CANCER

# EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps

Xiaotao Wang[1], Yu Luan[1], Feng Yue[1,2]*

**The Hi-C technique has been shown to be a promising method to detect structural variations (SVs) in human genomes. However, algorithms that can use Hi-C data for a full-range SV detection have been severely lacking. Current methods can only identify interchromosomal translocations and long-range intrachromosomal SVs (>1 Mb) at less-than-optimal resolution. Therefore, we develop EagleC, a framework that combines deep-learning and ensemble-learning strategies to predict a full range of SVs at high resolution. We show that EagleC can uniquely capture a set of fusion genes that are missed by whole-genome sequencing or nanopore. Furthermore, EagleC also effectively captures SVs in other chromatin interaction platforms, such as HiChIP, Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET), and capture Hi-C. We apply EagleC in more than 100 cancer cell lines and primary tumors and identify a valuable set of high-quality SVs. Last, we demonstrate that EagleC can be applied to single-cell Hi-C and used to study the SV heterogeneity in primary tumors.**

## INTRODUCTION

Structural variations (SVs), including deletions, inversions, duplications, and translocations, can directly contribute to tumorigenesis and other diseases through multiple mechanisms. SVs can lead to the deletion of tumor suppressor genes or duplication of proto-oncogenes (*1*) or promote the formation of oncogenic fusion genes (*2*). More recently, it has been shown that SVs can bring distal enhancers to the proximity of proto-oncogenes and cause the up-regulation of oncogenic gene expression through a mechanism termed enhancer hijacking (*3*, *4*). The discovery of recurrent SVs has greatly advanced our knowledge about tumorigenesis and led to effective targeted therapy (*5*).

Despite their importance, genome-wide detection of SVs remains a challenging problem. Traditionally, karyotyping has been the major method to detect various genetic disorders in the clinic; however, it is an inherently low-throughput and low-resolution method (*6*). Microarray has been used to identify gains and losses of genetic materials, but it has limitations in detecting copy number neutral events such as inversions and balanced translocations (*7*). More recently, short-read whole-genome sequencing (WGS) has been widely used to identify a variety of genomic variations due to their high resolution, high throughput, and simplicity (*8*–*13*). However, because of the mappability issue of short reads, it is difficult to detect SVs at repetitive regions using WGS (*11*). The advent of long-read sequencing such as PacBio and Nanopore has partly alleviated the mappability issue (*14*, *15*). However, these technologies have a relatively high sequencing error rate and also need deep sequencing for SV detection (>20×) (*16*).

Recently, we and other groups showed that Hi-C, a technique that was originally proposed to study three-dimensional (3D) genomic architectures, can also be used for systematic SV detection with as little as 1× genome coverage (*11*, *17*–*20*). As SVs induce de novo chromatin interactions across the breakpoints, when Hi-C reads are mapped to the reference genome, different types of SVs are characterized by aberrant interaction blocks with different orientations. Identifying SVs is essentially the same as identifying and annotating such blocks on a Hi-C map. Compared to WGS and nanopore that require direct breakpoint spanning reads to detect SVs, such property of Hi-C substantially decreases the sequencing depths that are needed for SV detection and also gives Hi-C higher chances to detect SVs at repetitive regions, as long as the adjacent regions of breakpoints are mappable. So far, three methods have been proposed to predict SVs with Hi-C data. The Hi-C breakfinder that we codeveloped is the first algorithm of this kind, where we use an iterative approach to search for abnormal interaction blocks with significantly higher interaction frequencies compared with a background model (*18*). HiCtrans identifies translocation breakpoints by searching for signal changepoints on interchromosomal contact matrices of each chromosome pair (*17*). More recently, Wang *et al.* (*19*) proposed a new method called HiNT-TL for translocation detection, which is based on the identification of regions with both unusually high interaction frequencies and uneven distribution of interaction strengths.

However, all current methods have their limitations. HiCtrans and HiNT-TL cannot predict intrachromosomal SVs, which usually accounts for a large portion of all SVs in a cancer genome (*13*). Although Hi-C breakfinder can identify interchromosomal translocations, it can only detect large intrachromosomal SVs with a size >1 Mb (Table 1). The challenge for short-range SV detection is that Hi-C maps typically contain features such as topologically associating domains (TADs) and chromatin loops (*18*), which are usually less than 1 Mb, and such patterns make the accurate detection of SV challenging. Furthermore, all three methods still have less-than-optimal resolution. Therefore, we develop EagleC, a framework that combines deep-learning and ensemble-learning strategies to predict a full range of SVs at high resolution. We show that EagleC outperforms existing methods in both precision and recall rates. Furthermore, we demonstrate that EagleC can be used as a general framework to predict SVs in many other 3C-based platforms, such as HiChIP, ChIA-PET, capture Hi-C, and even single-cell Hi-C (scHi-C). With the pretrained

[1] Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine Northwestern University, Chicago, IL, USA. [2] Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL, USA.
*Corresponding author. Email: yue@northwestern.edu

**Table 1. Comparing methods for detecting SVs.**

| | Interchromosomal translocation | Intrachromosomal SVs (>1 Mb) | Intrachromosomal SVs (<1 Mb) | Applicable to other 3C-based techniques (ChIA-PET, HiChIP, capture Hi-C, and scHi-C) | Support nonhuman genome | Gene fusions |
|---|---|---|---|---|---|---|
| HiCtrans | √ | | | | √ | |
| HiNT-TL | √ | | | | | |
| Hi-C breakfinder | √ | √ | | | | |
| EagleC | √ | √ | √ | √ | √ | √ |

models, we predicted SVs in over 100 cancer cell lines or primary tumors. Pan-cancer analysis of these datasets showed that the location and formation of SVs are closely associated with 3D chromatin architectures.

## RESULTS

### Overview of the EagleC framework

Identifying SVs from a Hi-C map is essentially a multilabel image classification problem in machine learning. There are multiple types of SVs, and each type is characterized by a unique pattern on a Hi-C contact map (figs. S1 to S4) (*18*). For example, we draw three consecutive fragments A, B, and C in fig. S1. A deletion of fragment B will result in the junction of the 3′-end of fragment A and the 5′-end of fragment C (left of fig. S1A). Because of the spatial proximity, there will be strong chromatin interaction signals between the 3′-end of fragment A and the 5′-end of fragment C. However, when we map the Hi-C reads of the sample to the reference genome, we will see an abnormal increase in the interactions between fragment A and fragment C (middle of fig. S1A). As a result, in the submatrix centered at the breakpoints, there will be an increase in interactions in the upper-right quadrant. Similarly, tandem duplication sequentially links the original DNA fragment and the duplicated fragment, resulting in strong interactions in the lower-left quadrant of the submatrix (fig. S1B); inverted duplication causes aberrant signals either in the upper-left or lower-right quadrants (fig. S1, C and D), depending on the direction of the inverted DNA fragment. For inversions (fig. S1E) and reciprocal translocations (fig. S2), de novo interactions are formed on the opposite sides of the breakpoints, resulting in a "butterfly shape" on the Hi-C map. In our framework, an SV with the "+−" label corresponds to the fusion of the 3′-end of a fragment to the 5′-end of another fragment, while the "++" label corresponds to the 3′-to-3′ fusion, "−+" corresponds to the 5′-to-3′ fusion, and "−−" corresponds to the 5′-to-5′ fusion.

Figure 1A describes the overall design of the EagleC framework. The positive training samples are defined as the Hi-C contact matrices surrounding a set of high-confidence SVs, which were detected by both WGS and optical mapping in eight cancer cell lines (A549, Caki2, K562, LNCaP, NCI-H460, PANC-1, SK-N-MC, and T47D) (*18*). We found that the original samples demonstrate severely imbalanced class distributions (Materials and Methods and table S1). To avoid the model biased toward any specific classes during the training, we proposed a data augmentation algorithm based on Poisson distributions to make sure each class has a similar number of samples (Materials and Methods). Furthermore, to make the model able to distinguish real SV signals from false-positive signals induced by normal 3D genomic features, we sampled similar numbers of intrachromosomal and interchromosomal submatrices from the Hi-C map of a normal cell line GM12878 (*21*) and labeled them as "intranegative" and "internegative," respectively. These negative samples include matrices surrounding random pixels, chromatin loops, and the transition points of A/B compartments. We also included matrices from the cancer Hi-C data that are located in an SV block but not overlapping with the breakpoint as an additional negative dataset.

Because the strong diagonal Hi-C signals can confound the detection of short-range SVs, in the preprocessing steps, EagleC corrects distance effects for intrachromosomal matrices by using the distance-averaged signals (fig. S3, A and B). To alleviate potential data noise, each input matrix is then convolved with a 2D Gaussian filter followed by min-max scaling.

The inputs to the convolutional neural network (CNN) are 21 × 21 grayscale images, which go through two convolutional layers, each followed by a max-pooling layer. The probabilities of each label (++, +−, −+, −−, intranegative, and internegative) are calculated from two fully connected layers using the sigmoid activation. Before the output layer, we insert a dropout layer with a dropout probability of 0.5 to avoid overfitting.

One important component of the EagleC framework is that it performs an iterative learning procedure to gradually improve the model specificity. After each round of training, the model is used to perform a genome-wide prediction in GM12878 Hi-C. As GM12878 is a karyotypically normal cell line, all the predictions will be considered as false positives and randomly selected as additional negative samples in the next round of training. Such processes are repeated until the convergence is observed.

To further optimize the sensitivity and specificity of the framework, we perform an ensemble learning procedure. In total, 50 models are independently trained using the same iterative approach described above, with each model randomly initialized with different set of training samples. When predicting SVs in a novel sample, the final probability scores are determined as the average across all 50 models, and a pixel will be reported as an SV breakpoint if the probability of at least one positive label (++, +−, −+, and −−) is greater than a predefined cutoff (Materials and Methods).

We trained a series of EagleC models optimized for various sequencing depths using down-sampled versions of the training samples (Materials and Methods). To investigate the performance of EagleC, we predicted SVs (unless noted, all SVs reported in this study are at the 5-kb resolution) in other cancer Hi-C datasets that were not used in the training procedure (table S2). EagleC successfully predicted different types of SVs, including short-range SVs with breakpoint distance

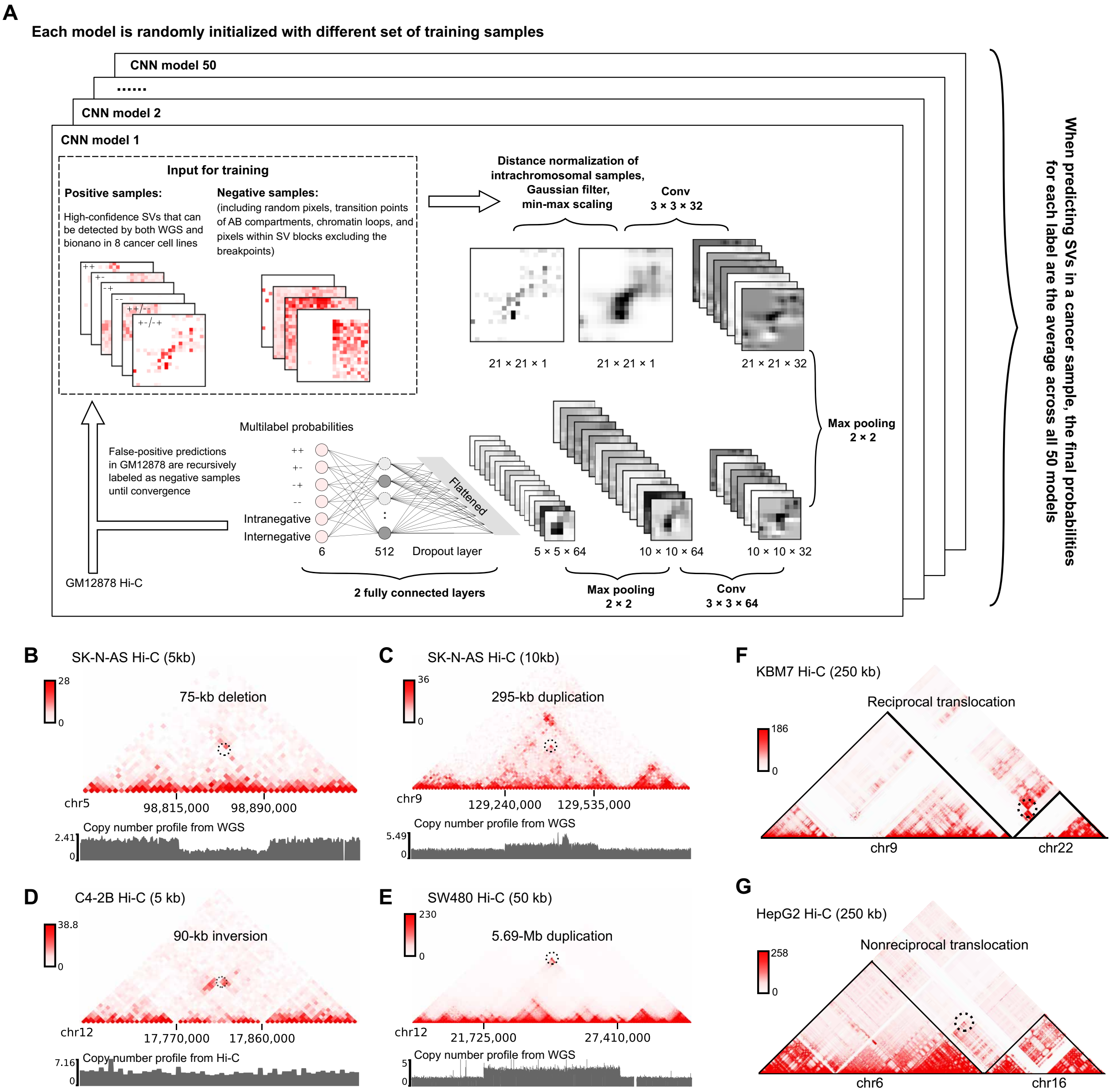
A
Each model is randomly initialized with different set of training samples
CNN model 50
......
CNN model 2
CNN model 1
Input for training
Positive samples:
High-confidence SVs that can be detected by both WGS and bionano in 8 cancer cell lines
Negative samples:
(including random pixels, transition points of AB compartments, chromatin loops, and pixels within SV blocks excluding the breakpoints)
++
+-
-+
--
++/--
+-/-+
Distance normalization of intrachromosomal samples, Gaussian filter, min-max scaling
Conv 3 × 3 × 32
21 × 21 × 1
21 × 21 × 1
21 × 21 × 32
Max pooling 2 × 2
Multilabel probabilities
False-positive predictions in GM12878 are recursively labeled as negative samples until convergence
Intranegative
Internegative
6
512
Dropout layer
Flattened
5 × 5 × 64
10 × 10 × 64
10 × 10 × 32
GM12878 Hi-C
2 fully connected layers
Max pooling 2 × 2
Conv 3 × 3 × 64
When predicting SVs in a cancer sample, the final probabilities for each label are the average across all 50 models
B
SK-N-AS Hi-C (5kb)
28
0
75-kb deletion
chr5
98,815,000
98,890,000
Copy number profile from WGS
2.41
0
C
SK-N-AS Hi-C (10kb)
36
0
295-kb duplication
chr9
129,240,000
129,535,000
Copy number profile from WGS
5.49
0
F
KBM7 Hi-C (250 kb)
Reciprocal translocation
186
0
chr9
chr22
D
C4-2B Hi-C (5 kb)
38.8
0
90-kb inversion
chr12
17,770,000
17,860,000
Copy number profile from Hi-C
7.16
0
E
SW480 Hi-C (50 kb)
230
0
5.69-Mb duplication
chr12
21,725,000
27,410,000
Copy number profile from WGS
5
0
G
HepG2 Hi-C (250 kb)
Nonreciprocal translocation
258
0
chr6
chr16

**Fig. 1. EagleC predicts a full range of high-resolution SVs from chromatin interaction data.** (**A**) Workflow of the EagleC framework. (**B** to **G**) Examples showing different types of SVs predicted by EagleC. The black dashed circle indicates the SV breakpoint position in each case. The resolution of each Hi-C map is labeled within the parentheses. (B) A short-range (75 kb) heterozygous deletion predicted in the SK-N-AS cells. (C) A short-range (295 kb) duplication predicted in the SK-N-AS cells. (D) A short-range (90 kb) inversion predicted in the C4-2B cells. (E) A long-range (5.69 Mb) duplication predicted in the SW480 cells. (F) A reciprocal translocation predicted in the KBM7 cells. (G) A nonreciprocal translocation predicted in the HepG2 cells.

less than 1 Mb or even 100 kb (Fig. 1, B to D), large intrachromosomal SVs (Fig. 1E), reciprocal interchromosomal translocations (Fig. 1F), and nonreciprocal interchromosomal translocations (Fig. 1G).

## EagleC outperforms existing methods in detecting SVs on Hi-C maps

We first visually inspected the predictions and found that nearly all blocks with abnormally high interaction frequencies were predicted as SVs, suggesting high sensitivity of the framework (Fig. 2A). We then examined closely individual loci and compared the predictions from EagleC and Hi-C breakfinder (*18*). In many cases, although EagleC and Hi-C breakfinder predicted the same SV blocks, the exact coordinates of the predicted breakpoints were different, and the EagleC-predicted breakpoints were more likely to be validated by WGS (Fig. 2A, regions "A," "C," "D," and "E"). Further, EagleC predicted more precise breakpoints at the 5-kb resolution than Hi-C breakfinder predictions, which are usually 100-kb resolution (block with a dashed line in regions "B" and "D" in Fig. 2A).

Next, we systematically evaluated the performance of EagleC by comparing it with all existing methods, HiCtrans (*17*), Hi-C breakfinder

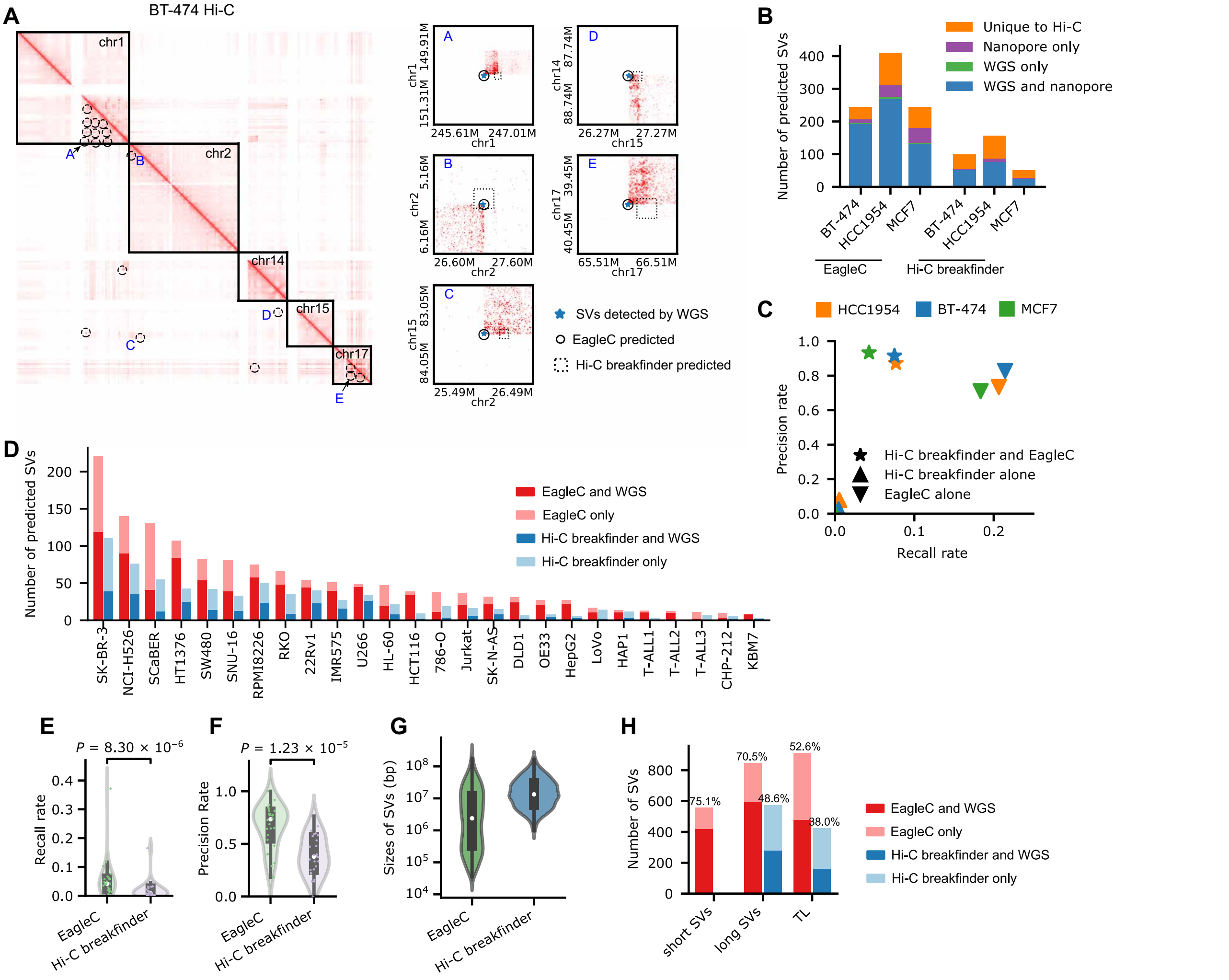

**Fig. 2. EagleC outperforms existing methods in both precision and recall.** (**A**) (Left) Genome-wide Hi-C map and the predicted SVs (black circles) in BT-474. (Right) Enlarged Hi-C maps of the indicated SV regions. The SV breakpoints detected by different methods are highlighted using different marks on each map. (**B**) Number of SVs predicted by EagleC and Hi-C breakfinder in BT-474, HCC1954, and MCF7 cells with levels of validation by orthogonal methods. For Hi-C breakfinder, we only counted SVs reported at the 10-kb resolution. (**C**) Precision and recall rates of SVs that can only be predicted by EagleC, SVs that can only be predicted by Hi-C breakfinder, and SVs that can be predicted by both methods. (**D**) The number of SVs predicted by EagleC and Hi-C breakfinder in additional 26 cancer samples with both Hi-C and WGS data available. (**E** and **F**) Recall rates and precision rates of SVs predicted by EagleC and Hi-C breakfinder in 26 cancer samples. Each dot represents an individual sample. The *P* values were computed using the two-sided Wilcoxon signed-rank test. (**G**) Size distributions of intrachromosomal SVs detected by EagleC and Hi-C breakfinder. Data were merged from 29 cancer samples. (**H**) Number of different range of SVs predicted by EagleC and Hi-C breakfinder with validation ratios by WGS. Short SVs, intrachromosomal SVs with breakpoint distance less than 1 Mb; long SVs, intrachromosomal SVs with breakpoint distance greater than 1 Mb; TL, interchromosomal translocations.

(*18*), and HiNT-TL (*19*) (Table 1). We used three breast cancer cell lines BT-474, HCC1954, and MCF7 as the benchmark datasets, as there are Hi-C, WGS, and nanopore data available in the same cell lines. Because HiCtrans and HiNT-TL can only detect interchromosomal translocations, we first focused on interchromosomal translocations alone. The reported results by each method differed greatly (fig. S5, A and B). First of all, different methods predicted a different number of translocation candidates. For example, in the MCF7 cell line, EagleC and Hi-C breakfinder predicted 154 and 116 translocations, respectively. HiCtrans reported the largest number of translocations ($n = 520$), while HiNT-TL detected the smallest number of translocations ($n = 28$). In terms of the resolutions at which the translocations were reported, EagleC predicted translocations at the highest resolution among all methods at 5 kb. The translocations reported by HiCtrans were at 10 or 20 kb; translocations reported by Hi-C breakfinder were at a mixture of 10-kb, 100-kb, and 1-Mb resolutions; and nearly all translocations reported by HiNT-TL were at the 100-kb resolution (fig. S5A). To further investigate the performance of each method, we compared the translocation predictions from each method with a reference translocation set defined by WGS and nanopore for each cell line (Materials and Methods). As shown in fig. S5B, EagleC outperforms all the other methods with both higher precision rates and higher recall rates in all three cell lines. Specifically, although HiCtrans detected three times as many interchromosomal translocations as EagleC, it recalled fewer validated SVs due to its redundant false-positive predictions within a single SV block (fig. S5C).

### More in-depth analysis between EagleC and Hi-C breakfinder that includes both inter- and intrachromosomal SVs

Then, we performed more in-depth comparisons between EagleC and Hi-C breakfinder, as they are currently the only methods that can identify intrachromosomal SVs. Notably, EagleC detected 2.4-fold (244 versus 100), 2.6-fold (410 versus 157), and 4.8-fold (244 versus 51) as many SVs (including interchromosomal translocations and intrachromosomal SVs) as Hi-C breakfinder in BT-474, HCC1954, and MCF7, respectively (Fig. 2B). At the same time, EagleC achieved notably higher precision rates than Hi-C breakfinder in these cell lines. When allowing 20-kb mismatches for either side of the breakpoints, 84.8, 76.3, and 73.8% of SVs predicted by EagleC in BT-474, HCC1954, and MCF7 can be validated by either WGS or nanopore, while corresponding rates for Hi-C breakfinder are only 55.0, 55.4, and 54.9% (Fig. 2B and fig. S6, A to C). When we increased the allowed mismatch from 20 to 100 kb, the validation rates for EagleC nearly stayed the same, while the rates for Hi-C breakfinder increased by 11.0, 10.8, and 7.8% in the three cell lines, respectively, which suggests that Hi-C breakfinder failed to predict the exact breakpoint positions within the SV block for a portion of SVs (Fig. 2A and figs. S6, A to C). In BT-474, 24.2% (59 of 244) of the EagleC-predicted SVs matched 59.0% (59 of 100) of the Hi-C breakfinder predictions. Of the 185 SVs that are unique to EagleC, 83.2% (154 of 185) can be validated by either WGS or nanopore, compared with 2.4% (1 of 41) for Hi-C breakfinder unique SVs (Fig. 2C). Similarly, in HCC1954 and MCF7, 73.4 (232 of 316) and 71.0% (152 of 214) of EagleC-unique SVs can be validated, compared with 7.9 and 0.0% for SVs that are specific to Hi-C breakfinder. On average, EagleC-unique SVs have a 21.9-fold higher precision rate and a 61.0-fold higher recall rate than Hi-C breakfinder unique SVs in these three cell lines (Fig. 2C).

Furthermore, we evaluated the performance of EagleC and Hi-C breakfinder at various sequencing depths by down-sampling the original BT-474 and HCC1954 Hi-C data to nine different depths (ranging from 5 to 175 million contact pairs) (fig. S6D). Notably, EagleC achieved obviously higher precision and recall rates than Hi-C breakfinder at all sequencing depths. In addition, while the recall rates for Hi-C breakfinder reached a plateau at the depth with around 75 million contact pairs, the rates for EagleC kept increasing along with higher sequencing depths, which suggests that the power of Hi-C in SV detection might have been underestimated by previous studies. To evaluate the impact of tumor heterogeneity on SV prediction, we simulated a series of Hi-C datasets by mixing the BT-474/HCC1954 Hi-C with HMEC (human mammary epithelial cells, a normal breast cell line) Hi-C at various fractions while keeping the total sequencing depth at around 200 million contact pairs. Similarly, we observed that EagleC predicted much more SVs with higher accuracy than Hi-C breakfinder at all tumor heterogeneity levels (fig. S6E).

We next extended the analysis to 26 additional cancer cell lines or patient samples with both Hi-C and WGS data available (table S2). Again, we observed that compared with Hi-C breakfinder, EagleC achieved significantly higher recall rates and precision rates in all the 26 cancer samples (Fig. 2, D to F, and fig. S6F). Because of the inherent limitations of the algorithm, Hi-C breakfinder can only detect large intrachromosomal SVs greater than 1 Mb. However, as shown in Fig. 2G, 39.5% of intrachromosomal SVs predicted by EagleC are short-range SVs, with a minimum size of 35 kb. To our surprise, although SVs at this range have been thought hard to be distinguished from other Hi-C contact patterns, they were predicted with even higher accuracy than long-range SVs and translocations (Fig. 2H).

### EagleC detects novel fusion genes in cancer

Because we noticed that a sizable portion of EagleC-predicted SVs were missed by both short-read WGS and nanopore (Fig. 2B), we investigated whether such detection can be supported by other evidence. As the RNA sequencing (RNA-seq) data are available for these three cell lines, we predicted fusion genes with the Arriba software (*22*). As shown in Fig. 3A, EagleC detected breakpoints inside the *ATXN7* and *BCAS3* genes in MCF7, while the arriba software also predicted the fusion of these two genes (Fig. 3A, right). We showed two more such examples in Fig. 3 (B and C), demonstrating that because of the high-resolution nature of EagleC, it can uniquely predict fusion genes that are missed by WGS and nanopore. We also wanted to point out that the sequencing depth of Hi-C data in these three cell lines is much lower (BT-474, 17×; HCC1954, 11×; and MCF7, 16×) than the WGS (BT-474, 44×; HCC1954, 38×; and MCF7, 38×) and nanopore (BT-474, 31×; HCC1954, 49×; and MCF7, 26×), suggesting that Hi-C can detect a unique set of SVs even with low sequencing depths. Last, we noticed that genes involved in these fusion events were significantly overexpressed in cancer cells, compared with their expression levels in nonmalignant cell lines without the fusion (Fig. 3D).

### EagleC can accurately predict SVs using other 3C-based techniques

In addition to Hi-C, there are several other 3C-derived techniques. Among them, pulldown-based 3C assays, including Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) (*23*), HiChIP (*24*), Proximity Ligation-Assisted ChIP-seq (PLAC-Seq) (*25*), and Capture Hi-C (*26*), are gaining more and more interests because of their efficiency in detecting genome-wide chromatin interactions mediated by a protein or a set of genes of interest. However, the potential of these techniques in SV detection has never been explored by previous studies (Table 1). We hypothesized that the rules we learned for predicting SVs on Hi-C maps are common among all 3C-based platforms. To validate this hypothesis, we focused on the breast cancer cell line MCF7, in which there are WGS, nanopore, Hi-C, CTCF ChIA-PET, and Pol2 ChIA-PET data available (tables S2 and S3). We directly applied the EagleC models trained on Hi-C data to CTCF ChIA-PET and Pol2 ChIA-PET. Overall, EagleC predicted a similar number of SVs in Hi-C, CTCF ChIA-PET, and Pol2 ChIA-PET, and there is a large overlap between the three datasets (Fig. 4, A and B). For instance, EagleC predicted 226 SVs in CTCF ChIA-PET, 66.4% of which were predicted in Hi-C as well. Similarly, 62.8% (123 of 196) of SVs predicted in Pol2 ChIA-PET matched 50.4% (123 of 244) of predictions from Hi-C. We found that EagleC achieved comparable precision rates in both ChIA-PET datasets (CTCF ChIA-PET, 65.5%; and Pol2 ChIA-PET, 68.2%) compared to Hi-C (73.8%) (Fig. 4C). Moreover, we observed that EagleC-predicted SVs have significantly higher recall rates and precision rates than Hi-C breakfinder in all the 10 HiChIP/ChIA-PET datasets with matched WGS data (Fig. 4, D to F, and table S3).

To investigate whether EagleC models are also transferable to other 3C-based platforms, we collected nine capture Hi-C datasets in mice, with each dataset containing one and only one known SV: (i) a series of duplications ranging from 420 kb to 1.74 Mb that were originally used to study the impact of duplications on TAD structures (fig. S7A) (*27*); (ii) a 115-kb inversion in mouse forelimb at

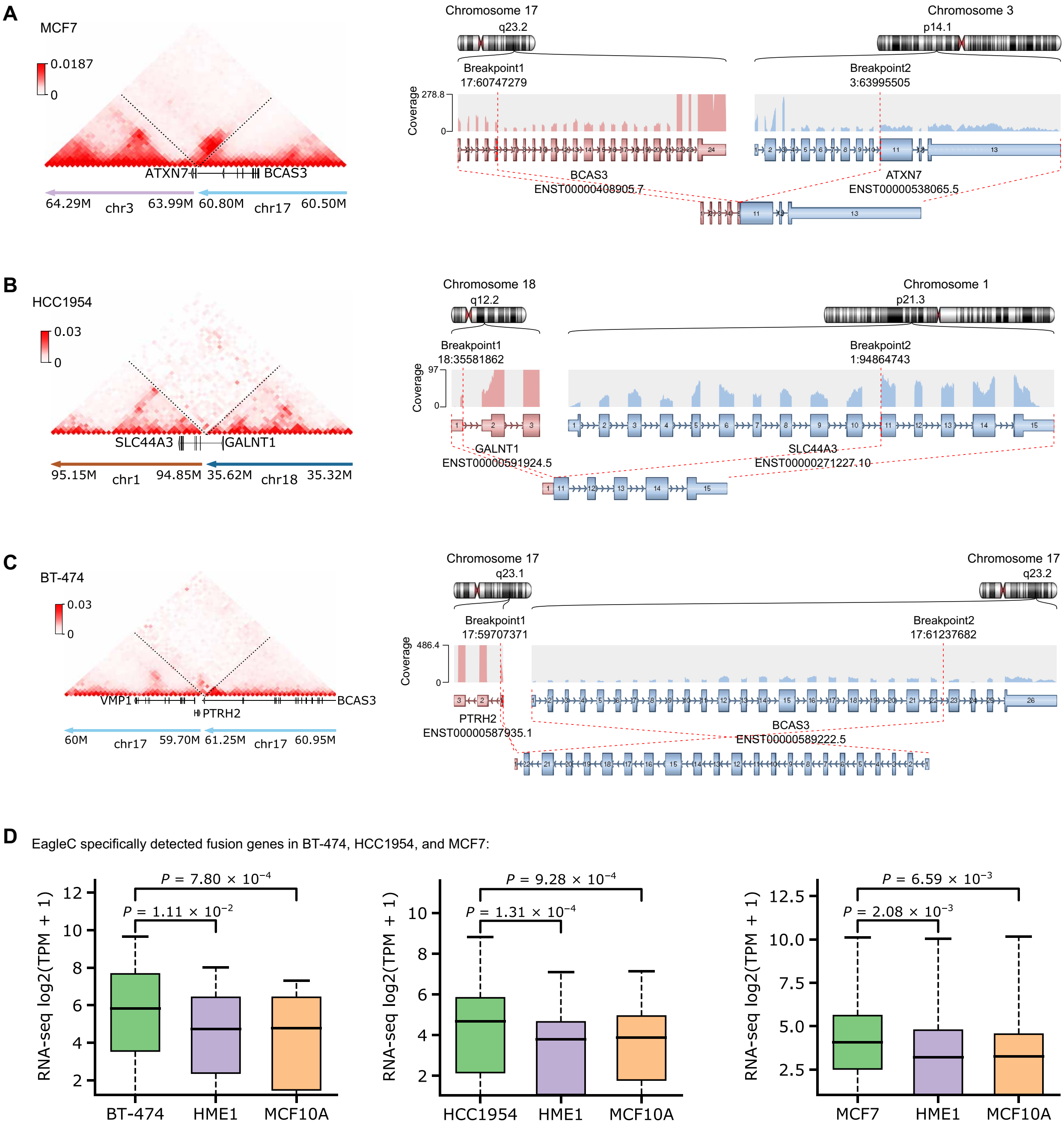

**Fig. 3. Fusion genes uniquely detected by EagleC are overexpressed in cancer cells.** In this analysis, we only included fusion events that were detected by EagleC but missed by both WGS and nanopore. (**A** to **C**) Examples of novel gene fusions detected by EagleC (left) and the supporting evidence from RNA-seq data (right). The fusion gene partners, their orientations, and the retained exons in the fusion transcripts were detected and plotted using the arriba software. (**D**) Normalized expression signals for novel fusion genes detected by EagleC in different breast cancer cell lines versus two nonmalignant breast cell lines (HME1 and MCF10A). TPM, transcripts per kilobase million. The *P* values were computed using the two-sided Wilcoxon signed-rank test. In each boxplot, the center line indicates the median, the box limits represent the upper and lower quartiles, and the box whiskers indicate the 1.5× interquartile range.

embryonic day 11.5 (E11.5) (fig. S7B) (*28*); (iii) a 1.14-Mb inversion in mouse limb buds at E12.5 (fig. S7C) (*29*); and (iv) a series of inversions ranging from 620 kb to 1.10 Mb, which have an invariable downstream breakpoint and a variable upstream breakpoint (fig. S7D) (*30*). We found that EagleC was able to predict the known SVs in all these datasets. No other pixels were predicted as SVs, suggesting both high sensitivity and high specificity of EagleC in predicting SVs on capture Hi-C maps.

## Detection of SVs in 105 cancer samples

After we have validated our framework in various 3C-based platforms, we applied the trained models to 91 Hi-C datasets and 25 HiChIP/ChIA-PET datasets from 105 cancer cell lines or primary tumors (tables S2 and S3). If multiple datasets are available in the same sample, we combined their results to achieve a more comprehensive set of SV annotations. In total, we predicted 5620 SVs across all the samples,

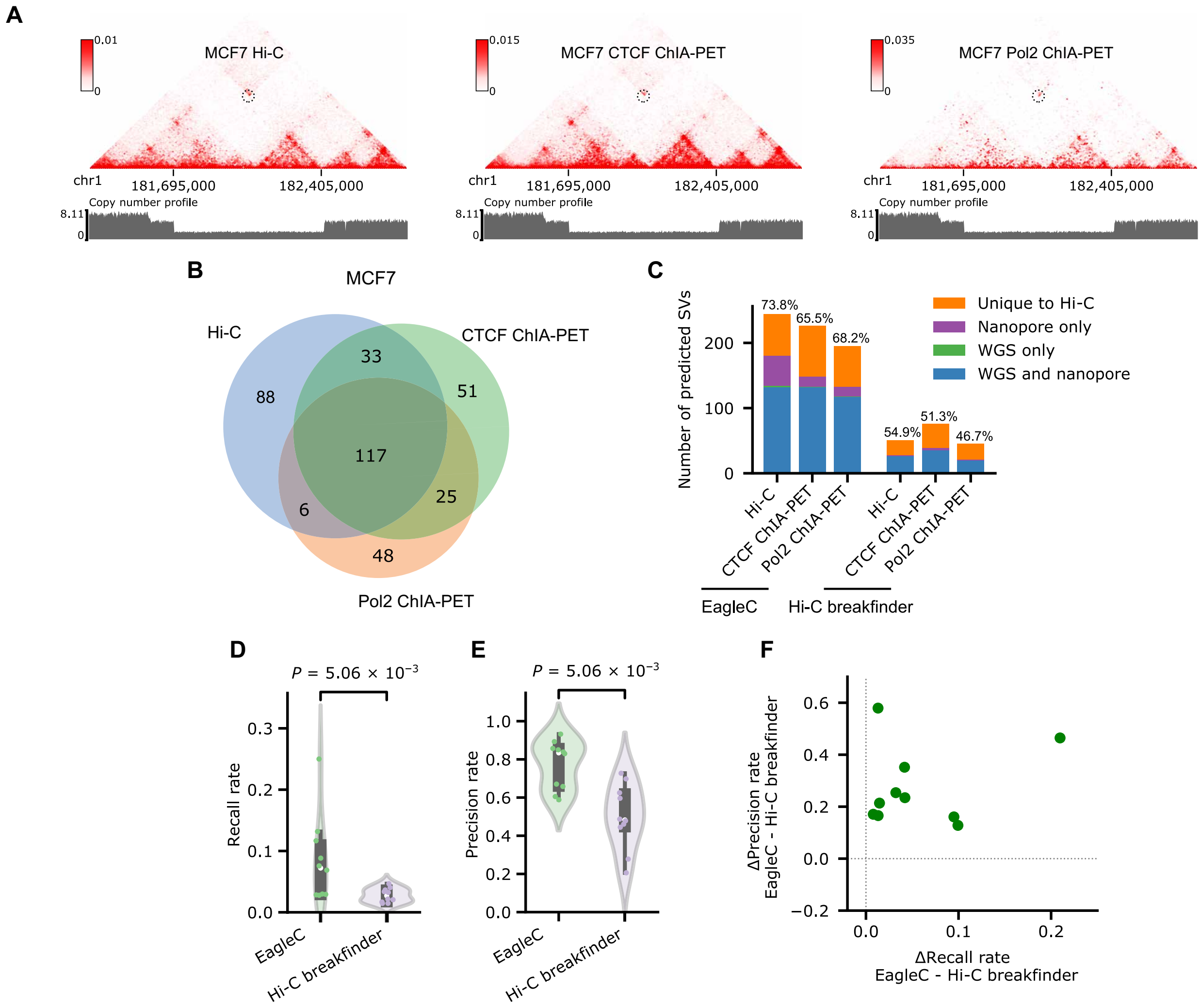

**Fig. 4. EagleC accurately predicts SVs on HiChIP and ChIA-PET contact maps.** (**A**) A heterozygous deletion (chr1, 181,695,000 to 182,405,000) on chromosome 1 is correctly predicted by EagleC on Hi-C, CTCF ChIA-PET, and Pol2 ChIA-PET contact maps of MCF7 cells. The same copy number profile calculated from WGS is shown below each contact map. (**B**) Venn diagram of SVs predicted by EagleC on different contact maps of MCF7 cells. (**C**) Number of SVs predicted by Hi-C breakfinder and EagleC on different contact maps of MCF7 cells with levels of validation by orthogonal methods. The ratio of SVs that can be validated by orthogonal methods is indicated above each bar. (**D** and **E**) Performance of EagleC and Hi-C breakfinder on 10 HiChIP/ChIA-PET contact maps in cancer samples. Each dot represents an individual sample. The *P* values were computed using the two-sided Wilcoxon signed-rank test. (**F**) Differences of precision rates and recall rates between EagleC-predicted and Hi-C breakfinder–predicted SVs on 10 HiChIP/ChIA-PET datasets in cancer.

with the number in each sample ranging from 2 to 410 (Fig. 5A and table S4). The highest numbers of SVs are observed in breast cancer cell lines, consistent with previous findings that breast cancer cells frequently contain genomic instability driven chromosomal variations (*31*). Combining data from all samples, 30.9% of the predicted SVs are short-range SVs (<1 Mb), 35.7% are long-range SVs, and 33.4% are interchromosomal translocations.

Next, we investigated how 3D genome architectures can influence the location and formation of SVs. As genomic variations such as SVs and copy number variations (CNVs) can confound the interpretation of contact maps in cancer, we computed 3D genome features including A/B compartments and TADs for different cancer types using Hi-C data in normal cells/tissues with similar cell of origin (table S5). It has been widely known that the genome can be partitioned into two compartments, with the A compartment associated with open chromatin, and the B compartment associated with closed chromatin, and chromatin interactions within the same compartments (A-A/B-B) are stronger than interactions between different compartments (A-B) (*32*). We hypothesized that the preexisting chromatin interactions between distal compartments would increase the probability of SV formation between these compartments. To this end, we quantified the proportions of SVs that occurred within the same compartments (A-A/B-B) and between different compartments (A-B). As a control, we randomly shuffled the SV breakpoints in the mappable genome regions 1000 times for each cancer sample, controlling for the ratio of interchromosomal versus intrachromosomal SVs and the sizes of the intrachromosomal SVs. Compared with random controls, SVs are preferentially formed between A-A

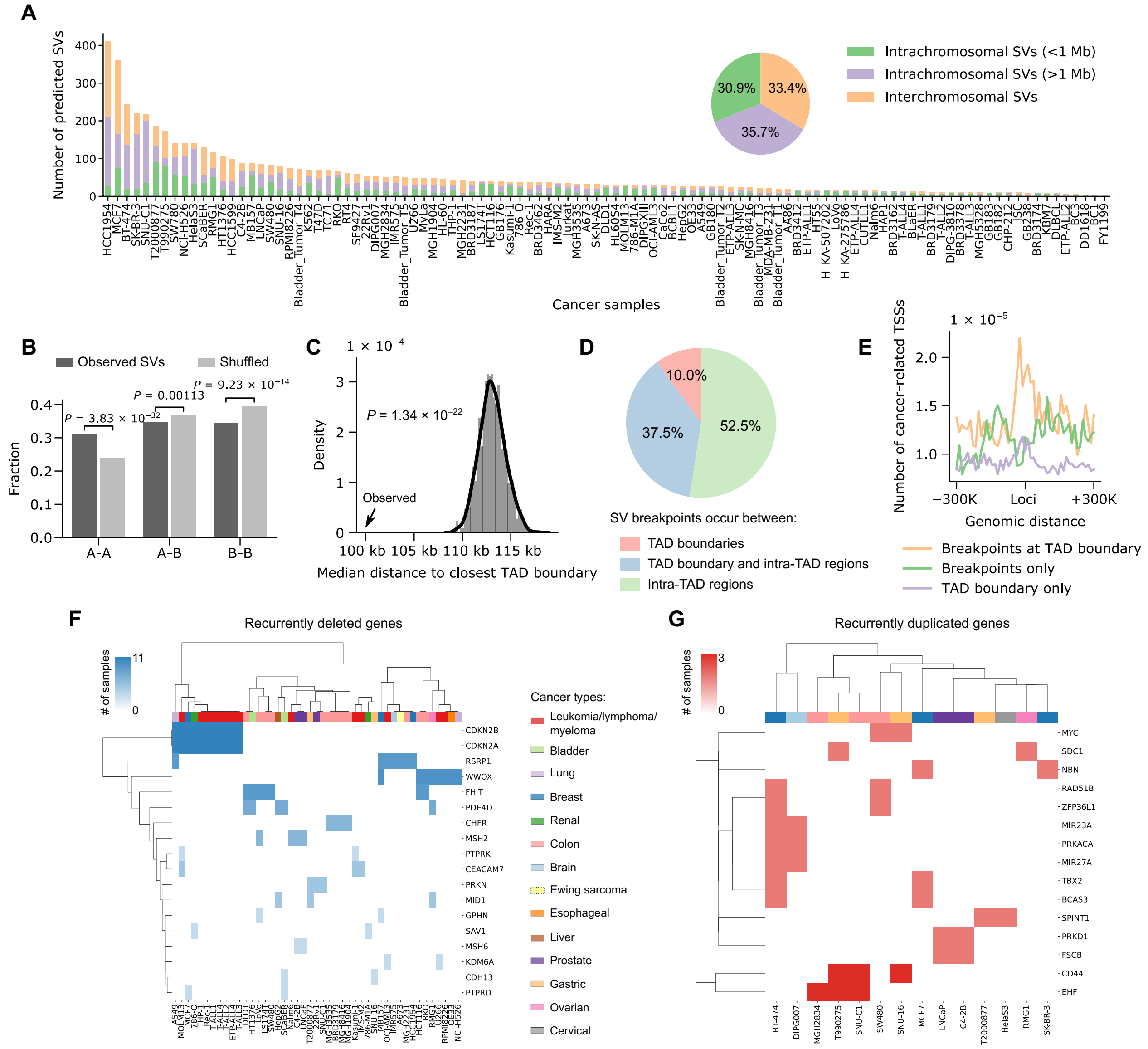

**Fig. 5. Pan-cancer analysis of SVs in 105 cancer cell lines or patient samples.** (**A**) Number of short-range intrachromosomal SVs, long-range intrachromosomal SVs, and interchromosomal translocations predicted in each sample. The pie chart shows the percentages of different SV categories based on the data combined from all samples. (**B**) SVs are significantly enriched in A-to-A and depleted in B-to-B compartments compared to randomly shuffled controls. The *P* values were calculated using two-sided *Z* test. (**C**) SV breakpoints occur significantly closer to TAD boundaries compared to random shuffled controls. The *P* values were calculated using two-sided *Z* test. (**D**) Percentages of SVs with breakpoints occurring between TAD boundaries, between a TAD boundary and intra-TAD regions, and between intra-TAD regions. (**E**) The TSSs of cancer-related genes are specifically enriched near SV breakpoints at TAD boundaries. (**F**) Unsupervised clustering of recurrently deleted genes and cancer samples based on the occurrence of genes in each sample. Different cancer types are coded by different colors. (**G**) Similar to (F), the clustering of recurrently duplicated genes and cancer samples.

compartments rather than B-B or A-B compartments (Fig. 5B), and such patterns are largely conserved for different cancer types and different ranges of SVs (figs. S8 and S9).

At the megabase scale, it has been shown that mammalian genomes are organized into TADs (*33*). TAD boundaries, which are enriched for CTCF binding sites, provide an insulated environment for proper gene regulation. In comparison to an expected distribution derived from randomly shuffled SVs, we found that SV breakpoints are located significantly closer to TAD boundaries, consistent with previous findings that DNA topoisomerase II beta (TOP2B)-mediated DNA double-strand breaks are enriched at anchors of chromatin loops (Fig. 5C and figs. S8 and S9) (*34*). Overall, around 10% of SVs are formed between TAD boundaries, 37.5% are formed between a TAD boundary and an intra-TAD region, and 52.5% are formed between intra-TAD regions (Fig. 5D and figs. S8 and S9). Moreover, we found that transcription start sites (TSSs) of cancer-related genes are specifically enriched at breakpoint-associated TAD boundaries (Fig. 5E), suggesting that the disruption of TAD boundaries by genomic rearrangements might be an important mechanism for oncogene dysregulation and tumorigenesis.

To further explore the value of our SV annotations, we identified genes that are recurrently affected by short-range SVs in different samples. As expected, we found that the majority of deleted genes are tumor suppressor genes (Fig. 5F), such as *CDKN2A/2B* (*35*), *WWOX* (*36*), *CHFR* (*37*), and *MSH2* (*38*) genes. On the other hand, a lot of genes within the duplicated regions are oncogenes (Fig. 5G), such as *MYC*, which has been reported to be associated with cell proliferation in multiple cancer types (*39*), and the *CD44* gene, which is a common biomarker of cancer stem cells and encodes a cell-surface glycoprotein involved in tumor initiation and progression (*40*).

## EagleC predicts known interchromosomal translocations in single cells

To make EagleC work for scHi-C with limited contact information per cell, we down-sampled contact maps of the same eight cancer cell lines and GM12878 cells to comparable sequencing depths, and retrained the models at the 500-kb resolution (Materials and Methods). Then, we tested EagleC on published scHi-C datasets in HAP1 and K562 (*41*), both of which are chronic myeloid leukemia cell lines. HAP1 cells contain a reciprocal translocation between chromosome 9 and chromosome 22 (*42*), while K562 cells contain a nonreciprocal translocation between chromosome 9 and chromosome 22 (*43*). The HAP1 dataset contains 256 single cells, with a median of 18,793 contacts per cell, while the K562 dataset contains 337 cells, with a median of merely 3974 contacts per cell (Fig. 6A and fig. S10A). Notably, we found that even with these extremely sparse contact matrices, EagleC was able to predict the known chr9-chr22 translocations in single cells (Fig. 6, B and C, and fig. S10, B and C).

To systematically investigate the lower limit of contact number for accurately predicting SVs in single cells, we ranked all the 256 HAP1 cells by their sequencing depths and generated a series of contact matrices (contact pairs ranging from 148,635 to 4.05 million) by pooling up to 99 deepest single cells (Fig. 6D). As expected, the number of predicted SVs decreases along with the increasing number of cells (Fig. 6E). By using SVs predicted from the merged Hi-C map as the gold standard SV set, we found that the $F_1$ scores increased with the cell number and reached a plateau of 1 when the cell number reached 25 (1.68 million contact pairs) (Fig. 6F). For K562 cells, we

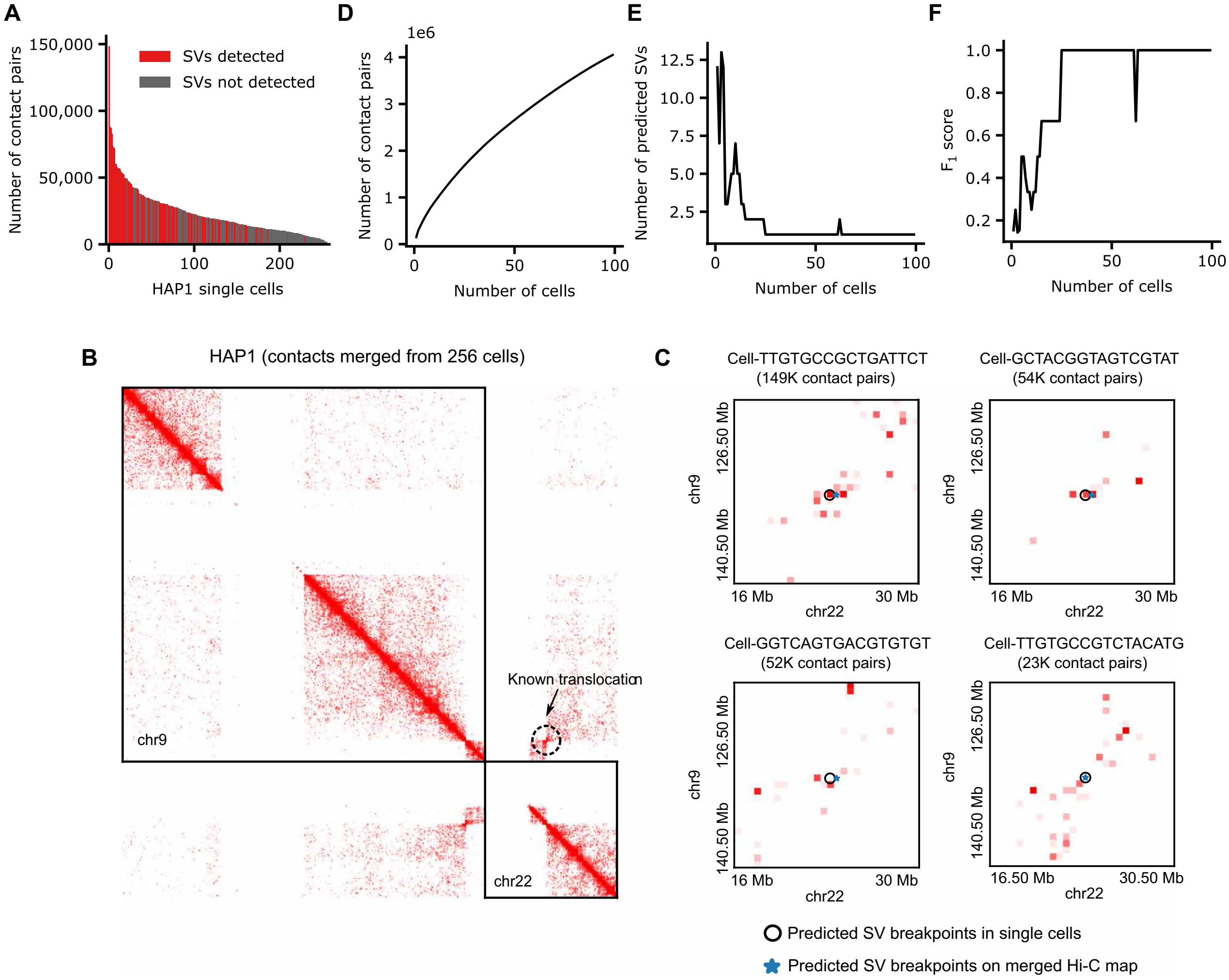

**Fig. 6. EagleC predicts interchromosomal translocations in HAP1 scHi-C.** (**A**) Number of Hi-C contact pairs per cell. Red bars and gray bars represent cells with and without SV detected, respectively. (**B**) There is a known reciprocal translocation between chr9 and chr22 in the HAP1 cell line. (**C**) EagleC accurately predicts breakpoint coordinates of the known translocation in single cells. The cell ID and corresponding number of contacts are indicated in the title of each contact map. (**D** to **F**) The number of contact pairs (D), the number of predicted SVs (E), and the $F_1$ scores of the predicted SVs (F) in a series of pooled single-cell maps.

performed a similar analysis by pooling up to 300 deepest K562 single cells, but this time, we counted interchromosomal translocations and intrachromosomal SVs separately (fig. S10, D to H). Again, by using SVs predicted from the merged Hi-C map of all 337 K562 single cells as the gold standard, we observed that the $F_1$ scores for both intrachromosomal SVs and interchromosomal translocations increased with the cell number. However, predicting intrachromosomal SVs needed a higher number of usable reads from more cells to achieve a reasonable performance (fig. S10F).

In conclusion, EagleC can identify both interchromosomal translocations and intrachromosomal SVs in scHi-C data. However, because of insufficient usable reads per cell, pooling contacts from multiple cells can help achieve the most accurate predictions at current stages.

## DISCUSSION

Although several methods have been developed to detect SVs using Hi-C data, the power of Hi-C in detecting short-range SVs with breakpoint distance less than 1 Mb has not been achieved, mainly due to the challenge of distinguishing SV signals from other chromatin interaction signals within this range. Here, by taking the advantage of CNNs in image recognition and ensemble learning in avoiding the overfitting problem, we developed EagleC to fill this important gap. For individual models, we applied an iterative training approach to gradually improve their specificity by incorporating negative samples from a normal cell line. We showed that EagleC not only predicted unique short-range SVs but also greatly improved the overall prediction power over existing methods. We demonstrated the feasibility of using Hi-C to detect fusion genes, some of which were missed by both WGS and nanopore. Although our current framework cannot achieve the base-pair resolution, we observed that Hi-C has unique ability in detecting fusion points within introns compared with RNA-seq (Fig. 3, A and B). Moreover, EagleC can serve as a general model to predict SVs using other 3C-based contact maps including ChIA-PET, HiChIP/PLAC-Seq, capture Hi-C, and even scHi-C. With unique properties of different platforms in enriching different set of chromatin interactions, we envision that the application of EagleC in these platforms will boost SV-related discoveries, such as enhancer hijacking (*4*). Furthermore, by applying EagleC to 116 Hi-C/HiChIP/ChIA-PET datasets, we predicted SVs in 105 cancer samples and found the distributions of SVs on the genome are closely associated with 3D chromatin architectures.

We note that existing methods such as Hi-C breakfinder (*18*) and HiNT-TL (*19*) can only be applied to human samples (Table 1), as they rely on the identification of interaction blocks that deviate from the expected interaction frequencies, which were only precalculated for human genomes. In comparison, the contact patterns learned by EagleC are genome agnostic and can be used to predict SVs or judge the accuracy of genome assemblies in any species (fig. S7) (*44*). Because the data we collected in this study had various sequencing depths and quality, we limited our analyses at the 5-kb resolution and predicted SVs with a minimum size of 35 kb. However, our framework should be able to predict SVs at higher resolutions (1 kb) when sequencing depths are sufficient.

Recent progress in single-cell sequencing techniques has enabled the studies of molecular changes and evolutionary trajectories during cancer development. In this course, multiple algorithms have been developed for identifying single-nucleotide variants (SNVs) and CNVs in single cells (*45*–*47*). However, predicting SVs at the single-cell level is still relatively unexplored. Here, by applying EagleC to scHi-C datasets in cancer cell lines (*41*), we demonstrated that the lower limit of contact number for EagleC to accurately predict SVs is between 1 and 2 million usable reads (Fig. 6F and fig. S10F). It has been shown that several biotin-free scHi-C protocols, such as Dip-C (*48*), can achieve such level of sequencing depths per cell. On the other hand, single-nucleus methyl-3C sequencing, another method without biotin pulldown of ligation junctions, can simultaneously measure chromatin contacts and DNA methylation levels in the same cells (*49*). Combining these technologies and EagleC in primary samples will enable the study of SV heterogeneity and potentially identify SVs that are critical for cancer studies.

## MATERIALS AND METHODS

### Hi-C data processing

For Hi-C/HiChIP/ChIA-PET datasets, if the data had been mapped to hg38 and processed into contact matrices at multiple resolutions, we directly downloaded and used the processed contact matrices in our study; if only raw sequencing data were available, we processed the data using the runHiC Python package (https://pypi.org/project/runHiC/), which is based on the 4DN Hi-C data processing pipeline; otherwise, if the data were originally mapped to hg19 and raw sequencing data were not available, then we converted the coordinates to hg38 using pairLiftOver (https://pypi.org/project/pairLiftOver/; see description below). For Hi-C datasets, we used the CNV-normalized matrices calculated by our recently developed toolkit NeoLoopFinder (*4*) as input for EagleC. For HiChIP and ChIA-PET datasets, we used the iterative correction and eigenvector decomposition (ICE)-normalized matrices as input for EagleC (*50*).

For other Hi-C–based SV detection methods, we installed and ran Hi-C breakfinder following https://github.com/dixonlab/hic_breakfinder. We installed and ran HiNT-TL (v2.2.8) following the official guidelines at https://github.com/parklab/HiNT. We downloaded, installed, and ran HiCtrans (hictrans.v3.R) following https://github.com/ay-lab/HiCtrans.

### pairLiftOver

To facilitate the processing of Hi-C/HiChIP data that were mapped to a different reference genome (hg19) and did not have raw sequencing data available, we developed a command line tool called pairLiftOver to convert the 2D genomic coordinates of chromatin contacts between assemblies. pairLiftOver is based on the UCSC chain files (https://genome.ucsc.edu/goldenPath/help/chain.html), which describes pairwise alignment between two assemblies. The input to pairLiftOver can be two kinds of pairs files: (i) the pairs format defined by 4DN DCIC (https://github.com/4dn-dcic) and (ii) allValidPairs defined by HiC-Pro (https://nservant.github.io/HiC-Pro/RESULTS.html). Both formats define contact pairs in plain text, with each row representing 2D coordinates of a single pair. pairLiftOver iterates each row of a pairs file and converts the coordinates of both sides using the pyliftover package (https://github.com/konstantint/pyliftover). A pair is retained only if both sides can be uniquely mapped to the target genome. For each row, only columns pertaining to genomic coordinates (columns 2 to 5 for 4DN pairs; columns 2 and 3 and columns 5 and 6 for allValidPairs) are converted and all other columns remain unchanged. The input pairs file can be plain text file, gzip/bgzip compressed file (.gz), or lz4 compressed file (.lz4). By default, pairLiftOver will output a sorted pairs file in the standard

4DN pairs format (https://github.com/4dn-dcic/pairix), containing seven columns: "readID," "chr1," "pos1," "chr2," "pos2," "strand1," and "strand2." However, users can also choose to output a matrix file in ".mcool" (*50*) or ".hic" (*51*) format by setting the parameter "--output-format."

### scHi-C data processing

The FASTQ files of the scHi-C data were downloaded from GSE84920. The cellular demultiplexing was performed by following the pipeline described in the original paper (*41*), and cells with a low number of reads were filtered out. Then, the demultiplexed reads were aligned to hg19 using BWA-MEM (*52*) with the parameter "-SP5M." The BAM files were then parsed into the 4DN pairs format, and the polymerase chain reaction duplications were removed by using pairtools (https://github.com/open2c/pairtools). Only contact pairs with UU/UR/RU flags in the pairs file and with both sides mapped to different restriction fragments were kept for further analysis. Last, we generated contact matrices at the 500-kb resolution by using the cooler Python package (*50*).

### WGS and nanopore data processing

For WGS, the paired-end reads were first mapped to hg38 by BWA-MEM (v0.7.17), and duplicate reads were removed by Picard (v2.6.0) (https://github.com/broadinstitute/picard). Then, we used two methods to detect SVs from the same BAM files: (i) We ran Delly (v0.8.7) with parameters "-t ALL -q 20 -s 15" (*53*), and (ii) we ran smoove (v0.2.6) (https://github.com/brentp/smoove) with default parameters, which is an optimized pipeline based on lumpy (*54*). When we evaluated the precision rate for SV predictions from Hi-C (Figs. 2, D, F, and H, and 4, E and F; and fig. S6F), the union of SVs detected by delly and smoove was used as the reference SV set; when we evaluated the recall rate (Figs. 2E and 4, D and F; and fig. S6F), the intersect of delly and smoove was used as the reference SV set. We also inferred copy number profiles from WGS using Control-FREEC (v11.6) (*55*). Multiple ploidy values ("ploidy = 1,2,3,4") were specified in the configuration file to enable the program to automatically select the one that explains the most observed copy number alterations.

For nanopore, we applied three methods for SV detection: sniffles (*56*), Picky (*57*), and svim (*58*). To run sniffles (v1.0.12) and svim (1.4.2), the reads were aligned to hg38 using minimap2 (v2.20) (*59*) with parameters "-ax map-ont -L," and after we have obtained the alignments in BAM format, we ran both methods with default parameters. For Picky, we used LAST (v1256) to align reads. To speed up the calculation, we followed the official pipeline to split the raw FASTQ files into multiple chunks, with each chunk containing 800,000 reads (https://github.com/TheJacksonLaboratory/Picky/wiki/Cluster-Support). We then ran Picky on these chunk files separately and combined results from all chunks. To evaluate the precision rates in Figs. 2 (B and C) and 4C and fig. S6 (A to E), the union of WGS-detected SVs and nanopore-detected SVs was used as the reference SV set, where WGS SVs were defined as the union of SVs from delly and smoove, and nanopore SVs were defined as the union of SVs from svim, sniffles, and Picky. To evaluate the recall rates in Fig. 2C and fig. S6 (D and E), the intersect of WGS-detected SVs and nanopore-detected SVs was used as the reference SV set.

### RNA-seq data processing

The RNA-seq data for BT-474, HCC1954, MCF7, HME1, and MCF10A cell lines were downloaded from GSE152908. The raw FASTQ files were first processed using fastp (v0.20.1) with parameters "--detect_adapter_for_pe --trim_poly_x --correction." The trimmed reads were then processed using the ENCODE long-read RNA-seq pipeline (https://github.com/ENCODE-DCC/long-rna-seq-pipeline) with default parameters to calculate both the genome-wide plus and minus strand signal tracks and gene quantifications.

The gene fusions were detected using Arriba (v2.2.1) (*22*) with suggested parameters using chimeric alignments outputted by STAR (v2.7.10a) as input (https://arriba.readthedocs.io/en/latest/workflow/).

### Down-sample Hi-C contact maps to a specified sequencing depth

Our down-sampling procedure assumes that the number of contacts between two genomic regions follows a binomial distribution. Suppose there are totally $N_{\text{total}}$ contact pairs in the original matrix $M$, and we want to generate a down-sampled matrix $M^{'}$ with around $N_{\text{sample}}$ contact pairs, i.e., with a down-sample rate of $\alpha = N_{\text{sample}}/N_{\text{total}}$, $N_{\text{sample}} < N_{\text{total}}$. To this end, for each nonzero pixel in $M$, we designate the corresponding contact frequency in $M^{'}$ a random integer number generated from a binomial distribution with parameters $M_{ij}$ and α, where $M_{ij}$ is the contact count of the 100% Hi-C matrix between bin $i$ and bin $j$. The same algorithm was also used when we mixed BT-474/HCC1954 Hi-C with HMEC Hi-C at different fractions.

### Collection of training samples and data augmentation

In our previous work (*18*), we have compiled comprehensive SV lists for eight cancer lines (A549, Caki2, K562, LNCaP, NCI-H460, PANC-1, SK-N-MC, and T47D) from multiple experimental platforms. To create a high-quality positive training set for EagleC, we manually curated a set of high-confidence SVs that can be detected by both WGS and optical mapping, and have Hi-C signals surrounding the breakpoints. In total, we obtained 243 such SVs in eight cell lines. We noticed that this original SV set demonstrated severely imbalanced distributions in two aspects: (i) the numbers of SVs with different orientations (++, 37; +−, 96; −+, 61; −−, 29; ++/−−, 15; +−/−+, 5) and (ii) the numbers of SVs at different ranges (short-range SVs, 67; and long-range SVs and translocations, 176). To avoid any biases introduced by such imbalance during the training, and to boost the number of samples, we proposed a data augmentation algorithm as follows: Given a submatrix $M_{ij}$ with a size of 21 × 21, where each entry represents the raw contact frequency between bin $i$ and bin $j$, we can generate a matrix $M_{\text{ij}}^{\prime}$ of the same size, where the value of each entry follows a Poisson distribution with $\lambda = M_{ij}$. By using this algorithm as the core, we increased the positive training set to ~3000 samples for each individual model of the EagleC framework and made sure that these samples had balanced distributions in both different SV orientations and different genomic ranges.

For negative training samples, the chromatin loops in GM12878 were downloaded from a previous study (*21*) with the coordinates converted from hg19 to hg38 using LiftOver (*60*), and A/B compartments were identified using cooltools (v0.3.2, https://github.com/open2c/cooltools) at the 10-kb resolution.

### Implementation of the EagleC framework

EagleC is an ensemble-learning framework that makes predictions based on 50 different models and uses CNN as the individual model. Each CNN model takes 21 × 21 grayscale images as input. Sequentially, the CNN architecture includes the following components: (i) convolution with 32 filters of a kernel size 3 × 3 and stride size 1, followed by the ReLU activation and a 2 × 2 max pooling; (ii) convolution with 64 filters of a kernel size 3 × 3 and stride size 1, followed by the ReLU activation and a 2 × 2 max pooling; and (iii) two fully connected layers with 512 hidden units. The first fully connected layer is followed by the ReLU activation and a dropout layer with a dropout probability of 0.5 to avoid overfitting, and the second fully connected layer acts as the final sigmoid output layer, which computes probability scores for each of the six labels: ++, +−, −+, −−, intranegative, and internegative. Each individual model is randomly initialized with a different set of training samples and trained using an iterative approach (Fig. 1A). During each round of training, the model is optimized against the accuracy using the Adam algorithm. We built the whole framework in Python and the neural network part was implemented using the TensorFlow Keras API (v2.3.0).

Computationally, it is impractical to perform predictions for the submatrix surrounding every pixel of a genome-wide contact map at high resolutions. To speed up the calculation, we perform several prefiltering procedures based on our prior knowledge that SVs usually induce abnormal signals with both high intensity and high density: (i) We filter out pixels where there are fewer than five nonzero values within their 21 × 21 window, and (ii) for intrachromosomal maps, we only consider pixels within the 3 × 3 window of significant interactions. Here, the significant interactions are identified using a model that accounts for the distance-dependent decay of interaction frequencies. Specifically, the expected interaction frequency at given genomic distance $k$ is calculated as follows

$$E_k^* = \begin{cases} \dfrac{1}{n-k} \displaystyle\sum_{|i-j|=k} M_{ij}^*, k < 100 \\ \dfrac{1}{n-100} \displaystyle\sum_{|i-j|=100} M_{ij}^*, k \geq 100 \end{cases}$$

where $M_{ij}^*$ is a CNV-normalized or ICE-normalized intrachromosomal contact matrix with a size of $n \times n$. Note that all pixels with genomic distance greater than 100 bins at a given resolution will have the same expected background. Then, the $P$ value for each observed interaction frequency $M_{ij}$ is calculated on the basis of the Poisson process with expected value $\lambda_{ij} = E_{|i-j|}^*/(W_i \times W_j)$, where $W_i$ is the bias vector extracted either from the "weight" (*50*) (in case of ICE normalization) or "sweight" (*4*) (in case of CNV normalization) column of the ".cool" file. To reduce potential false negatives, we apply a loose $P$ value cutoff of 0.05 to include as many pixels as possible. In addition to the filtering procedures, different intrachromosomal and interchromosomal matrices can be automatically processed in parallel. All that users need to do is to submit the same command for a certain number of times. According to our test on a computational cluster [CPU information: Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz], with 16 parallelized jobs, the program can be generally finished within 3 hours for a dataset with ~200 million contact pairs, which is comparable to or even faster than existing methods (table S6).

The final probability scores for each pixel are calculated by averaging the values across all 50 models. A pixel is identified as a candidate SV breakpoint if the probability of at least one positive label (++, +−, −+, and −−) is greater than a predefined cutoff (we set different cutoffs for different resolutions; see details below). We perform DBSCAN to identify any local clusters of highly scored pixels, and within each cluster, the pixel with the highest probability score will be reported in the final SV list.

To optimize the prediction performance for various sequencing depths, we down-sampled contact matrices of training samples to a series of sequencing depths and independently trained EagleC models for each depth. Specifically, we trained models for six levels of sequencing depths, including 300-800 million (M), 200-300M, 100-200M, 50-100M, 10-50M, and 5-10M contact pairs. In prediction, the most appropriate models were selected according to the number of contacts in the target contact map; for example, SVs in BT-474 and HCC1954 were predicted using the "100-200M" models because the Hi-C maps in these two cell lines contain 192M and 188M contact pairs, respectively.

### Combining SV predictions from multiple resolutions

To optimize the performance of our SV prediction pipeline in both specificity and sensitivity, we propose a strategy that combines predictions from multiple resolutions including 5, 10, and 50 kb. Basically, high-resolution contact matrices (5 or 10 kb) usually achieve higher accuracy and have unique advantages in predicting short-range SVs, while low-resolution contact matrices at 50 kb can complement the predictions when sequencing depths are not sufficient to cover the real SV breakpoints at high resolutions. In our pipeline, we first predict SVs at 5-, 10-, and 50-kb resolutions separately. The default probability cutoffs are empirically set to 0.8, 0.8, and 0.99999 for 5, 10, and 50 kb, respectively. However, according to our test against the benchmark datasets used in this study, EagleC is pretty robust to different cutoff values, but generally tuning down the cutoffs detects more SVs with slightly lower accuracy, while tuning up the cutoffs detects fewer SVs with slightly higher accuracy (fig. S11). For 10- and 50-kb predictions, we further search for the most probable breakpoint coordinates within a local region on 5-kb contact maps so that all the reported SVs will have the 5-kb resolution. After we have obtained SV predictions from individual resolutions, we merge the SV coordinates from different resolutions together and only report nonredundant SVs for each sample.

### Application of EagleC to scHi-C

We made the following modifications when we applied EagleC to scHi-C: (i) All the training and predicting steps were based on contact matrices at the 500-kb resolution; (ii) raw contact signals were used instead of ICE/CNV-normalized signals; (iii) the models were trained for eight levels of sequencing depths with lower number of contacts, including 10-20M, 5-10M, 3-5M, 1-3M, 750K-1M, 500-750K, 250-500K, and 100-250K contact pairs; and (iv) in prediction, the probability cutoff was set to 0.95.

### Annotations of duplications and deletions

We defined duplications and deletions by using both the orientation information of SV breakpoints and copy number profiles. Specifically, duplications were defined as intrachromosomal SVs with −+, ++, or −− orientations, and the genomic interval between breakpoints had a copy number ratio larger than 1.5, while deletions were defined as intrachromosomal SVs with the +− orientation, and the genomic interval had a copy number ratio smaller than 0.3. Copy number profiles calculated from WGS were used if WGS was available; otherwise, we used copy number profiles inferred from Hi-C in this calculation (*4*). In addition, in Fig. 5 (F and G), we only considered short-range SVs with a breakpoint distance of less than 1 Mb.

### Identification of compartments and TADs

We identified A/B compartments and TADs on Hi-C contact maps of several normal cell lines or tissues (table S5) to investigate the associations between 3D genomic architectures and SV formation. The A/B compartments were identified using cooltools (v0.3.2). Briefly, the eigenvalue decomposition was performed on the 100-kb intrachromosomal contact maps, and the first eigenvector (PC1) was used to capture the "plaid" contact pattern. The original PC1 was oriented according to gene densities (Ensembl 93) so that positive values correspond to active genomic regions (A compartment) and negative values correspond to inactive regions (B compartment). For TADs, we ran HiTAD at the 25-kb resolution and defined TADs as the bottom-level domains returned by HiTAD (*61*).

## SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abn9215

View/request a protocol for this paper from *Bio-protocol*.

## REFERENCES AND NOTES

1. R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. McHenry, R. M. Pinchback, A. H. Ligon, Y. J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M. S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, M. Meyerson, The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
2. F. Mitelman, B. Johansson, F. Mertens, The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).
3. J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stutz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl, H. Brenner, O. T. Brustugun, S. M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger, R. Sotillo, H. Glimm, J. O. Korbel, Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
4. X. Wang, J. Xu, B. Zhang, Y. Hou, F. Song, H. Lyu, F. Yue, Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
5. P. Hasty, C. Montagna, Chromosomal rearrangements in cancer: Detection and potential causal mechanisms. *Mol. Cell. Oncol.* **1**, e29904 (2014).
6. T. S. Wan, Cancer cytogenetics: Methodology revisited. *Ann. Lab. Med.* **34**, 413–425 (2014).
7. T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhsng, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
8. P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. Edwards, G. R. Bignell, M. R. Stratton, P. A. Futreal, Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
9. E. R. Mardis, R. K. Wilson, Cancer genome sequencing: A review. *Hum. Mol. Genet.* **18**, R163–R168 (2009).
10. S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjaerde, A. Langerod, M. Ringner, S. M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. Hooijer, S. J. Jang, D. R. Jones, H. Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J. Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporte, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodriguez-Gonzalez, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. 't van Veer, A. Tutt, S. Knappskog, B. K. Tan, J. Jonkers, A. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. Martens, A. L. Borresen-Dale, A. L. Richardson, G. Kong, G. Thomas, M. R. Stratton, Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
11. L. Harewood, K. Kishore, M. D. Eldridge, S. Wingett, D. Pearson, S. Schoenfelder, V. P. Collins, P. Fraser, Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.* **18**, 125 (2017).
12. M. Ghandi, F. W. Huang, J. Jane-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald III, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet, B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, W. R. Sellers, Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
13. K. C. Akdemir, V. T. Le, S. Chandran, Y. Li, R. G. Verhaak, R. Beroukhim, P. J. Campbell, L. Chin, J. R. Dixon, P. A. Futreal; P. S. V. W. Group, P. Consortium, Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305 (2020).
14. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
15. Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
16. A. Zhou, T. Lin, J. Xing, Evaluating nanopore sequencing data processing pipelines for structural variation identification. *Genome Biol.* **20**, 237 (2019).
17. A. Chakraborty, F. Ay, Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics* **34**, 338–345 (2018).
18. J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardimci, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, F. Yue, Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
19. S. Wang, S. Lee, C. Chu, D. Jain, P. Kerpedjiev, G. M. Nelson, J. M. Walsh, B. H. Alver, P. J. Park, HiNT: A computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.* **21**, 73 (2020).
20. K. Kim, M. Kim, Y. Kim, D. Lee, I. Jung, Hi-C as a molecular rangefinder to examine genomic rearrangements. *Semin. Cell Dev. Biol.* , (2022).
21. S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
22. S. Uhrig, J. Ellermann, T. Walther, P. Burkhardt, M. Frohlich, B. Hutter, U. H. Toprak, O. Neumann, A. Stenzinger, C. Scholl, S. Frohling, B. Brors, Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
23. M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, Y. Ruan, An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
24. M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, H. Y. Chang, HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
25. R. Fang, M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, B. Ren, Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
26. B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, B. Herman, S. Happe, A. Higgs, E. LeProust, G. A. Follows, P. Fraser, N. M. Luscombe, C. S. Osborne, Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
27. M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schopflin, K. Kraft, R. Kempfer, I. Jerkovic, W. L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, S. Mundlos, Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

28. B. K. Kragesteen, M. Spielmann, C. Paliou, V. Heinrich, R. Schopflin, A. Esposito, C. Annunziatella, S. Bianco, A. M. Chiariello, I. Jerkovic, I. Harabula, P. Guckelberger, M. Pechstein, L. Wittler, W. L. Chan, M. Franke, D. G. Lupianez, K. Kraft, B. Timmermann, M. Vingron, A. Visel, M. Nicodemi, S. Mundlos, G. Andrey, Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* **50**, 1463–1473 (2018).

29. A. Despang, R. Schopflin, M. Franke, S. Ali, I. Jerkovic, C. Paliou, W. L. Chan, B. Timmermann, L. Wittler, M. Vingron, S. Mundlos, D. M. Ibrahim, Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).

30. K. Kraft, A. Magg, V. Heinrich, C. Riemenschneider, R. Schopflin, J. Markowski, D. M. Ibrahim, R. Acuna-Hidalgo, A. Despang, G. Andrey, L. Wittler, B. Timmermann, M. Vingron, S. Mundlos, Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.* **21**, 305–310 (2019).

31. P. H. G. Duijf, D. Nanayakkara, K. Nones, S. Srihari, M. Kalimutho, K. K. Khanna, Mechanisms of genomic instability in breast cancer. *Trends Mol. Med.* **25**, 595–611 (2019).

32. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

33. J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

34. A. Canela, Y. Maman, S. Jung, N. Wong, E. Callen, A. Day, K. R. Kieffer-Kwon, A. Pekowska, H. Zhang, S. S. P. Rao, S. C. Huang, P. J. McKinnon, P. D. Aplan, Y. Pommier, E. L. Aiden, R. Casellas, A. Nussenzweig, Genome organization drives chromosome fragility. *Cell* **170**, 507–521.e18 (2017).

35. B. E. Clurman, M. Groudine, The *CDKN2A* tumor-suppressor locus—A tale of two proteins. *N. Engl. J. Med.* **338**, 910–912 (1998).

36. S. Khawaled, G. Nigita, R. Distefano, S. Oster, S. S. Suh, Y. Smith, A. Khalaileh, Y. Peng, C. M. Croce, T. Geiger, V. L. Seewaldt, R. I. Aqeilan, Pleiotropic tumor suppressor functions of WWOX antagonize metastasis. *Signal Transduct. Target. Ther.* **5**, 43 (2020).

37. L. Kashima, M. Toyota, H. Mita, H. Suzuki, M. Idogawa, K. Ogi, Y. Sasaki, T. Tokino, CHFR, a potential tumor suppressor, downregulates interleukin-8 through the inhibition of NF-kappaB. *Oncogene* **28**, 2643–2653 (2009).

38. D. Zink, C. Mayr, C. Janz, L. Wiesmuller, Association of p53 and MSH2 with recombinative repair complexes during S phase. *Oncogene* **21**, 4788–4800 (2002).

39. H. Chen, H. Liu, G. Qing, Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct. Target. Ther.* **3**, 5 (2018).

40. C. Chen, S. Zhao, A. Karnad, J. W. Freeman, The biology and role of CD44 in cancer progression: Therapeutic implications. *J. Hematol. Oncol.* **11**, 64 (2018).

41. V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, J. Shendure, Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).

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

43. Y. Shibata, A. Malhotra, A. Dutta, Detection of DNA fusion junctions for BCR-ABL translocations by Anchored ChromPET. *Genome Med.* **2**, 70 (2010).

44. H. Yang, Y. Luan, T. Liu, H. J. Lee, L. Fang, Y. Wang, X. Wang, B. Zhang, Q. Jin, K. C. Ang, X. Xing, J. Wang, J. Xu, F. Song, I. Sriranga, C. Khunsriraksakul, T. Salameh, D. Li, M. N. K. Choudhary, J. Topczewski, K. Wang, G. S. Gerhard, R. C. Hardison, T. Wang, K. C. Cheng, F. Yue, A map of cis-regulatory elements and 3D genome structures in zebrafish. *Nature* **588**, 337–343 (2020).

45. O. Poirion, X. Zhu, T. Ching, L. X. Garmire, Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nat. Commun.* **9**, 4892 (2018).

46. X. F. Mallory, M. Edrisi, N. Navin, L. Nakhleh, Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol.* **21**, 208 (2020).

47. R. Gao, S. Bai, Y. C. Henderson, Y. Lin, A. Schalck, Y. Yan, T. Kumar, M. Hu, E. Sei, A. Davis, F. Wang, S. F. Shaitelman, J. R. Wang, K. Chen, S. Moulder, S. Y. Lai, N. E. Navin, Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).

48. L. Tan, D. Xing, C. H. Chang, H. Li, X. S. Xie, Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).

49. D. S. Lee, C. Luo, J. Zhou, S. Chandran, A. Rivkin, A. Bartlett, J. R. Nery, C. Fitzpatrick, C. O'Connor, J. R. Dixon, J. R. Ecker, Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* **16**, 999–1006 (2019).

50. N. Abdennur, L. A. Mirny, Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).

51. N. C. Durand, M. S. Shamim, I. Machol, S. S. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016).

52. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 (2013).

53. T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, J. O. Korbel, DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

54. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).

55. V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, E. Barillot, Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).

56. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

57. L. Gong, C. H. Wong, W. C. Cheng, H. Tjong, F. Menghi, C. Y. Ngan, E. T. Liu, C. L. Wei, Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat. Methods* **15**, 455–460 (2018).

58. D. Heller, M. Vingron, SVIM: Structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).

59. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

60. R. M. Kuhn, D. Haussler, W. J. Kent, The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

61. X. T. Wang, W. Cui, C. Peng, HiTAD: Detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res.* **45**, e163 (2017).

**Acknowledgments**
**Funding:** F.Y. is supported by NIH grants 5R01HG011207, 5R35GM124820, 5R01HG009906, and 1U24HG012070. **Author contributions:** F.Y. conceived and supervised the project. X.W. implemented the EagleC framework and performed the data analysis. Y.L. helped process the scHi-C data. **Competing interests.** F.Y. and X.W. are listed as inventors of a provisional patent based on this work. F.Y. is a scientific cofounder for Sariant Therapeutics Inc. The remaining author declares no competing interests. **Data and materials availability:** A list of high-confidence SVs we compiled for training EagleC models is provided in table S1. Detailed information of cancer Hi-C datasets collected in this study is summarized in table S2. Detailed information of HiChIP/ChIA-PET datasets is summarized in table S3. Breakpoint coordinates predicted in each sample are provided in table S4. Hi-C datasets in normal cell lines or tissues are summarized in table S5. The list of cancer-related genes was obtained from the Bushman Lab (http://bushmanlab.org/assets/doc/allOnco_May2018.tsv). The EagleC software and detailed documentation are available at Zenodo (https://doi.org/10.5281/zenodo.6482060) and GitHub (https://github.com/XiaoTaoWang/EagleC).

Submitted 31 December 2021
Accepted 28 April 2022
Published 15 June 2022
10.1126/sciadv.abn9215