience indicates that these events are much more likely for any distances less than 1Mb in size. As a result, we do not consider any bins separated by less than 1Mb. Most rearrangements identified by our method are larger than 1Mb, so this removes a limited number of calls. Lastly, as the search space for intra-chromosomal interactions is considerably smaller than the search space for inter-chromosomal interactions, our first pass begins with a bin size of 100kb instead of 1Mb. We then proceed to smaller bin sizes in a similar manner to what was described above for inter-chromosomal interactions.

II.b.vi. Post-Processing Strand Determination

The above description of breakpoint finding makes no determination of which direction or strand the re-arrangement occurs with. However, this pattern is often readily apparent as a “peak” of signal in one of the corners of the sub-matrix. We use this pattern to identify the direction of the re-arrangement. Our experience suggests that simple peak finding or correlation can yield inaccurate strand predictions. Therefore, we estimate the strandedness using multiple different correlation metrics, and use an aggregated result for strand prediction. Specifically, we compute 12 separate spearman correlations of the interaction frequencies within the sub-matrix. The simplest of these are row-wise and column-wise correlations. For example, in column-wise correlation, we compute the average of all normalized interaction frequencies within a given column of the sub-matrix and then compute the correlation between this average and the bin position of the column. Positive correlations suggest that the distal aspect, or the “right” side of the sub-matrix, is where the re-arrangement has occurred (we term these positive stranded), while negative correlations indicate that the “left” side of the sub-matrix is involved in the re-arrangement (we term these negative stranded). The same can be applied to row wise correlations. For column and row wise correlations, we compute a spearman correlation using the average of the normalized interaction frequencies and the average of the log of the normalized interaction frequencies. This gives four of the 12 correlation coefficients we calculate.

The remaining eight correlations are computing using the diagonal of the matrix as the distance value. There are two diagonal axis in each sub-matrix, one extending from the upper-left hand to the lower right hand portion of the matrix, and one from the lower left hand to the upper right hand portion of the matrix. In this case, the distance of each bin from is equal to the sum of the distance of the row-wise and column-wise bin from the corner of the matrix. We then compute an average interaction frequency at each distance using the normalized interaction frequencies and use this for calculating the spearman correlation. Similar to what we do for column or row-wise correlation, we also compute an average of the log-transformed normalized interaction frequencies. In addition, we also express the data as binary variables, with the data represented as 1 if the signal from that bin is greater than zero and as zero otherwise. We compute the correlation for these normalized, log-transformed normalized, and binary matrices for both diagonal axis, giving us six more correlation coefficients. Finally, we also compute spearman correlation considering all values instead of for distanced averaged values. This is done for both diagonal axes using the normalized values. This yields an additional 2 correlation coefficients.

To aggregate these correlation coefficients into a final strandedness score, we sum all correlation coefficients that are informative for left/right strandedness and all correlation coefficients that are information for upper/lower strandedness. For left/right strandedness, this includes all column wise correlations and all diagonal correlations. If the sum of these values is less than zero, we consider this sub-matrix with a left-handed (negative) strand, and if it is positive we consider this as having a right-handed (positive) strand. For upper/lower strandedness, we consider all row-wise correlations and all diagonal correlations. Any correlation coefficient derived from the lower left hand to upper right hand diagonal is multipled by -1. This score is then summed, and negative values are considered upper (negative) and positive values are considered lower (positive) stranded. Our experience is that this additive approach is less sensitive to any anomaly in interaction frequency or breakpoint calls and more reliably identifies the strandedness patterns seen in the Hi-C data.

All software for calling re-arrangements in Hi-C data is available through https://github.com/dixonlab/hic_breakfinder.

II.c. Cross-method comparison and integration of structural variants

The methods that we use to identify SVs appear to have different sensitivity for detecting SVs of different sizes. Specifically, Hi-C only rarely identifies SVs smaller than 1Mb. Therefore, we perform comparisons of SVs by dividing SVs into three different categories, namely, 1) inter-chromosomal translocations identified by Hi-C, WGS and optical mapping, 2) large intra-chromosomal SVs (≥1Mb) identified by Hi-C, WGS, and optical mapping 3) and intra-chromosomal SVs < 1Mb that involves WGS and optical mapping. For the first two groups, we also included SV calls from additional methods, including karyotyping [23-31], fusion transcripts, and paired-end tag sequencing (PET-seq)[32, 33]. Data from all six methods are available only for the T47D and K562 cell lines, we hence perform the cross-six-method comparisons in these samples. For six cell lines (Caki2, A549, NCI-H460, PANC-1, LNCaP and SK-N-MC), we have data from Hi-C, WGS, optical mapping, karyotyping, and RNA-seq, therefore we perfom a five-method comparison. For MCF7 cells, we have Hi-C, PET-seq (from two separate studies), and RNA-seq data, so we compared between these three methods in MCF7 cells. Finally, we have Hi-C data and fusion transcript data for PC3, SK-N-SH, SK-N-DZ, RPMI-7951 and G401 cells lines. Finally, we have Hi-C, optical mapping, and WGS data for the karyotypically normal cell line NA12878 that we use as a non-cancer cell line control. The details of which methods were obtained from each cell line are available in Supplementary Table 1.

We converted the strand orientation for SVs detected from different methods to a unified system, in which “+” indicates the breakpoint locates at the 3’ end of the joined arm, and “-” indicates the breakpoint at the 5’ end of the joined arm. For WGS data, this dictates that SV originally classified as deletions are given the strand orientation of “+-”, inversions as “++ and - -”, duplications as “-+” and unclassified intra-chromosomal rearrangement as “++” or “- -”. Optical mapping originally reports deletions, which are assigned a strand orientation of “+-”, inversions, which are assigned as “++” or “- -”. Optical mapping also reports intra-chromosomal rearrangements >5Mb as “unclassified intra-chromosomal rearrangements” for which the software reports the strand orientation.

To determine whether the SVs detected by different methods reflect the same event, we set criteria for SV matching when comparing inter-chromosomal translocations and large intra-chromosomal SVs: 1) They have the same loci for both ends of the breakpoint. 2) They have the same strand orientation. Because the different methods have very different resolutions for SV detection, we use variable criteria for determining whether two methods identify SVs at the “same loci”. This overlap is set such that break ends within +/- 500Kb are considered as overlapping when comparing Hi-C, WGS, optical mapping, fusion transcripts and PET-seq. For karyotyping, an overlap of +/- 10Mb was set to accommodate for its low resolution. For specifically comparing deletions smaller than 1Mb, for calling to deletions as overlapping, we require that at least 50% of deletion defined by WGS must overlap with the deletion defined by optical mapping, and the size of the deletion detect by optical mapping must be within 80-120% of total length detected by WGS.

After identifying matched SVs between methods, we can resolve some unclassified SV types. Since we require SVs to have the same orientation, we can confirm certain Hi-C-detected intra-chromosomal SVs to be deletions, insertions or inversions if the same event was specified by optical mapping or WGS. Likewise, we can resolve unclassified intra-chromosomal variants from WGS to be inversions detected by optical mapping or Hi-C, and we can determine the SV type for unclassified large intra-chromosomal SVs identified by optical mapping as deletions, inversions and duplications if the orientation and SV type are determined by WGS or Hi-C. In addition, in our comparison of smaller scale of SVs, we found that insertions detected by optical mapping may be resolved as duplications in WGS, which we annotate as duplications.

We then calculated confidence levels for each SV and refine the SV coordinates based on the integration of different methods. Confidence levels are presented as the number methods by which each SV is detected. For refining the SV breakpoint coordinates, we choose loci determined by the highest resolution method for final breakpoint refinement. We consider WGS as the highest resolution method, followed by optical mapping, fusion transcripts, PET-seq, Hi-C, and then karyotyping.

II.d. Circos genome profiling

Genome profiles of cancer cell lines and GM12878 were generated using Circos [71]. Copy number is plotted according to the normalized CNV predicted by Control-freec for each 50Kb region. Duplications and deletions plotted if identified as high-confidence calls detected by at least two methods between Hi-C, WGS and optical mapping. Plotted rearrangements includes inter-chromosomal translocations, intra-chromosomal inversions and unclassified intra-chromosomal rearrangements, all of which are high-confidence calls that are identified at least twice between Hi-C, WGS, optical mapping, karyotyping, fusion transcripts, or PET-seq.

II.e. Size distribution of deletions and un-mappable translocations transitions

II.e.i. Deletions

The size of deletion detected by WGS is simply the distance between the start and end of a deletion event. The size of deletion detected by optical mapping is calculated as: $Size_{deletion} = Size_{reference} - Size_{sample} = (Reference_{end} - Reference_{start}) - (Contig_{end} - Contig_{start})$. The size of final merged deletions detected by both WGS and optical mapping was defined by the size from WGS. Then we performed Wilcoxon rank sum test to examine the difference of deletion size detected by WGS and optical mapping.

II.e.ii. Translocation un-mappable transition

In the detection of translocations, certain SVs will include a “transition” region between the two resolved portions of the rearrangement. The size of the un-mappable transition of a translocation detected by WGS is the number of basepairs that fail to align to either of the two rearranged regions. For a translocation detected by optical mapping between two chromosomes, chrA and chrB, is the distance between the closest two labels ($L_{A,} L_B$) that map to chrA and chrB respectively. There may be multiple un-mappable labels between $L_{A,} L_{B,}$ which are $L_{A+1}, L_{A+2}, L_{A+3}\ldots L_{A+M}, L_{B-N} \ldots L_{B-3}, L_{B-2}, L_{B-1}$. To provide minimum size estimation of un-mappable transitions, we assume that the DNA from the last mappable labels to their nearest un-mappable labels (L_A *to* $L_{A+1,}$ L_B *to* L_{B-1}) are all mappable. Therefore, the size of an un-mappable transition in a translocation with no or one un-mappable label will be calculated as zero basepairs. For translocations with at least two un-mappable lables, the minimal size of the unmappable transition will be $|L_{B-1}- L_{A+1}|$. If an un-mappable region is detected by in a translocation by both WGS and optical mapping, we defined the size of the un-mappable regions as the size defined by WGS.

II.f. Characterization of deletions

II.f.i. Overall disruption of genes, repeats, enhancers and insulators

We evaluated the disruption of number genes, repeats, enhancers, and insulators that were deleted by high confidence deletions. High confidence deletions are defined as those that are detected by at least two methods out of WGS, Hi-C and optical mapping from in each cell lines: A549, T47D, Caki2, K562, LnCAP, PANC-1, SK-N-MC, NCI-H460, and NA12878. The number deleted genes or repetitive elements are simply calculated by intersecting the positions of deletions with gene annotations (NCBI RefSeq) and repeat annotations (UCSC repeatMasker) in the hg38 reference genome in each cell line.

In contrast to genes and repetitive elements, enhancers and insulators can potentially have cell type specific annotations. Therefore, to identify the number of deleted enhancers in each cell line, we first match each cancer cell line with a control normal cell type from the same or similar tissue type. We use H3K27ac as a mark for enhancers and CTCF binding sites as insulators. Specifically, we use human normal mammary epithelial (HMEC) cells as a control for T47D cells, blood mononuclear cells as a control for K562 and NA12878 cells, primary pancreatic tissue as a control for PANC-1 cells, and Normal human lung fibroblasts (NHLF) as a control for NCI-H460 and A549 cells. The only exception is that we use CTCF binding sites from NA12878 to annotate insulators in K562 cells, as no CTCF is available in mononuclear cells. By intersecting high confidence deletions in cancer cell lines with enhancers or insulators in matched control cell lines or tissues, we can evaluate how many enhancers or insulators are disrupted in the cancer cell line by deletions. Further, since the overall abundance of deletions can vary in each cancer cell line, we calculate the number of lost enhancers per 100Kb of deleted genome, and then normalize this number to a constant value of 100,000 enhancers per genome.

II.f.ii. Estimates of enhancer deletion enrichment relative to random controls

To estimate whether enhancers were preferentially deleted or retain, we performed simulation by randomly distributing the high confidence deletions each cancer genome 1000 times and then examining their overlap with enhancers. The distribution of the overlap between deletions and enhancers can then be summarized and plotted. The empirical P value is calculated based on how many times the simulated number of deleted enhancer is smaller than that number in fact observed from a given cell line.

II.f.iii. Identifying polymorphic and novel deletions

High confidence deletions are stratified into two categories: known polymorphc deletions and novel variants. This is accomplished by intersecting deletions with variants reported in DGV SVs annotated as “deletion”, “loss”, and “loss and gain” using *bedtools* [75]. A detected deletion must have at least 90% reciprocal overlap between the detected deletion and deletions documented in DGV dataset to be considered as polymorphic. Some deletions reported in DGV are overlapping with each other. In such cases, if these deletions overlapped with exactly same region across the nine cell lines, these were treated as a single deletion event. Deletions that do not overlap with variants reported in DGV are defined as novel variants.

II.f.iv. Enrichment analysis of polymorphic deletions and novel deletions

To evaluate the enrichment of various genomic featurs with polymorphic or novel deletions, we first began by sorting and merging all polymorphic and novel deletions detected by both WGS and optical mapping in K562, T47D, Caki2, and GM12878 cells. The number of polymorphic and novel deletions were then counted in each cells, and the proportion of polymorphic vs. novel deletions was then compared between cancer cell lines and NA12878 cells. The overall loss of DNA content caused by polymorphic deletions or novel deletions was also calculated by summing the length of all non-redundant deletions identified in each cell. To determine if there is an enrichment of either class of deletion with genes, polymorphic and novel deletions from the nine cell lines were intersected with RefSeq genes. Genes were further annotated using the list of COSMIC-tumor related genes, considering only genes with clear annotations as oncogenes or tumor suppressors. The overlap of different classes of deletions with exons was evaluated by comparing polymorphic and novel deletions with non-redundant exons from refFlat records of GENCODE24. The overlap of different classes of deletions with repetitive elements was evaluate by comparing deletions with non-redundant repetitive elements obtained from the UCSC repeatMasker. For example, for polymorphic deletions in K562 cells containing *i* events, if the size of each deletion is DEL_i, and if the size of overlap with repeats from each deletion is Rep_i, the enrichment of repeats ($Enrich_{repeats}$)was calculated as:

$$Enrich_{repeats} = \frac{\sum Rep_i}{\sum DEL_i}$$

We also determined whether there was an enrichment for deletion of enhancers by polymorphic or novel enhancers. This was accomplished by randomly permuting deletions 1000 times in each cell type, and calculating the overlap with H3K27ac defined enhancers in the same control normal cell lines listed above. The empirical P-value was calculated based the random shuffling. The results from the two classes of deletions was then compared across each cell type to test whether the enhancer loss is preferentially associated with novel or polymorphic deletions.

II.g. Fluorescence *in situ* hybridization (FISH)

FISH probes were prepared from BAC clones targeting breakpoint proximal regions and ordered from the Children's Hospital of Oakland Research Institute (CHORI). Probes were generated using nick translation using amino-allyl-dUTP and directly labeled using amine-reactive dyes as previously described [34]. For detection of re-arrangements, cells were grown in the presence of nocodazole for 4 hours to induce mitotic arrest. Cells were then fixed in 3:1 methanol acetic acid and metaphase spreads were prepared as previously described [35]. FISH was performed as previously described using 20ng of each labeled probe [36]. Slides were visualized using Zeiss Axioimager Z1 microscope.

Clone Name	Genomic Coordinates (GRCh38)
RP4-591B8	chr1:114317729-114460280
RP11-549D23	chr6:136684046-136869512
RP11-552O4	chr18:26699051-26869173
RP5-1184F4	chr20:32435227-32554099
RP11-510J16	chr16:81921455-82098760
RP1-259A10	chr6:17166672-17317191
RP11-136C6	chr6:39145400-39247240
RP11-548O1	chr3:138878614-138950820

II.h. Breakpoint PCR

PCR across predicted breakpoints was performed using the Qiagen Long-Range PCR kit. PCR products amplified from K562 template were cloned into TOPO-XL cloning vectors and sequenced using conventional Sanger sequencing. In the event that the breakpoint did not fall within the Sanger sequenced regions, primers were re-designed and the process was repeated.

Cell	SV type	Name	Sequence
K562	Translocation	K_chr9_22_F	AAAGAGCCTTTTGTTGGCTATGTTGTT
K562	Translocation	K_chr9_22_R	CAGAAGGAAGAGCTATGCTTGTTAGGG
K562	Translocation	K_chr3_10_F	CTGCCATAAAGAGTTCACAAACACACC
K562	Translocation	K_chr3_10_R	CTGAGACCTGGAAAACAGAGCAAGAC
K562	Translocation	K_chr5_6_F	AGCAATTTTAGAGGCACTTCTCCTTGT
K562	Translocation	K_chr5_6_R	AGGCATTTGGGATCTTGCTGGATTATG

K562	Translocation	K_chr9_13_F	TTGAGATGTCTGTTTCATTTCCCGACT
K562	Translocation	K_chr9_13_R	GAACCACTGCTCCTGGACTTCATCTT
T47D	Translocation	T_[chr6_chr22]_F	CACATAACCAAGGGAGAGTT
T47D	Translocation	T_[chr6_chr22]_R	GTGAGGTGAATTCAAATGTT
T47D	Translocation	T_chr4]_chr5]_F	TTGCACACCGGCTCCATGAG
T47D	Translocation	T_chr4]_chr5]_R	GATCTCTACTTAATCTGCAT
T47D	Translocation	T_9]_[15_F	TAAAAGATAAAGGCATCTGT
T47D	Translocation	T_9]_[15_R	ACCAACCAAAAAAAGCCCAG
T47D	Translocation	T_5]_[5_F	CTTCCCGTCTAAGCAGACCT
T47D	Translocation	T_5]_[5_R	CTTTCATCATGTTAGTCATG
T47D	Translocation	T_9]_[9_F	GGTTTGGGCATTCTATTTTC
T47D	Translocation	T_9]_[9_R	GCCTTCAGAAAGTTCTCAGT
T47D	Translocation	T_chr10]_[chr10_F	ATATAAATGCGATGCTTTTTCCT
T47D	Translocation	T_chr10]_[chr10_R	GAGTTGTTTTGAGTTCCTTGGAG
T47D	Translocation	T_chr10]_[chr3_F	GCAAAGTTCTTCTTAAGAATGT
T47D	Translocation	T_chr10]_[chr3_R	ACAGATTAATTGACTCCCTTC
T47D	Translocation	T_chr3]_[chr9_F	GTGCTAGGATTACAGGAATGAGC
T47D	Translocation	T_chr3]_[chr9_R	GGAAACCCTTGTACACTATTGGT
Caki2	Translocation	C_chr12]_[chr4_F	TTCCCTTTAAAAGCACAATGCCC
Caki2	Translocation	C_chr12]_[chr4_R	ATTTCCTATAATTGGGTTTTCCT
Caki2	Translocation	C_chr9]_[chr19_F	AGTCAGTCTTGTACCTTGGGATG
Caki2	Translocation	C_chr9]_[chr19_R	AGAAAGCTTCCAGTCACAAAACT
Caki2	Translocation	C_ [chr6_[chr8_F	GGTATGGAGATGATCAACCCAAG
Caki2	Translocation	C_ [chr6_[chr8_R	TTGACAAAAGAATAAACAAATAGAT
T47D	Deletion	T_chr2_212590110_F	GTGGGATAAACAAGTGACTAACC
T47D	Deletion	T_chr2_212720073_R	ACCACGAAGCCACCAGAAGGAAG
T47D	Deletion	T_chr2_97188517_F	AATTAACTCCTAAAATGGTAATT
T47D	Deletion	T_chr2_97190465_R	ATCAATGTGGATATGCCGAGTGA
T47D	Deletion	T_chr14_104948976_F	GCATCTGCAGCTTGGGCAGGTGC
T47D	Deletion	T_chr14_104951429_R	AAAGTGGACCTCAAGGGCCCCCA
T47D	Deletion	T_chr3_58586154_F	TTTCCTGAATAGAAAAGAAACAC
T47D	Deletion	T_chr3_58586217_R	CAATCCTCACGTCATTCTTTTTA
T47D	Deletion	T_chr4_165081464_F	CCACCTAGGAACCTCCCACTCTT
T47D	Deletion	T_chr4_165083902_R	GAAAAAAACATGACTGGGCGCGG
T47D	Deletion	T_chrX_42652746_F	CCACTGCAAAAACATGCCAA
T47D	Deletion	T_chrX_42656304_R	AGTTTTCAAAGGGAATGCTT
T47D	Deletion	T_chr2_28466613_F	AATTATAAAAGTATCATGGG
T47D	Deletion	T_chr2_28469693_R	CCAGGCAAATCAGAGGTGTC
T47D	Deletion	T_chr7_6861596_F	CTTTACTGGTGTTGGACTCG
T47D	Deletion	T_chr7_6887316_R	ATTAAAGCAGTTGGATTTTT
T47D	Deletion	T_chr1_207523594_F	AAAAGCAATAGGACAAAGGC
T47D	Deletion	T_chr1_207546536_R	GCTCATCTCCTTTCAAGTCT

T47D	Deletion	T_chr12_58325913_F	TGAGTTCCCTTAGTATTTAT
T47D	Deletion	T_chr12_58339245_R	ATAGGTGGGGATTATGGGAG
T47D	Deletion	T_chr11_107361838_F	GAAGCCTCAGGAGCTGATGA
T47D	Deletion	T_chr11_107374676_R	GTCACCAATCTTGTCTTCCT
T47D	Deletion	T_chr7_97762466_F	ACTGGATCCCTTCCTTACAG
T47D	Deletion	T_chr7_97773481_R	GGCAAGCTGCTGAATTGCCT
T47D	Deletion	T_chr7_70969523_F	TGAGCCAATTAAACCTCTAT
T47D	Deletion	T_chr7_70979773_R	GTATTCATGCTTCAAAGAAG
T47D	Deletion	T_chr6_85998091_F	TGCAGTGTTTGGTTTTCTAT
T47D	Deletion	T_chr6_86007304_R	AAAAAGTGGGCAAAGGATAT
T47D	Deletion	T_chr1_53126296_F	GGACTACAGGTGCCCACCAT
T47D	Deletion	T_chr1_53129986_R	CCAGTGGTGGCTTCATCTGT
T47D	Deletion	T_chr13_69400712_F	CTACAGAAAGACTGAATAGC
T47D	Deletion	T_chr13_69404714_R	ATTATATTTGGGGAATCTAC

III. Public datasets used in this study

Cell Type	Data Type	Accession	Source
A549	Hi-C	ENCSR444WCZ	ENCODE
B-ALL	Hi-C	GSM1906333	GEO
Caki2	Hi-C	ENCSR401TBQ	ENCODE
G401	Hi-C	ENCSR079VIJ	ENCODE
K562	Hi-C	GSM1551618	GEO
K562	Hi-C	GSM1551619	GEO
K562	Hi-C	GSM1551622	GEO
KBM7	Hi-C	GSM1551624	GEO
KBM7	Hi-C	GSM1551625	GEO
KBM7	Hi-C	GSM1551626	GEO
KBM7	Hi-C	GSM1551627	GEO
KBM7	Hi-C	GSM1551628	GEO
LNCaP	Hi-C	ENCSR346DCU	ENCODE
MCF7	Hi-C	GSM1631185	GEO
MCF7	Hi-C	GSM1942100	GEO
MCF7	Hi-C	GSM1942101	GEO
MHH-CALL-4	Hi-C	GSM1906334	GEO
NCI-H460	Hi-C	ENCSR489OCU	ENCODE
Panc1	Hi-C	ENCSR440CTR	ENCODE
PC3	Hi-C	GSM1902605	GEO
PC3	Hi-C	GSM1902606	GEO
RL	Hi-C	GSM1906332	GEO
RPMI-7951	Hi-C	ENCSR862OGI	ENCODE

SJCRH30	Hi-C	ENCSR998ZSP	ENCODE
SK-MEL-5	Hi-C	ENCSR312KHQ	ENCODE
SK-N-DZ	Hi-C	ENCSR105KFX	ENCODE
SK-N-MC	Hi-C	ENCSR834DXR	ENCODE
SK-N-SH	Hi-C	GSM1826481	GEO
SK-N-SH	Hi-C	GSM1826482	GEO
T47D	Hi-C	ENCSR549MGQ	ENCODE
H1 hESC	Hi-C	GSM1267196	GEO
Mes	Hi-C	GSM1267199	GEO
MSC	Hi-C	GSM1267200	GEO
NPC	Hi-C	GSM1267202	GEO
Troph	Hi-C	GSM1267205	GEO
HMEC	Hi-C	GSM1551608	GEO
HMEC	Hi-C	GSM1551609	GEO
HMEC	Hi-C	GSM1551610	GEO
HMEC	Hi-C	GSM1551611	GEO
HMEC	Hi-C	GSM1551612	GEO
HUVEC	Hi-C	GSM1551630	GEO
IMR90	Hi-C	GSM1551602	GEO
GM12878	Hi-C	GSM1551597	GEO
AA86	Hi-C	GSM2176962	GEO
GB176	Hi-C	GSM2176966	GEO
GB180	Hi-C	GSM2176967	GEO
GB182	Hi-C	GSM2176968	GEO
GB183	Hi-C	GSM2176969	GEO
GB238	Hi-C	GSM2176970	GEO
A549	RNA-seq	ENCFF000EJJ	ENCODE
A549	RNA-seq	ENCFF000EJV	ENCODE
Caki2	RNA-seq	ENCFF185BLE	ENCODE
Caki2	RNA-seq	ENCFF859XNV	ENCODE
Caki2	RNA-seq	ENCFF917KJE	ENCODE
Caki2	RNA-seq	ENCFF272LRD	ENCODE
G401	RNA-seq	ENCFF757UTO	ENCODE
G401	RNA-seq	ENCFF780REB	ENCODE
HMEC	RNA-seq	ENCFF000GDZ	ENCODE
HMEC	RNA-seq	ENCFF000GEO	ENCODE
HMEC	RNA-seq	ENCFF000GDA	ENCODE
HMEC	RNA-seq	ENCFF000GDT	ENCODE
K562	RNA-seq	ENCFF001RFE	ENCODE
K562	RNA-seq	ENCFF001RFF	ENCODE
LNCaP	RNA-seq	ERR361060_1	ENA
LNCaP	RNA-seq	ERR361060_2	ENA

MCF7	RNA-seq	ENCFF002DKR	ENCODE
MCF7	RNA-seq	ENCFF002DKU	ENCODE
NCI-H460	RNA-seq	ENCLB297PER_1	ENCODE
NCI-H460	RNA-seq	ENCLB297PER_2	ENCODE
NCI-H460	RNA-seq	ENCLB794DVD_1	ENCODE
NCI-H460	RNA-seq	ENCLB794DVD_2	ENCODE
PANC-1	RNA-seq	SRR1736496_1	SRA
PANC-1	RNA-seq	SRR1736496_2	SRA
PC-3	RNA-seq	ENCFF186UTV	ENCODE
PC-3	RNA-seq	ENCFF884SCR	ENCODE
Primary Kidney	RNA-seq	SRR2087309_1	SRA
Primary Kidney	RNA-seq	SRR2087309_2	SRA
Primary Kidney	RNA-seq	SRR2087322_1	SRA
Primary Kidney	RNA-seq	SRR2087322_2	SRA
T47D	RNA-seq	SRR5808857_1	SRA
T47D	RNA-seq	SRR5808857_2	SRA
T47D	RNA-seq	SRR925736_1	SRA
T47D	RNA-seq	SRR925736_2	SRA
SK-N-MC	RNA-seq	SRR1594020_1	SRA
SK-N-MC	RNA-seq	SRR1594020_2	SRA
RPMI-7951	RNA-seq	ENCFF002DLX	ENCODE
RPMI-7951	RNA-seq	ENCFF002DLY	ENCODE
SK-N-DZ	RNA-seq	ENCFF482SFO	ENCODE
SK-N-DZ	RNA-seq	ENCFF691TRA	ENCODE
SK-N-SH	RNA-seq	ENCFF000IMC	ENCODE
SK-N-SH	RNA-seq	ENCFF000IMS	ENCODE
HG00268-FIN-F	WGS	SRR1293236_1	SRA
HG00268-FIN-F	WGS	SRR1293236_2	SRA
HG00096-GBR-M	WGS	SRR1291026_1	SRA
HG00096-GBR-M	WGS	SRR1291026_2	SRA
HG00419-CHS-F	WGS	SRR1295433_1	SRA
HG00419-CHS-F	WGS	SRR1295433_2	SRA
NA12878	WGS	ERR194147_1	ENA
NA12878	WGS	ERR194147_2	ENA
NA19625-AA-F	WGS	SRR1295538_1	SRA
NA19625-AA-F	WGS	SRR1295538_2	SRA
LNCaP	WGS	SRR1977632_1	SRA
LNCaP	WGS	SRR1977632_2	SRA

IV. New deposit of dataset to SRA under project PRJNA380394:

Cell lines	Data type
T47D	WGS
Caki2	WGS
K562	WGS
A549	WGS
PANC-1	WGS
SK-N-MC	WGS
NCI-H460	WGS
K562	Hi-C
SK-N-AS	Hi-C

V. References

1. Kumar-Sinha, C., S.A. Tomlins, and A.M. Chinnaiyan, *Recurrent gene fusions in prostate cancer.* Nat Rev Cancer, 2008. **8**(7): p. 497-511.
2. Maher, C.A., et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009. **458**(7234): p. 97-101.
3. Tomlins, S.A., et al., *Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer.* Nature, 2007. **448**(7153): p. 595-9.
4. Cai, C., et al., *ETV1 is a novel androgen receptor-regulated gene that mediates prostate cancer cell invasion.* Mol Endocrinol, 2007. **21**(8): p. 1835-46.
5. Hollenhorst, P.C., et al., *The ETS gene ETV4 is required for anchorage-independent growth and a cell proliferation gene expression program in PC3 prostate cells.* Genes Cancer, 2011. **1**(10): p. 1044-1052.
6. Harenza, J.L., et al., *Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines.* Sci Data, 2017. **4**: p. 170033.
7. Kryh, H., et al., *Comprehensive SNP array study of frequently used neuroblastoma cell lines; copy neutral loss of heterozygosity is common in the cell lines but uncommon in primary tumors.* BMC Genomics, 2011. **12**: p. 443.
8. Peifer, M., et al., *Telomerase activation by genomic rearrangements in high-risk neuroblastoma.* Nature, 2015. **526**(7575): p. 700-4.
9. Valentijn, L.J., et al., *TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors.* Nat Genet, 2015. **47**(12): p. 1411-4.
10. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic cancer.* Nature, 2015. **518**(7540): p. 495-501.
11. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**(7605): p. 47-54.
12. Holland, D.G., et al., *ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium.* EMBO Mol Med, 2011. **3**(3): p. 167-80.

13. Wang, Z., et al., *The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types.* PLoS One, 2013. **8**(3): p. e58793.
14. Lindsley, R.C. and A.S. LaCasce, *Biology of double-hit B-cell lymphomas.* Curr Opin Hematol, 2012. **19**(4): p. 299-304.
15. Hayward, N.K., et al., *Whole-genome landscapes of major melanoma subtypes.* Nature, 2017. **545**(7653): p. 175-180.
16. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
17. Klijn, C., et al., *A comprehensive transcriptional portrait of human cancer cell lines.* Nat Biotechnol, 2015. **33**(3): p. 306-12.
18. Pendleton, M., et al., *Assembly and diploid architecture of an individual human genome via single-molecule technologies.* Nat Methods, 2015. **12**(8): p. 780-6.
19. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.
20. Imakaev, M., et al., *Iterative correction of Hi-C data reveals hallmarks of chromosome organization.* Nat Methods, 2012. **9**(10): p. 999-1003.
21. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
22. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre.* Nature, 2012. **485**(7398): p. 381-5.
23. Peng, K.J., et al., *Characterization of two human lung adenocarcinoma cell lines by reciprocal chromosome painting.* Dongwuxue Yanjiu, 2010. **31**(2): p. 113-21.
24. Struski, S., et al., *Identification of chromosomal loci associated with non-P-glycoprotein-mediated multidrug resistance to topoisomerase II inhibitor in lung adenocarcinoma cell line by comparative genomic hybridization.* Genes Chromosomes Cancer, 2001. **30**(2): p. 136-42.
25. Strefford, J.C., et al., *A combination of molecular cytogenetic analyses reveals complex genetic alterations in conventional renal cell carcinoma.* Cancer Genet Cytogenet, 2005. **159**(1): p. 1-9.
26. Naumann, S., et al., *Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization.* Leuk Res, 2001. **25**(4): p. 313-22.
27. Beheshti, B., et al., *Identification of a high frequency of chromosomal rearrangements in the centromeric regions of prostate cancer cell lines by sequential giemsa banding and spectral karyotyping.* Mol Diagn, 2000. **5**(1): p. 23-32.
28. Liu, J., et al., *Modeling of lung cancer by an orthotopically growing H460SM variant cell line reveals novel candidate genes for systemic metastasis.* Oncogene, 2004. **23**(37): p. 6316-24.

29. Espino, P.S., et al., *Genomic instability and histone H3 phosphorylation induction by the Ras-mitogen activated protein kinase pathway in pancreatic cancer cells.* Int J Cancer, 2009. **124**(3): p. 562-7.
30. Sirivatanauksorn, V., et al., *Non-random chromosomal rearrangements in pancreatic cancer cell lines identified by spectral karyotyping.* Int J Cancer, 2001. **91**(3): p. 350-8.
31. Rondon-Lagos, M., et al., *Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis.* Mol Cytogenet, 2014. **7**(1): p. 8.
32. Hillmer, A.M., et al., *Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes.* Genome Res, 2011. **21**(5): p. 665-75.
33. Hampton, O.A., et al., *Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines.* Cancer Genet, 2011. **204**(8): p. 447-57.
34. Bolland, D.J., et al., *Robust 3D DNA FISH using directly labeled probes.* J Vis Exp, 2013(78).
35. Bangs, C.D. and T.A. Donlon, *Metaphase chromosome preparation from cultured peripheral blood cells.* Curr Protoc Hum Genet, 2005. **Chapter 4**: p. Unit 4 1.
36. Knoll, J.H. and P. Lichter, *In situ hybridization to metaphase chromosomes and interphase nuclei.* Curr Protoc Hum Genet, 2005. **Chapter 4**: p. Unit 4 3.