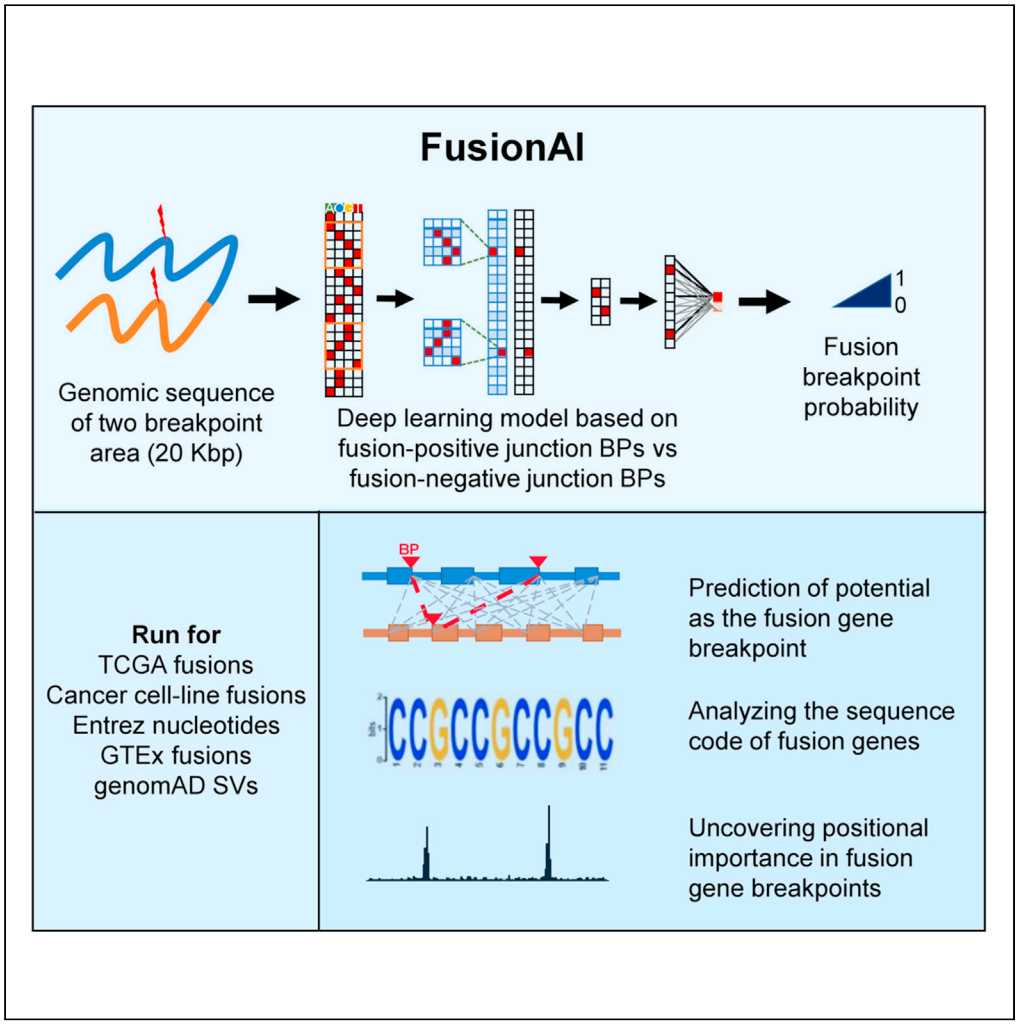
iScience

CellPress
OPEN ACCESS

Article

FusionAI: Predicting fusion breakpoint from DNA sequence with deep learning



Pora Kim, Hua Tan, Jiajia Liu, Mengyuan Yang, Xiaobo Zhou

xiaobo.zhou@uth.tmc.edu (X.Z.)
pora.kim@uth.tmc.edu (P.K.)

Highlights

FusionAI predicts fusion gene breakpoints from a DNA sequence

FusonAI reduce the effort for validating fusion genes with other tools

High feature importance regions were apart 100nt from the exon junction BPs

High feature importance regions were overlapped with 44 human genomic features

Kim et al., iScience *24*, 103164
October 22, 2021
https://doi.org/10.1016/j.isci.2021.103164

iScience

CellPress
OPEN ACCESS

Article

FusionAI: Predicting fusion breakpoint from DNA sequence with deep learning

Pora Kim,[1,6,7,*] Hua Tan,[1,6] Jiajia Liu,[1,4] Mengyuan Yang,[1,5] and Xiaobo Zhou[1,2,3,*]

SUMMARY

Identifying the molecular mechanisms related to genomic breakage is an important goal of cancer mechanism studies. Among diverse locations of structural variants, fusion genes, which have the breakpoints in the gene bodies and are typically identified from the split reads of RNA-seq data, can provide a highlighted structural variant resource for studying the genomic breakages with expression and potential pathogenic impacts. In this study, we developed FusionAI, which utilizes deep learning to predict gene fusion breakpoints based on DNA sequence and let us identify fusion breakage code and genomic context. FusionAI leverages the known fusion breakpoints to provide a prediction model of the fusion genes from the primary genomic sequences via deep learning, thereby helping researchers a more accurate selection of fusion genes and better understand genomic breakage.

INTRODUCTION

Identifying the molecular mechanisms related to the genomic breakage is one of the important goals of disease biology studies to understand the origin of new genes and aberrant functional features from broken genomes. Among diverse locations of structural variants, breakpoints of fusion genes are located in the gene bodies. Fusion genes are formed mainly through the chromosomal rearrangements initiated by DNA double-strand breakages. Due to the cost-effectiveness (data creation) and analysis (interpretation), and usage (diagnosis), there is a huge amount of RNA-seq data accumulated to date. Fusion genes are usually identified from the split reads (unmapped reads) of RNA-seq data as the form of chimeric transcripts. These expressed fusion genes can provide a highlighted structural variant resource for studying the genomic breakages with expression and thereby potential pathogenic impacts. Indeed, the broken gene context of fusion genes provided the aberrant functional clues to study disease pathogenesis, specifically in cancer (Kim et al., 2020). However, to predict fusion genes correctly, an inherent limitation of RNA-seq data and analyses are restricted by diverse combinations of limiting factors, such as different conditions of sequencing depth, read length, read alignment tools and software options, filtering criteria, and etc., which create many false positives. Most of all, even though, if there is a robust, reproducible, and unbiased method, we cannot identify the fusion genes that were lowly expressed. Therefore, developing the sequencing-free prediction method of fusion genes would be helpful and may provide new insights into the genomic breakage phenomenon in the cell.

Motivated by recent success in the use of deep learning approaches to predict the genomic regulatory elements and alternative splice events from the genomic context (Jaganathan et al., 2019; Zhou and Troyanskaya, 2015), we hypothesized that the exon junctional breakpoints of known fusion genes identified from the split reads of RNA-seq data can be used to construct a deep-learning model of predicting the breakage tendency. Because it is very rare to have and analyze the real genomic breakpoints of fusion genes from the matched samples' RNA-seq and whole-genome sequencing (WGS) data, it is hard to have enough number of fusion events that have the information about the real genomic breakpoints to build a deep learning model. Furthermore, real genomic breakpoints of fusion genes are usually located in the introns from a larger percentage across the genome than exons as 1% versus 24% or across transcriptome (4% versus 96%), respectively. Accordingly, the breakpoints at the fusion transcripts are located at the exon junction boundaries. Therefore, if we build the model from the RNA-sequence context, there would be no chance to identify the genomic sequence features relating to the DNA double-strand breakage but may have more chance to identify some features relating to alternative splicing events, which is very sensitive to the exon junction boundaries. However, if we study based on the intron sequence that has the real genomic breakage near to the chimeric exon junctions, then there is a chance to identify genomic breakage features.

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[3]School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[4]College of Electronic and Information Engineering, Tongji University, Shanghai, Shanghai 201804, China

[5]School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[6]These authors contributed equally

[7]Lead contact

*Correspondence: xiaobo.zhou@uth.tmc.edu (X.Z.), pora.kim@uth.tmc.edu (P.K.)

https://doi.org/10.1016/j.isci.2021.103164

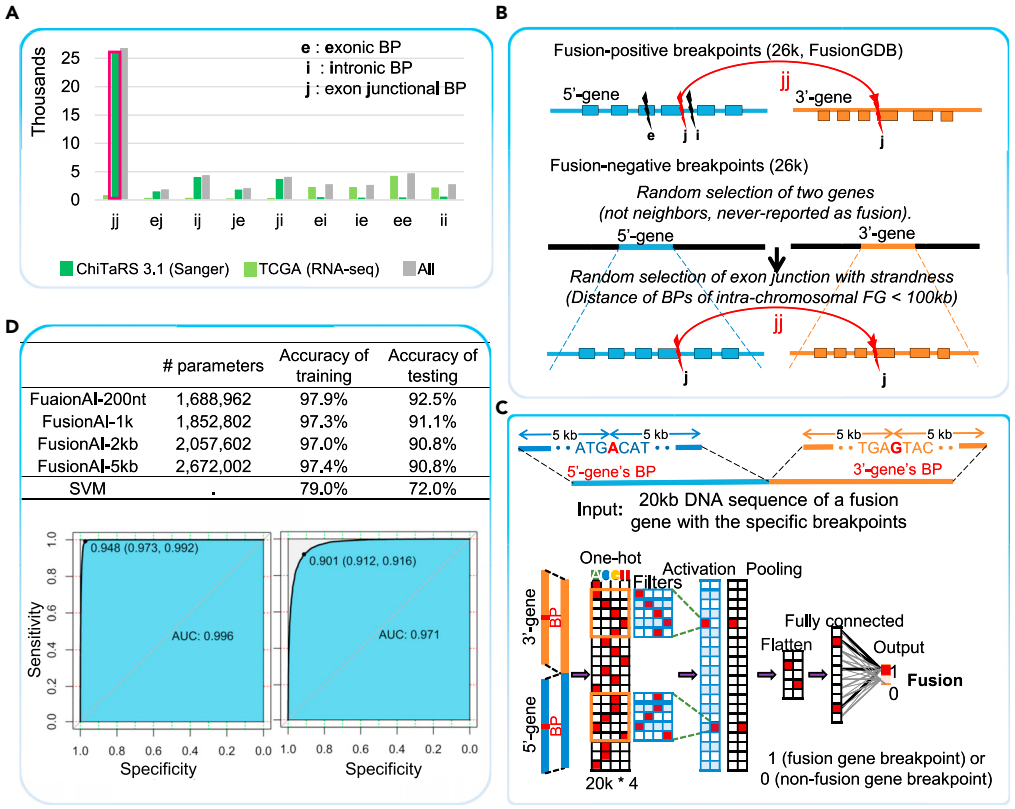


Figure 1. Overview of FusionAI

(A) The investigation of fusion gene breakpoints of 48K FGs from FusionGDB identified the BP location across the human genome.

(B) Making training and test datasets of fusion-positive and -negative breakpoints.

(C) Diagram of fusion gene breakpoints classification by FusionAI.

(D) Effect of the size of the input sequence context on the accuracy.

To test our hypothesis, we developed FusionAI, a deep residual neural network that predicts whether a fusion gene breakpoint at the exon junction-junction area predicted from RNA-seq data will be a potential fusion gene breakpoint or not, from the sole input of the primary DNA sequence. Through FusionAI, we also want to understand the genomic features that were highly enriched in the genomic breakpoint area. FusionAI consists of a deep neural network (DNN) model that classifies between fusion-positive and -negative breakpoints on the basis of DNA sequences. We used the fusion genes that have exon junction-junction breakpoints in both fusion partner genes of The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research et al., 2013) among nine different combinations of fusion breakpoint location annotated from FusionGDB (Kim and Zhou, 2019) on hg19 genome version (Figure 1A). Among ~48K human fusion gene breakpoints from FusionGDB, there were ~33K of fusion breakpoints from TCGA. Of these TCGA fusions, there were ~26K fusions which have the breakpoints located at the exon junction-junction positions. We used these 26K fusions as the positive data. For the negative data, we made pseudo fusion breakpoint sequences by stringent criteria with a similar number of ~26K (Figure 1B). Then, we divided these total ~52K fusions (~26K positives and ~26K negatives) into 70% (~18K positives and ~18K negatives) and 30% (~8K positives and ~8K negatives) for the training and test data sets, and trained and built the FusionAI. The input is total 20 kb length DNA sequence from combining of ± 5kb flanking sequence from two breakpoints (Figure 1C). The accuracies for training and test datasets were 97.4% (AUROC = 0.9962) and 90.8% (AUROC = 0.9706) with 0.12 and 0.42 error rate, respectively (Figure 1D).

Except for this validation using the internal independent test dataset of 16K fusion events, we also evaluated the performance of FusionAI by applying it to multiple external independent datasets. Those are experimentally validated fusion gene datasets such as validated 2200 fusion genes from 675 human cancer

cell-lines (Klijn et al., 2015), and Sanger sequencing-based fusion genes in Entrez from ChiTaRs-3.1 (Gorohovski et al., 2017). Furthermore, to see the difference of FusionAI scores between cancer and normal population, we applied FusionAI to the fusion events from a non-disease population of Genotype-Tissue Expression (GTEx) (Consortium et al., 2017; Singh et al., 2020). Next, to check the difference of FusionAI scores between RNA-seq data-based fusions and potential fusions from the WGS data, we ran FusionAI to the potential fusion genes that might be derived from the structural variants of The Genome Aggregation Database (genomAD) (Collins et al., 2020) annotated from FGviewer (Kim et al., 2020). From these validations to the real data, we noticed that the application of FusionAI increased the specificity of fusion gene prediction. To identify the genomic context that has an impact on classifying fusion or not, we investigated the feature importance scores of 20 nt window across 20 kb sequence. Naturally, the feature importance (FI) scores reflected the output scores of FusionAI well for classifying fusion-positives and -negatives. Specifically, we studied whether there is a significant enrichment of top 1% (high) FI scores in the fusion-positive sequence compared to the fusion-negative sequence. Across the fusion-positive 20kb sequence, we investigated the overlap with 44 human genomic sequence features in diverse categories such as virus integration sites, multiple types of repeats, structural variant regions, specific chromatin stated regions, and expression regulating regions. Most of all, from these top 1% FI scored regions of 322 transcription factor fusion genes, we identified a GC high DNA sequence motif, which might be targeted by SP1. The low percentage of FusionAI prediction scores in the healthy population derived fusion genes might reflect a potential validity or tumorigenicity of individual fusion gene breakpoint. In summary, FusionAI is an example of interpretable scientific deep learning in studying the human genomic breakages with diverse potential genomic regions related to different cellular mechanisms.

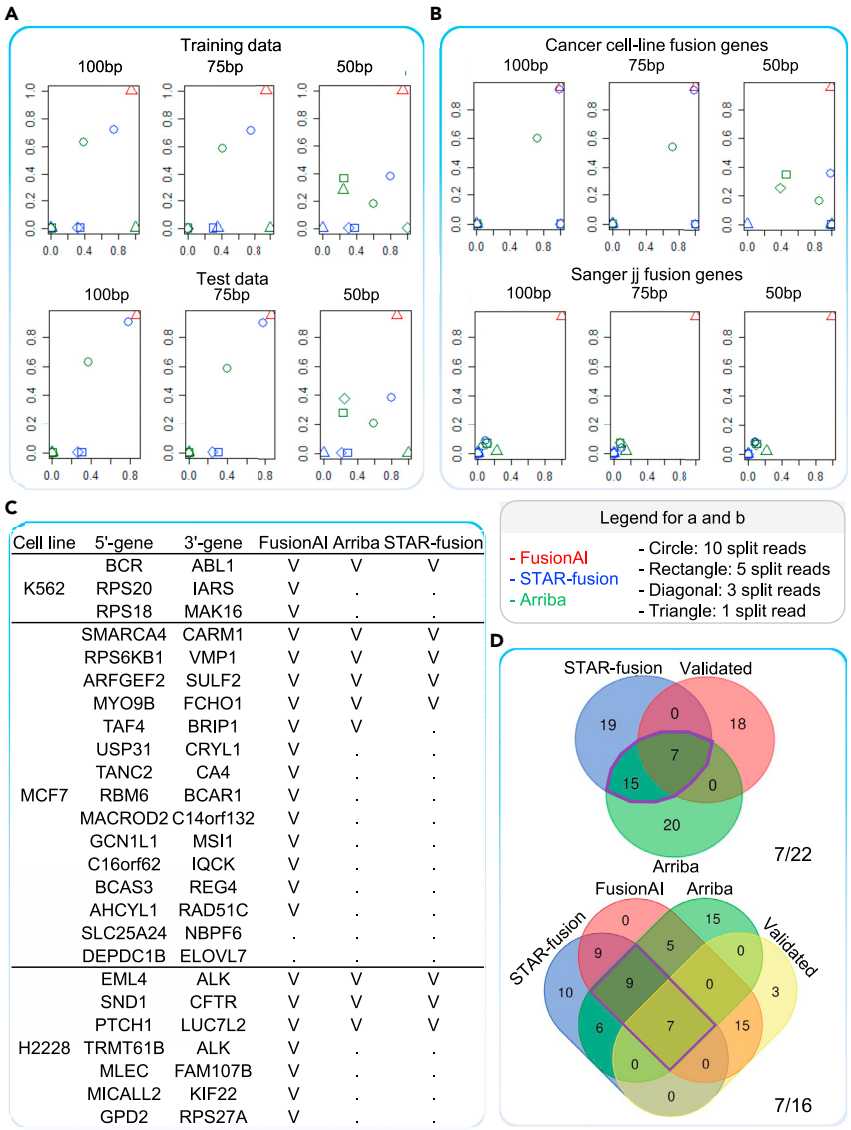
RESULTS

Overview of FusionAI

Prior to constructing the model, we investigated the distribution of the fusion breakpoint location on gene structures. The majority of the fusion genes are predicted from RNA-seq data, specifically from the unmapped split reads which aligned at exon junction-junction positions of two different genes (Figure 1A). This provides the evidence of the hypothesis that the breakpoints of fusion genes would be located in the intron since the exons cover the human genome only about 1%, but the introns cover more than 24% of the reference genome, which is equivalent to 4% and 96% in the transcriptome (Venter et al., 2001). From this context, we used breakpoint information of fusion genes that have both breakpoints at the exon junction-junction sites among nine different combinations of breakpoints from the TCGA cohort as the fusion-positive datasets (~26k fusion breakpoints in Table S1) (Figure 1A). We made a similar number of fusion-negative data with strict criteria (Figure 1B. See the STAR Methods section). Using the divided data from the mixture of fusion-positives and -negatives into 70% and 30% (36K and 16K), we trained and test FusionAI, respectively. The input is a 20 kb length DNA sequence from combining of $\pm$ 5 kb flanking sequence from two breakpoints. The transformed one-hot encoded input resultant into the probability of fusion breakpoints through passing the deep learning processes including filtering, activation, pooling, flattening, and fully connected functions (Figure 1C). To examine long-range and short-range specificity determinants of the input sequence, we compared the scores assigned to fusion-positives by the models trained on 200 nt, 1k, and 2k of the sequence context versus the full model that is trained on 5k of context (Figure 1D). Overall, there was no big difference in the accuracy across the results using different sequence range from the fusion breakpoints. However, to identify the breakpoint-specific genomic context, we decided to use 5k based model. Then, the accuracies for training and test datasets were 97.4% (AUROC = 0.9962) and 90.8% (AUROC = 0.9706) with 0.12 and 0.42 error rates, respectively (Figure 1D). This performance is much better than the traditional machine learning methods like support vector machine (SVM). To identify the hidden genomic context features specific fusion gene breakpoints, we performed our studies based on the 5k-based model with a convolutional neural network approach.

Primary sequence-based FusionAI improves the identification of fusion genes

To compare the performance of prediction of fusion genes of FusionAI with other RNA-seq based fusion gene prediction tools, we chose STAR-fusion and Arriba based on the best performance results from the accuracy assessment study (Haas et al., 2019; Uhrig et al., 2021). Our training and test datasets are constructed based on the genomic breakpoint information. Since the usual RNA-seq based tools require the input of RNA-seq data, we made the simulation RNA-seq data of the split reads at the exon junction-junction breakpoints with different read lengths (50, 75, and 100 bp) and different numbers of split reads (1, 3, 5 split reads, and 10 random around breakpoints) for the fusion-positive and -negative



Cell line	5'-gene	3'-gene	FusionAI	Arriba	STAR-fusion
	BCR	ABL1	V	V	V
K562	RPS20	IARS	V	.	.
	RPS18	MAK16	V	.	.
	SMARCA4	CARM1	V	V	V
	RPS6KB1	VMP1	V	V	V
	ARFGEF2	SULF2	V	V	V
	MYO9B	FCHO1	V	V	V
	TAF4	BRIP1	V	V	.
	USP31	CRYL1	V	.	.
	TANC2	CA4	V	.	.
MCF7	RBM6	BCAR1	V	.	.
	MACROD2	C14orf132	V	.	.
	GCN1L1	MSI1	V	.	.
	C16orf62	IQCK	V	.	.
	BCAS3	REG4	V	.	.
	AHCYL1	RAD51C	V	.	.
	SLC25A24	NBPF6	.	.	.
	DEPDC1B	ELOVL7	.	.	.
	EML4	ALK	V	V	V
	SND1	CFTR	V	V	V
	PTCH1	LUC7L2	V	V	V
H2228	TRMT61B	ALK	V	.	.
	MLEC	FAM107B	V	.	.
	MICALL2	KIF22	V	.	.
	GPD2	RPS27A	V	.	.

Figure 2. Performance of FusionAI

(A) Comparison of FusionAI to other methods for fusion gene prediction including 38,000 TCGA fusion genes from training and test datasets. The plots show the number of true positives and sensitivity from the left.

(B) Comparison of the number of the true positive fusions in ~2200 validated fusion genes in ~530 cancer cell-lines, and 862 Sanger sequence-based fusion genes that have fusion breakpoints at the exon junction position from ChiRTaRS3.1.

(C) Comparison of predicting fusion genes in three cancer cell-lines (H2228, K562, and MCF7).

(D) Identification of validated fusion genes in three cell-lines.

breakpoints in our training and test data sets (Figure 2A). We also made simulation RNA-seq data for the experimentally validated fusion gene breakpoints by Sanger sequencing and RT-PCR from ChiTaRS-3.1(Gorohovski et al., 2017) and cancer cell-line study(Klijn et al., 2015), respectively. To apply FusionAI to these data sets, we made the input sequence of 20 kb long (Table S1). As shown in these comparisons in Figures 2A and 2B, the typical RNA-seq's limiting factors (i.e., read-length and the number of split reads) were not problems to FusionAI compared with the general RNA-seq-based fusion prediction tools as shown in the red triangles in the plots (Figures 2A and 2B; Tables S2–S4). We also checked the individual fusion genes validated from the most famous fusion gene-positive cell-lines such as K562, MCF7, and NCI-H660, which are the *BCR-ABL1*, *BCAS4-BCAS3*, and *EML4-ALK*-positive cells, respectively (Figure 2C). Figure 2D shows that FusionAI has the biggest number of validated fusion genes among the three tools.

Typically, the researchers use multiple fusion gene prediction tools to select the candidates before experimental validation. Compared to the results using only two tools of STAR-fusion and Arriba, additional use of FusionAI reduced the number of false positives effectively but remained true positives. This can reduce the cost and efforts for validation. As the sequencing-free fusion prediction method, FusionAI can be a useful last step filtering scheme.

FusionAI facilitates a better understanding of the genomic context of fusion gene breakpoints

We investigated the feature importance (FI) scores using 20 bp window size every 20 bp position along the 20kbp input sequence of our training and test fusion genes. These feature importance scores are the values reflecting how big impact individual 20 bp length sequences along 20k have to distinguish fusion-positive and -negative breakpoints. The details on the feature importance scores are in the method section. As shown in the six most famous fusion genes (*BCR-ABL1*, *EML4-ALK*, *TMPRSS2-ERG*, *PML-RARA*, *RUNX1-RUNX1T1*, *and FGFR3-TACC3*), the feature importance scores were most high at the breakpoint area (Figure 3). For the intuitive validation, we checked the performance of FusionAI how much the output scores classify the fusion-positives and -negatives using the logistic regression approach. As shown in Figure 3B, FusionAI classifies the fusion-positive and -negative breakpoints significantly (p value $<2 \times 10^{-18}$). We also wondered about the relationship between FI scores and FusionAI output scores. Because some fusion genes had very small values of FI scores (i.e., *BCR-ABL1* in Figure 3A), we transformed the original values of FusionAI and FI scores to the quantile normalized values. Then, we identified the existence of grouping tendency between fusion-positives and -negatives among the FusionAI output scores, and second and third principal components as shown in Figure 3C.

High feature importance scored regions provide a landscape of the genomic feature aspect of fusion gene breakpoints

To date, there were many trials to understand the gnomic features of breakage and to study multiple effects of the genomic breakage (Ballinger et al., 2019; Chakraborty et al., 2020; Fungtammasan et al., 2012; Peng et al., 2006). In this study, we sought to identify the genomic features of the fusion gene breakpoint area across the human genome sequences. Overall, the top 1% feature importance scored regions were enriched near to the breakpoints among 20Kbp sequence (Figure S1), which is the distribution of median values of the top 1% FI scores per nucleotide across 20Kbp sequence. We integrated 44 different human genomic features belong five important cellular mechanism categories such as integration site category of 6 viruses, 13 types of repeat category, 5 types of structural variant category, 15 different types of chromatin state category, and 5 gene expression regulatory category to have the landscape of genomic features in the fusion breakpoint area. For individual features of these five categories, we counted the unique number of the overlap between feature loci with the top 1% FI scored regions in every nucleotide across 20k sequence of all fusion genes in both fusion-positive and -negative groups (Figure 3Di). The overall distribution of overlaps was enriched in the fusion gene breakpoint area. Furthermore, overall, the top 1% feature importance scored regions in fusion-positive breakpoints were more overlapped with genomic features than one of fusion-negative breakpoints. We checked the difference between the two groups whether the number of overlaps between individual features out of 44 and fusion-positive and -negative group using chi-square test (Table S5). From this study, we found that there was a significant difference in the overlap number of features. Next, we counted the unique number of the overlapped loci of the individual feature with all regions of the 20Kbp breakpoint sequence in both fusion-positive and -negative groups to see without potential confounding factors in the genome (Figure S2). Then, the overall patterns between fusion-positive and -negative breakpoints were similar. These similar patterns might be able to explain the rigorous approach to create fusion-negative datasets as the reliable combination of two different genomic regions.

Specifically, we counted the unique number of the overlapped loci of the individual feature with all regions of the 20Kbp breakpoint sequence focusing on the fusion-positive group to see without potential confounding factors in the genomes (Figure 3Dii). From this distribution, we identified several genomic features that showed different distribution around the breakpoint area. In the repeat category (green background), two repeats like G-Quadruplex forming repeats and low complexity AT-rich regions were increased to the breakpoint area. G-Quadruplex (G4) is formed in nucleic acids by sequences that are rich in guanine. G-Quadruplex is divided into two groups of telomeric quadruplexes and non-telomeric quadruplexes. The former quadruplexes have been shown to decrease the activity of the enzyme telomerase. A large number of the latter quadruplexes were found within gene promoters. On the other hand, Alu, L1, and L2 repeats were decreased to the breakpoint area. As for the DNA double-strand break repair

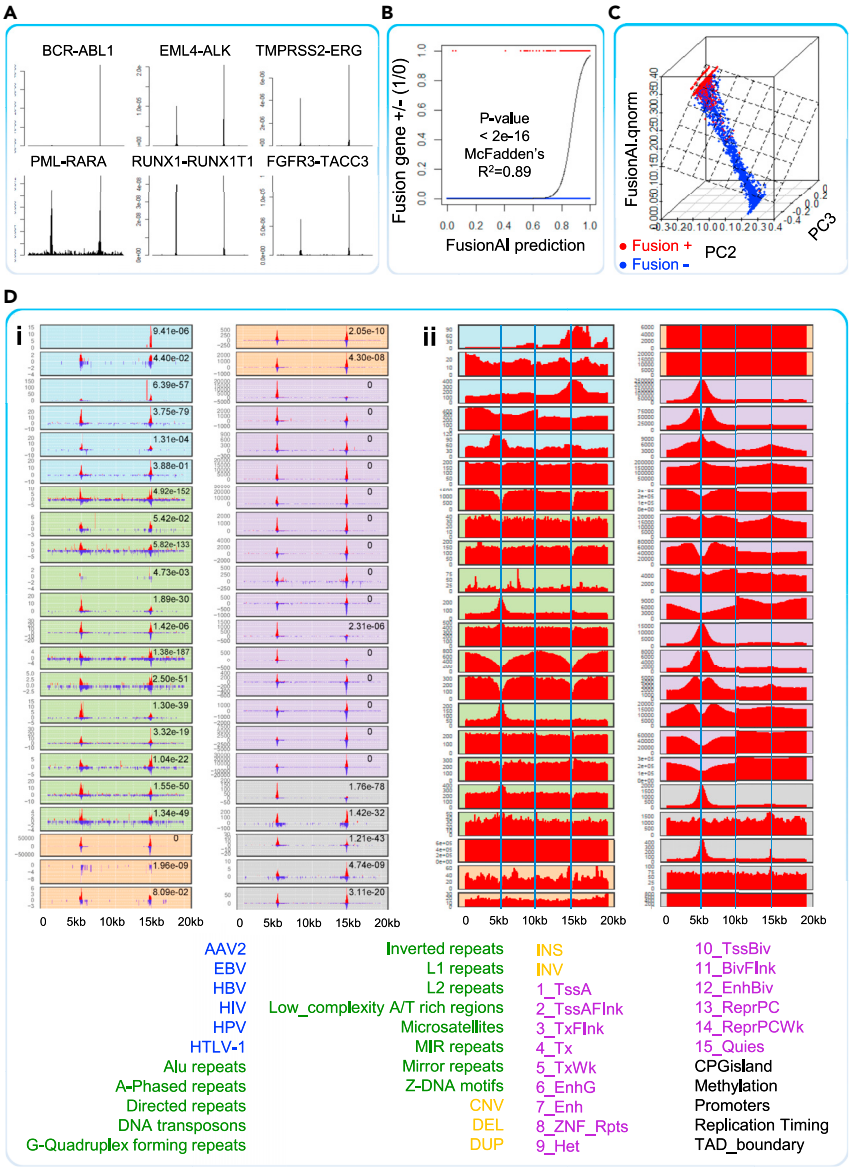


Figure 3. Feature importance score for understanding genomic breakage

(A) Distribution of FI scores across 20 kb long of six representative fusion gene breakpoints.

(B) Logistic regression result of FusionAI prediction.

(C) Classification between fusion-positive and -negative from FusionAI and FI scores.

(D) (i) Distribution of overlaps between top 1% FI scored regions and 44 different types of human genomic features in both positive and negative data. (ii) Distribution of overlaps between all regions and 44 different types of human genomic features.

process, a study observed that Alu elements followed the global genome repair kinetics, while LINE-1 elements repaired at a slower rate (Natale et al., 2018). In the chromatin state category (purple background), 1_TssA and 10_TssBiv chromatin states showed increased distribution to the breakpoints. Those states represent active TSS and bivalent/poised TSS. In the expression regulation category (gray background), CpG island and promoter regions showed increased distribution around the breakpoints. CpG islands are normally found at promoters (Sleutels and Barlow, 2002). From these plots, we can see that LINE repeats and promoter regions were decreased and increased, respectively, toward both breakpoints of two fusion partner genes. Long interspersed nuclear element-1 (LINE-1) retrotransposition is a major hallmark of cancer accompanied by global chromosomal instability, genomic instability, and genetic

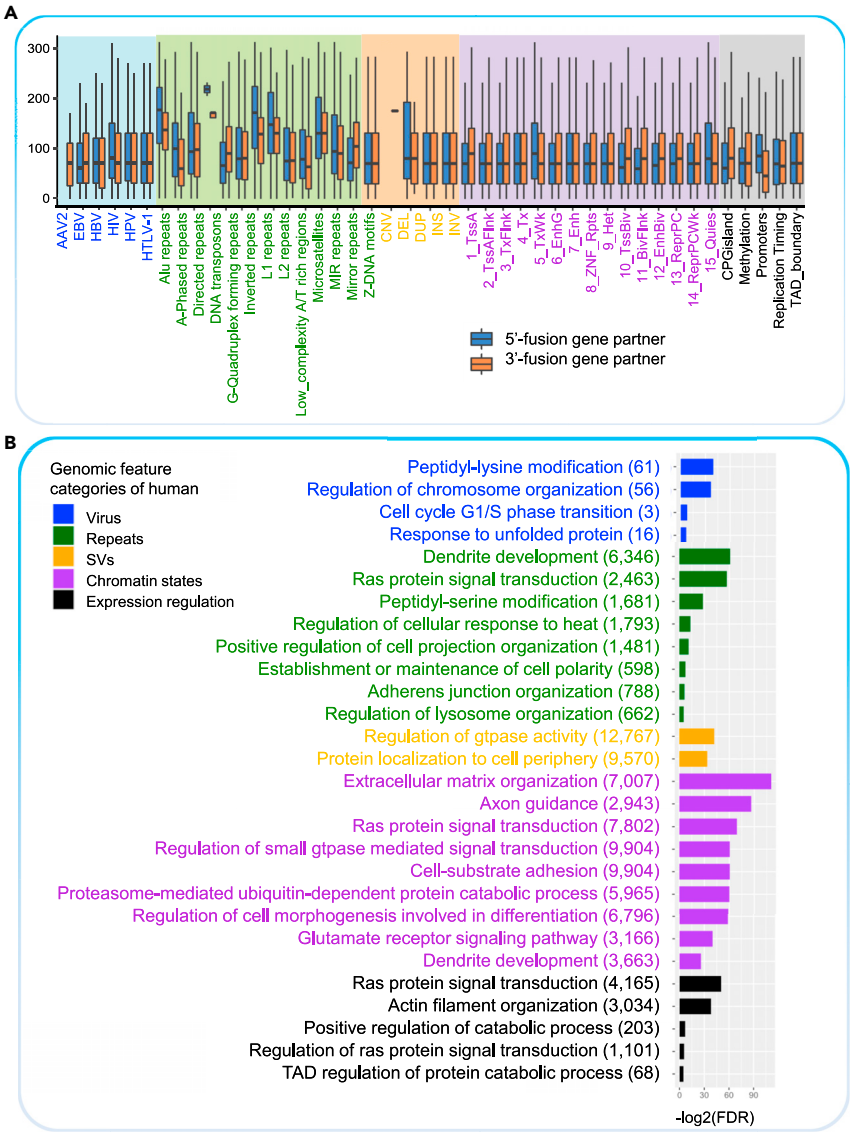


Figure 4. High feature importance scored regions

(A) Distribution of the distance between the high FI scored regions and the exon junctional breakpoints.

(B) Enriched biological processes in the genes that have overlap with high FI scored regions per individual genomic feature categories.

heterogeneity, and has become one indicator for the occurrence, development, and poor prognosis of many diseases (Zhang et al., 2020).

Overall, the distance between the top 1% feature importance scored regions and breakpoints was 70 nt of the median, 99.54 nt of mean with 211.28 nt of standard deviations (SDs) as shown in Figure 4A. This statistic seems to explain the outperformance of the FusionAI-200nt model among others. More interpretations of the individual features are described below. For the gene list functional enrichment with biological processes of Gene Ontology through ToppGene Suite (Chen et al., 2009), we selected the genes that have the overlap between top 1% feature importance scored regions and individual genomic features (Table S6).

Virus integration sites

Specifically, fusion-positive breakpoints were enriched in the virus integration sites of hepatitis B virus (HBV) and human immunodeficiency virus (HIV) (Figure 3Di). Gene ontology enrichment test identified

that the high FI scored region genes overlapped with the HIV integration sites were enriched in "peptidyl-lysine modification" and "regulation of chromosome organization" (Figure 4B). Multiple post-translational modifications (PTMs) of viral and cellular proteins gain increasing attention as modifying enzymes regulate virtually every step of the viral replication cycle (Chen et al., 2018). For HBV, "cell cycle G1/S phase transition" and "response to unfolded protein" were the enriched biological pathways. A hallmark of chronic HBV infection is known as containing excessive hepatitis surface antigen (HBsAg) in the ER which is linked to unfolded protein response (Li et al., 2019). A previous study showed that HBV-infected primary human hepatocytes are enriched in the G2/M phase compared to the predominantly G0/G1 phase of cultured primary human hepatocytes (Xia et al., 2018).

Repeats

Next, fusion-positive breakpoints were enriched in multiple types of repeats as shown in green background plots of Figure 3Di. Among 13 types of repeats, Alu repeats, direct repeats, and L1 repeats showed very significantly different distribution between fusion-positive and -negative groups. Alu elements are the most abundant transposable elements, containing over 1.2 million copies, which comprise 11% of the human genome (Deininger, 2011). It is reported as enriched in the common fragile sites (Fungtammasan et al., 2012). Direct repeats are known for eliciting genetic instability by both exploiting and eluding DNA double-strand break repair systems in mycobacteria (Wojcik et al., 2012). The human LINE-1 retrotransposon is known to create DNA double-strand breaks (Gasior et al., 2006). There was no specific difference in the structural variants compared to other features. This might be explained by that our training dataset is made up with the RNA-seq based exon junction breakpoints, which is different from WGS-based structural variants.

Chromatin states

The purple background plots are the overlaps with the chromatin state calls, using a 15-state model from Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics et al., 2015). As shown in the first four plots of the purple background ones in the right panel of Figure 3Di, the top 1% feature importance scored fusion-positive breakpoint area of 5′-genes enriched in the active chromatin states, which were associated with the expressed genes such as active transcription start site (Tss) proximal promoter states (TssA, TssAFlnk), a transcribed state at the 5′ and -3′ end of genes showing both promoter and enhancer signatures (TxFlnk), actively transcribed states (Tx). However, the fusion-negatives were relatively more enriched with the high FI scored regions related to the repressed chromatin states. In other words, the breakpoints of fusion-positive breakpoints are located at the transcriptionally active chromatin states' peak regions, but the ones of fusion-negatives are located at the transcriptionally repressed chromatin states' peak regions. This pattern might be related to the typical roles of the driver fusion genes as the transcriptional activation itself or downstream target genes (Kim et al., 2017, 2018, 2020). This makes sense since the 5′-gene partner's promoters will be used as the promoter of the fusion genes.

Gene expression regulatory

The first plot in the last category of gene expression regulatory with gray color background shows the more breakpoints of 5′-genes (3,684) are located in the CpG island area than 3′-genes (678) (Figure 3Di). This might provide additional evidence for the previous finding that initial chromosomal breakage occurs directly at or near CpGs (Tsai et al., 2008). Four thousand one hundred sixty five genes in these regions were enriched in the "Ras protein signal transduction" pathway. The imbalance of the Ras signaling pathway is a major hallmark of human cancer (Irimia et al., 2004). Furthermore, 3,849 of 4,165 genes were mainly targeted by MAX interactor 1, dimerization protein (MXI1), and enhancer of zeste 2 polycomb repressive complex 2 subunit (EZH2). MXI1 is the MYC antagonist, also regarded as a tumor suppressor. EZH2 is the enzymatic subunit of polycomb repressive complex 2 (PRC2), a complex that methylates lysine 27 of histone H3 (H3K27) to promote transcriptional silencing (Kim and Roberts, 2016). EZH2 is known to regulate MXI1 by targeting from the harmonizome, a collection of processed data sets gathered to serve and mine knowledge about genes and proteins (Rouillard et al., 2016). Through the DNA breakage of fusion genes, these CpGs located genes might be expected to have aberrant regulation governed by EXH2. The last feature, replication timing has been notified as being associated with the nature of chromosomal rearrangements in cancer (Du et al., 2019). Similar to CpGisland, there were 1,101 genes enriched in the "regulation of ras protein signal transduction" pathway. In the methylation and TAD_boundary features, 203 and 68 genes were enriched in the "positive regulation of catabolic process" and "regulation of protein catabolic process". Overall, this analysis identified enriched biological pathways of ras signal transduction and cellular regulation of catabolic processes in the gene expression regulatory that share

the genomic breakages. From this, we can infer the involvement of genomic breakage in the cancer cell proliferation or maturation, and metabolic reprogramming.

The GC-rich motif is enriched in the breakpoint area of transcription factor fusion genes

We wondered what DNA motif sequences exist in high FI scored regions, which are located about 100 nt ± 200 SD apart from the exon junction breakpoint. We made a fasta file composed of high FI score regions' DNA sequences of all fusion-positive breakpoints in training data. Then, we ran MEME to find the motifs among our input sequences with a maximum number of motifs of 100 and a threshold of E-value of 0.05. Then, there were 21 motifs that were located in more than 1000 high FI score region sequences. Here, the E-value of a motif is based on its log likelihood ratio, width, sites, background letter frequencies, and the size of the training set in their explanation (Bailey et al., 2015) (Figure 5A). From there, we further investigated the top three motifs of "CCGCCGCCGCC", "TTMWTTTTTTTTTTTTTTYYT" and "CGGCGG CGGCGGCGGCGGCGG" E-value of 2.1×10^{-165}, 9.3×10^{-244}, and $1.9 \times 10^{-}140$ (sorted by GOMO's criteria), respectively. Using Gene Ontology for Motifs (GOMOs) tool, we scanned all promoters using these three motifs individually and identified the targeted genes (Table S7). Then, the genes that have "CCGCCGCCGCC" motif in their promoter regions were enriched in these biological processes such as "insulin receptor signaling pathway", "activation of protein kinase A activity", and "Wnt receptor signaling pathway, calcium modulating pathway", based on the criteria (E-value $\leq$ 0.05 and q-value $\leq$ 0.01 and specificity = 100% from the GO tree). "TTMWTTTTTTTTTTTTTTYYT" motif was enriched in the genes involved in these biological processes, such as "negative regulation of alpha-beta T cell differentiation", "transcription initiation from RNA polymerase II promoter", "DNA damage checkpoint". Last, the genes with CGGCGGCGGCGGCGGCGGCGG motif in the promoter regions were enriched in these biological pathways like "activation of protein kinase A activity", "insulin receptor signaling pathway", and "positive regulation of phosphoinositide 3-kinase activity".

Because 2 of 3 known fusion genes are intra-chromosomal rearrangement events, we also searched DNA motif sequences focusing on these events. Then, there was only one motif CCGCSGCCGCCSC SGCCSCCGCCGCCGCCSCCSCCSCCGCC, which was enriched in these biological processes – "activation of protein kinase A activity", "protein amino acid autophosphorylation", and "insulin receptor signaling pathway". In terms of the clinical view of fusion genes, to date, there are two types of the most frequent driver fusion genes like kinase fusion genes and transcription factor fusion genes for tumorigenesis (Kim et al., 2017, 2018). Since the major working mechanism of enhanced proliferation through fusion gene is working based on the combination between dimerization domain at 5′-partner gene and kinase domain at 3′-partner gene, we investigated the motif sequence in the high feature importance regions of the fusion genes between the genes that have these domain combinations. Then we identified the "HTTTTBTTTTT" motif from 389 and this motif was enriched in the "negative regulation of alpha-beta T cell differentiation" pathway with an E-value of 9.3×10^{-3}. Next, from the high FI score regions of 322 transcription factor fusion genes, we found a relatively long (28 nt) GC-rich motif, GCBGGSSGSGGSSGSSSGGGGSGSBGGG (Table S8). The genes with this motif in their promoter regions are mainly involved in "insulin receptor signaling pathway", "Wnt receptor signaling pathway, calcium modulating pathway", "positive regulation of phosphoinositide 3-kinase activity", "positive regulation of transforming growth factor β receptor signaling pathway", "positive regulation of osteoblast differentiation". For example, Figure 5B shows the distribution of this GC-rich motif sequence across 20kb sequence of *TMPRSS2-ERG* fusion gene (Table S9). Interestingly, this GC-rich motif sequence was located around the breakpoint of the 5′-gene. *TMPRSS2-ERG* is known as dependent on the expression of Sp1 (Meisel Sharon et al., 2016), which is a transcription factor binding to GC-rich motifs (Koutsodontis et al., 2002). Figure 5C shows 22 recurrent transcription factor fusion genes, which are expressed in more than 3 samples. Specifically, 17 fusion genes of these 22 were bound by Sp1 as shown in the searching result from ENCODE transcription factor target database and TRRUST (Consortium, 2011; Han et al., 2018). In total, 220 genes involved in transcription factor fusion genes have this GC-rich motif. Of these, 51 and 38 genes were transcription factors and epigenetic factors, respectively. Unique 80 genes of these two group genes were annotated as mainly involved in the diverse epigenetic mechanisms as shown in Figure 5D such as "DNA methylation and alkylation", "histone modification", "transcription regulation", "cellular response to ROS", and etc. From this result, we may infer that the transcription factor fusion genes might regulate their downstream targets by using epigenetic mechanisms.

Chimeric transcripts from the non-disease tissues have a low FusionAI score

Fusion genes formed by the chromosomal rearrangements are the hallmark of cancers from genomic instability. However, chimeric transcripts are also existing in non-cancerous cells and tissues

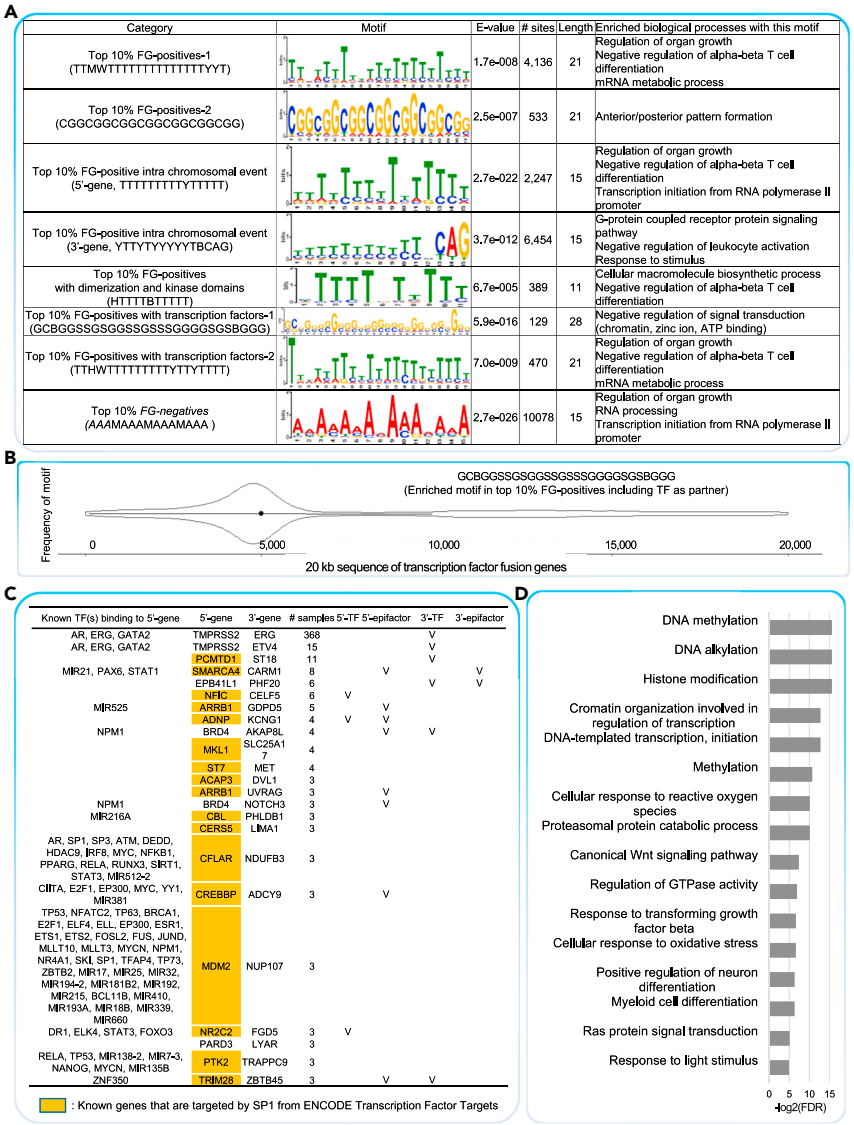


Figure 5. Consensus motif sequences in the high FI scored FG-positive regions and enriched biological processes.

(A) Identified DNA sequence motifs in fusion-positive breakpoint area of multiple groups such as all fusion-positives, intra-chromosomal events of fusion-positives, kinase fusion genes with dimerization and kinase domain, transcription factor fusion genes.

(B) Distribution of the GC-rich motif across 20 kb length sequence in the isoforms of TMPRSS2-ERG fusion gene.

(C) Transcription factor fusion genes that have GC-rich motifs.

(D) Enriched biological processes of those genes that have GC-rich motifs.

(Babiceanu et al., 2016; Finta and Zaphiropoulos, 2002; Li et al., 2009; Yuan et al., 2013). We wondered how FusionAI output scores would reflect this difference of the sample sources of fusion genes. To compare the prediction of FusionAI from diverse cohorts, we integrated the fusion genes whose breakpoints are located at the exon junction-junction loci from TCGA, cancer cell-lines, and Sanger transcripts from Entrez, GTEx, genomAD (Tables 1 and S2). We compared the ratio of the fusion genes whose FusionAI score is greater than or equal to 0.5. Here GTEx representing the healthy population was used to see how FusionAI prediction is different between cancerous and non-cancerous samples. Overall, fusion genes of the cancer cohorts, the top four groups in Table 1, had FusionAI output scores bigger than 0.5 in 93% of fusion gene breakpoints. However, if the fusion gene was common or exists in a healthy population only, then the ratio of the fusion genes with exon junctional breakpoints as predicted fusion-positive by FusionAI was

Table 1. Comparison of FusionAI scores among the fusion genes in pan-cancer and healthy tissues

Dataset	Desc.	Sequencing type	# jj fusion genes	# jj fusion genes (FusionAI score >0.5)	Percentage (%)
TCGA	Fusion genes in training data	RNA-seq	18,210	18,207	99.98
TCGA	Fusion genes in test data	RNA-seq	7,759	7,383	95.15
Klijin et al.	Fusion genes from cancer cell-lines	RNA-seq	2,162	2,066	95.56
ChiTaRS 3	Fusion transcripts	Sanger sequencing	862	807	93.62
GTEx	Fusion genes common in cancer	RNA-seq	646	537	83.13
GTEx	Fusion genes not in cancer	RNA-seq	925	634	68.54
genomAD	WGS based predicted fusions	Whole genome sequencing	923	46	4.98

decreased to 83% and 64%, respectively. Currently, we do not fully understand the meaning of this different ratio, but we guess this difference might be reflecting a validity or tumorigenicity of individual fusion gene breakpoint. genomAD, the whole genome-based structural variant data from diverse clinical cohorts, was used to see how FusionAI prediction is different between RNA-seq or WGS-based fusion genes. From the previous study for visualizing the functional features of fusion genes at four different levels, FGviewer, we found 1,037 potential fusion genes whose breakpoints of the structural variants detected from genomAD v2.1 of 15K population WGS data were located on the gene bodies. Of 1,037 fusion genes, 823 were the cases that have the exon junction-junction breakpoints. Among these, only 46 cases have been predicted as fusion-positive candidates by FusionAI (5%). Here, we used fusion genes anticipated as having exon junction-junction breakpoints from structural variants for genomAD. The small ratio might be from the different sequencing types of data and not expressed as the fusion transcript.

DISCUSSIONS

Our study suggests that FusionAI predicts fusion gene breakage with high specificity and expands the findings beyond a conventional RNA-seq fusion gene analysis by combining deep-learning predictions with empirical evidence in user-specific RNA-seq data. From the study on high feature importance scored regions, we found that the overall distance between the high FI scores and breakpoints was 70 nt median, 99.54 nt mean with 211.28 nt standard deviations. This might explain one of the reasons why the FusionAI-200nt model outperformed other models using different flanking sequence lengths.

Before finalizing the FusionAI model using the exon junction-junction breakpoints of both fusion-positive and -negative groups, we tried to build the initial version model by comparing the 20 Kbp sequence of fusion-positive breakpoints located at exon junction-junction boundaries versus one of fusion-negative breakpoints located in any location of the gene body (any regions of exon and intron). This initial model showed better performance than the current finalized model with an accuracy of 99.7% and 98.2% for training and test data, respectively. However, we recognized a potential issue regarding the design of the comparison since this model can highlight and give more weights on the exon junction regions (exon-intron boundary) compared with the intronic breakpoints. The learned features from the initial version model might mostly be the ones that are significantly related to the exon junctions like splicing signals. To avoid this wrong conclusion, we redesigned our modeling using only exon junction-junction breakpoints for both fusion-positive and -negative data and finalized FusionAI. Even though we are comparing the same conditions of exon junction-junction data, FusionAI found fusion-positive breakpoints well. In the future, enough validated data of the chimeric transcripts with specific breakpoints formed by the

trans-splicing mechanism in the RNA level would better explain the detailed exon junctional features that we found in this study.

To ensure the independence of the datasets and avoid overfitting, we checked the consistency by re-training our model with modified sample grouping strategies which ensured complete independence between training and test sets. First, to select the test datasets, we checked the number of fusion gene events per chromosome. We picked up the top-five chromosomes, which have the largest number of fusion events, including chr1, 11, 12, 17, and 19. For the fusion events of each chromosome out of these five chromosomes, we held out these as the test data and randomly picked 20k fusion events (10k positive +10k negative events) from the remaining fusion events as the training data set. We trained and tested using these new data sets. The average accuracy for the new training and test data sets was 82.9% (AUROC = 0.904) and 81.6% (AUROC = 0.889), respectively (Table S10). As shown in this result, our model per chromosome shows a consistent result.

Our trial for comparing the fusion gene breakpoints of different disease state, such as cancer versus normal, found that FusionAI prediction gives higher scores for the cancerous fusion breakpoints. If there are a plentiful number of fusion gene breakpoints at the DNA level that will be accumulated in the future, we will be able to build the genomic level breakpoint-based model and can bring deeper insight into the breakage mechanisms. Having enough amount of the real genomic breakpoints of the massive fusion genes to perform artificial intelligence approaches will need a long time and many efforts. We hope we can make more precise new models based on the genomic breakpoint-based training data, not by exon junctional breakpoints in near future.

We identified a high GC motif around the breakpoint area of the transcription factor fusion genes. To do this we investigated the most frequent DNA motif sequences based on the genomic breakpoint DNA sequence of hg19. To ensure this motif from the most recent human genome GRCh38, we also searched the consensus DNA motif sequences from the 20 Kbp sequence at the breakpoints after lifting over from hg19 to GRh38. Then, we identified multiple high GC motifs around the high feature importance scored regions of 330 TF fusion genes (Table S11), which shows the consistency and reliability of our findings based on the hg19 version human genome.

Our study used the deep learning method to better understand the genomic breakage context focused on the fusion genes, which are the highlighted ones as expressed structural variants. Much work remains to be done to understand the human genome breakage in diverse diseases, a greater understanding of the genomic breakage mechanisms could pave the way for novel candidates for therapeutic intervention. Our findings using the FusionAI model would enhance our understanding of the fusion gene context. We hope FusionAI could serve as the initial platform for the efficient investigation of genomic breakage events.

Limitations of the study

If we had enough data of real genomic breakpoint information, then we could identify correct genomic breakpoint features. However, due to the limited data of fusion breakpoint information replying on mainly RNA-seq data, we could not have the real genomic breakpoint information. We hope we can build a more accurate fusion genomic breakpoint classifier in the future based on the genomic breakpoint information.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - FusionAI architecture and training using deep learning
  - Feature importance score
  - Creating simulation RNA-seq data of training and test data to run STAR-fusion and Arriba
  - Model evaluation on ChiTaRS-3.1
  - Model evaluation on 2,200 fusion genes from 520 cancer cell-lines

- Comparison with existing fusion gene prediction tools for three cell-lines
- Identification of DNA sequence motif
- Making FusionAI input data
- Human genomic features information
- Gene ontology enrichment analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103164.

ACKNOWLEDGMENTS

This work was partially supported by the National Institutes of Health grants [NIH R01GM123037, U01AR069395, and R01CA241930] to X.Z. and [R35GM138184] to P.K. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. Funding for open access charge: Startup Fund to Dr. Kim from the University of Texas Health Science Center at Houston.

AUTHOR CONTRIBUTIONS

P.K. arranged the training and test data by making fusion-positive and -negative data; made simulation RNA-seq data; performed downstream analyses of DNA sequence motif and annotated findings; and wrote the paper. H.T. performed the deep learning analysis. J.L. integrated human genomic feature information. M.Y. ran other fusion prediction tools. P.K., H.T., and X.Z. supervised the analyses and designed the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 11, 2021
Revised: July 16, 2021
Accepted: September 21, 2021
Published: October 22, 2021

REFERENCES

Akdemir, K.C., Le, V.T., Chandran, S., Li, Y., Verhaak, R.G., Beroukhim, R., Campbell, P.J., Chin, L., Dixon, J.R., and Futreal, P.A. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat. Genet. *52*, 294–305.

Avvaru, A.K., Sharma, D., Verma, A., Mishra, R.K., and Sowpati, D.T. (2020). MSDB: a comprehensive, annotated database of microsatellites. Nucleic Acids Res. *48*, D155–D159.

Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., et al. (2016). Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Res. *44*, 2859–2872.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). Meme SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME suite. Nucleic Acids Res. *43*, W39–W49.

Ballinger, T.J., Bouwman, B.A.M., Mirzazadeh, R., Garnerone, S., Crosetto, N., and Semple, C.A. (2019). Modeling double strand break susceptibility to interrogate structural variation in cancer. Genome Biol. *20*, 28.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11.

Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome Atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

Chakraborty, A., Jenjaroenpun, P., Li, J., El Hilali, S., McCulley, A., Haarer, B., Hoffman, E.A., Belak, A., Thorland, A., Hehnly, H., et al. (2020). Replication stress induces global chromosome breakage in the fragile X genome. Cell Rep. *32*, 108179.

Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305–W311.

Chen, L., Keppler, O.T., and Scholz, C. (2018). Post-translational modification-based regulation of HIV replication. Front. Microbiol. *9*, 2131.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alfoldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.

Consortium, E.P. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.

Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods Groups-Analysis Working, G., Enhancing, G.G., et al.; Fund, N.I.H.C., NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Deininger, P. (2011). Alu elements: know the SINEs. Genome Biol. *12*, 236.

Du, Q., Bert, S.A., Armstrong, N.J., Caldon, C.E., Song, J.Z., Nair, S.S., Gould, C.M., Luu, P.L., Peters, T., Khoury, A., et al. (2019). Replication

timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. Nat. Commun. *10*, 416.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492.

Finta, C., and Zaphiropoulos, P.G. (2002). Intergenic mRNA molecules resulting from trans-splicing. J. Biol. Chem. *277*, 5882–5890.

Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K.A., and Makova, K.D. (2012). A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? Genome Res. *22*, 993–1005.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. J. Mol. Biol. *357*, 1383–1393.

Gorohovski, A., Tagore, S., Palande, V., Malka, A., Raviv-Shay, D., and Frenkel-Morgenstern, M. (2017). ChiTaRS-3.1-the enhanced chimeric transcripts and RNA-seq database matched with protein-protein interactions. Nucleic Acids Res. *45*, D790–D795.

Haas, B.J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. *20*, 213.

Han, H., Cho, J.W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. *46*, D380–D386.

Irimia, M., Fraga, M.F., Sanchez-Cespedes, M., and Esteller, M. (2004). CpG island promoter hypermethylation of the Ras-effector gene NORE1A occurs in the context of a wild-type K-ras in lung cancer. Oncogene *23*, 8695–8699.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting splicing from primary sequence with deep learning. Cell *176*, 535–548.e24.

Kim, K.H., and Roberts, C.W. (2016). Targeting EZH2 in cancer. Nat. Med. *22*, 128–134.

Kim, P., Ballester, L.Y., and Zhao, Z. (2017). Domain retention in transcription factor fusion genes and its biological and clinical implications: a pan-cancer study. Oncotarget *8*, 110103–110117.

Kim, P., Jia, P., and Zhao, Z. (2018). Kinase impact assessment in the landscape of fusion genes that retain kinase domains: a pan-cancer study. Brief Bioinform. *19*, 450–460.

Kim, P., Yiya, K., and Zhou, X. (2020). FGviewer: an online visualization tool for functional features of human fusion genes. Nucleic Acids Res. *48*, W313–W320.

Kim, P., and Zhou, X. (2019). FusionGDB: fusion gene annotation DataBase. Nucleic Acids Res. *47*, D994–D1004.

Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., et al. (2015). A comprehensive transcriptional portrait of human cancer cell lines. Nat. Biotechnol. *33*, 306–312.

Koutsodontis, G., Moustakas, A., and Kardassis, D. (2002). The role of Sp1 family members, the proximal GC-rich motifs, and the upstream enhancer region in the regulation of the human cell cycle inhibitor p21WAF-1/Cip1 gene promoter. Biochemistry *41*, 12771–12784.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., et al. (2013). DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. *41*, D936–D941.

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011). The sequence read archive. Nucleic Acids Res. *39*, D19–D21.

Li, H., Wang, J., Ma, X., and Sklar, J. (2009). Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle *8*, 218–222.

Li, Y., Xia, Y., Cheng, X., Kleiner, D.E., Hewitt, S.M., Sproch, J., Li, T., Zhuang, H., and Liang, T.J. (2019). Hepatitis B surface antigen activates unfolded protein response in forming ground glass hepatocytes of chronic Hepatitis B. Viruses *11*, 386.

Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C.C., de Hoon, M., Severin, J., Oki, S., Hayashizaki, Y., et al. (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Res. *47*, D752–D758.

Meisel Sharon, S., Pozniak, Y., Geiger, T., and Werner, H. (2016). TMPRSS2-ERG fusion protein regulates insulin-like growth factor-1 receptor (IGF1R) gene expression in prostate cancer: involvement of transcription factor Sp1. Oncotarget *7*, 51375–51392.

Natale, F., Scholl, A., Rapp, A., Yu, W., Rausch, C., and Cardoso, M.C. (2018). DNA replication and repair kinetics of Alu, LINE-1 and satellite III genomic repetitive elements. Epigenetics Chromatin *11*, 61.

Navarro Gonzalez, J., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Raney, B.J., Powell, C.C., Nassar, L.R., Maulding, N.D., Lee, C.M., et al. (2021). The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. *49*, D1046–D1057.

Peng, Q., Pevzner, P.A., and Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. PLoS Comput. Biol. *2*, e14.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford) *2016*, baw100.

Singh, S., Qin, F., Kumar, S., Elfman, J., Lin, E., Pham, L.P., Yang, A., and Li, H. (2020). The landscape of chimeric RNAs in non-diseased tissues and cells. Nucleic Acids Res. *48*, 1764–1778.

Sleutels, F., and Barlow, D.P. (2002). The origins of genomic imprinting in mammals. Adv. Genet. *46*, 119–163.

Tang, D., Li, B., Xu, T., Hu, R., Tan, D., Song, X., Jia, P., and Zhao, Z. (2020). VISDB: a manually curated database of viral integration sites in the human genome. Nucleic Acids Res. *48*, D633–D641.

Tsai, A.G., Lu, H., Raghavan, S.C., Muschen, M., Hsieh, C.L., and Lieber, M.R. (2008). Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. Cell *135*, 1130–1142.

Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U., Neumann, O., Stenzinger, A., Scholl, C., et al. (2021). Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Research *31*, In this issue, 448–460. In this issue. https://doi.org/10.1101/gr.257246.119.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Wojcik, E.A., Brzostek, A., Bacolla, A., Mackiewicz, P., Vasquez, K.M., Korycka-Machala, M., Jaworski, A., and Dziadek, J. (2012). Direct and inverted repeats elicit genetic instability by both exploiting and eluding DNA double-strand break repair systems in mycobacteria. PLoS One *7*, e51064.

Xia, Y., Cheng, X., Li, Y., Valdez, K., Chen, W., and Liang, T.J. (2018). Hepatitis B virus deregulates the cell cycle to promote viral replication and a premalignant phenotype. J. Virol. *92*, e00722-18.

Yuan, H., Qin, F., Movassagh, M., Park, H., Golden, W., Xie, Z., Zhang, P., Sklar, J., and Li, H. (2013). A chimeric RNA characteristic of rhabdomyosarcoma in normal myogenesis process. Cancer Discov. *3*, 1394–1403.

Zhang, X., Zhang, R., and Yu, J. (2020). New understanding of the relevant role of LINE-1 retrotransposition in human disease and immune modulation. Front. Cell Dev. Biol. *8*, 657.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Open reading frame annotation of known fusion genes	Kim and Zhou, 2019	https://ccsm.uth.edu/FusionGDB/
Fusion gene breakpoint information of TCGA fusion genes	Kim and Zhou, 2019	https://ccsm.uth.edu/FusionGDB/
Fusion gene breakpoint information of cancer cell-lines	Klijn et al., (2015)	N/A
Fusion gene breakpoint information of Sanger sequencing	Gorohovski et al., (2017)	http://biodb.md.biu.ac.il/chitars.prv
Fusion gene breakpoint information of GTEx cohorts	Singh et al., (2020)	N/A
Fusion gene breakpoint information of genomAD cohorts	Kim et al. (2020)	https://ccsmweb.uth.edu/FGviewer
Simulation RNA-seq data of all validation sets	This study	https://compbio.uth.edu/FusionGDB2/FusionAI
Fastq files for RNA-seq of K562	Sequence Read Archive (SRA) in NCBI	Sequence Read Archive accession: SRR521460
Fastq files for RNA-seq of MCF7	Sequence Read Archive (SRA) in NCBI	Sequence Read Archive accession: SRR064286
Fastq files for RNA-seq of H2228	Sequence Read Archive (SRA) in NCBI	Sequence Read Archive accession: DRR016705.1s
Virus integration site information	Tang et al., (2020)	https://bioinfo.uth.edu/VISDB
Repeatmasker	Bao et al., (2015)	http://www.repeatmasker.org
MicroSatellite DataBase (MSDB)	Avvaru et al., (2020)	https://data.ccmb.res.in/msdb
Structural variant breakpoint information of genomAD	Lappalainen et al., (2013)	https://www.ncbi.nlm.nih.gov/dbvar
Chromatin state calls using a 15-state model	Roadmap Epigenomics et al. (2015)	N/A
Location of CpGisland, Methylation, Promoters	Lizio et al. (2019)	https://fantom.gsc.riken.jp/5
Replication timing-specific peak regions	Davis et al. (2018)	https://www.encodeproject.org
Common TAD boundaries of five human cell-lines	Akdemir et al., (2020)	N/A
TRRUST2.0	Han et al., (2018)	http://www.grnpedia.org/trrust
ENCODE Transcription Factor Targets	ENCODE Project Consortium, 2011	https://maayanlab.cloud/Harmonizome/dataset/ENCODE+Transcription+Factor+Targets
Software and algorithms		
FusionAI software	This study	https://compbio.uth.edu/FusionGDB2/FusionAI
FusionAI model training	This study	https://compbio.uth.edu/FusionGDB2/FusionAI
STAR-fusion	Haas et al. (2019)	https://github.com/STAR-Fusion/STAR-Fusion
Arriba	Uhrig et al. (2021)	https://arriba.readthedocs.io/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the Lead Contact, Dr. Pora Kim (Pora.kim@uth.tmc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The training and test data, and simulation RNA-seq data are available on https://compbio.uth.edu/FusionGDB2/FusionAI.

FusionAI model and preprocessing codes available on https://compbio.uth.edu/FusionGDB2/FusionAI.

Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

METHOD DETAILS

FusionAI architecture and training using deep learning

To train the FusionAI DNN, we downloaded the fusion gene breakpoint information of 48K of fusion genes from FusionGDB. Since most of the fusion genes are predicted from the split reads of RNA-seq data and the real genomic breakpoints would be located in the intron, we used the sequence of known fusion genes that have the exon junction-junction breakpoints to train the FusionAI model. Out of ~48K known fusion events, there were ~33K fusion genes from the TCGA cohort and ~26K fusions had the breakpoints at the exon junction-junction position (j-j BP combination). To make fusion negative breakpoints data, we excluded 17,110 genes, which are involved in 48K known human fusion genes, among 43K GENCODE genes. From the rest of those genes (27,116), which are not known as involved in any fusion genes, we randomly chose two genes as fusion partners. Then, we filtered out potential false cases from the unnecessary multiply mapped cases and breakpoints belong to the repeat region, paralogs, or pseudogenes using RepeatMasker, Duplicated Genes Database, and HUGO database's pseudogenes. This is the typical pre-process by the fusion prediction tools to filter out false positives. We also excluded the gene pairs with neighboring gene relationships to exclude the potential read-through cases. In the case of the intra-chromosomal fusion genes, we set the minimum distance as 100Kbp between randomly chosen two breakpoints across gene bodies. Then, from the chosen two breakpoints after several filtering steps, a 20Kbp long DNA sequence was made by conjugating $\pm$ 5K bp sequence from each BP of two partner genes. Through this procedure, we created ~26K fusion-negative breakpoint sequence data. Based on these DNA sequences, we trained a multiple-layer DNN. The input of the model is a sequence of 20 kb one-hot encoded nucleotides. The output is two probabilities corresponding to fusion-positive and -negative breakpoints that sum to one. Our DNN consists of two convolutional layers with filter sizes (20, 4) and (200, 1), one max pooling, one flatten, and two dense layers preceding the output layer. The model involves 2,672,002 parameters including both weight matrix and bias at related layers (Figure S3). 36.4K BPs (~70%) from a combined total of 52K BPs (26K j-j combination BPs and 26K non-FGBPs) were used in the training step (further divided to 80% for training and 20% for validation), and the rest 15.6K BPs (~30%) was used for an independent test. The performance (accuracy and loss) during the training process is illustrated in Figure S4. We then tested the trained model on both the 26K original training samples and the 15.6K test samples. The accuracies for training and test datasets were 97.4% (AUROC = 0.9962) and 90.8% (AUROC = 0.9706) with 0.12 and 0.42 error rates, respectively. This performance is much better than the traditional machine learning method, SVM, which yielded an accuracy of 79% and 72% for training and test data, respectively (Figure 1D).

Feature importance score

To calculate the feature importance score (FIS), we masked 20 bp each time by setting all the 20 values to zero and measured the change of prediction outcome upon this masking. We slide this 20 bp window one nucleotide each time (i.e., stride = 1) along the whole 20K input sequence and repeated the procedure to obtain the FIS for all the 20bp segments. In this way, we got 20,000-20 + 1 = 19,981 FIS for each input sequence.

Creating simulation RNA-seq data of training and test data to run STAR-fusion and Arriba

We made simulation RNA-seq data of the split reads at the exon junction-junction breakpoints with different read lengths (50, 75, and 100 bp) and a different number of split reads (1, 3, 5 split reads, and 10 random around breakpoints) based on the fusion-positive and -negative breakpoints in training and

test datasets. Using random module of python, we chose random numbers once, three, and five times based on the seed length of 25bp as the transcript's broken position with 0 to 5 varied distance between the 5′-genes' exon sequence to the breakpoint and the 3′-genes' exon sequence from the breakpoint to make the split reads at the exon junction site with different read length. We also made 10 random split read sequences with a 10bp distance gap among the read alignments.

Model evaluation on ChiTaRS-3.1

We downloaded the fusion gene information from ChiTaRS-3.1 (Gorohovski et al., 2017). Among these, we only used the validated fusion genes by the Sanger sequencing approach, which is the typical way of validation of identified fusion genes from the Entrez transcript database by the National Center for Biotechnology Information (NCBI). Among these, 862 fusion genes had the breakpoints at the exon junction-junction positions. For these cases, we made a 20kb long DNA sequence as the input of FusionAI and ran it. Furthermore, to compare the performance with STAR-fusion and Arriba, we also made the simulation RNA-seq data of the split reads at the exon junction breakpoints with different read lengths and the different numbers of split reads. We used the default option based on GENCODE v19 genome for running STAR-fusion and Arriba.

Model evaluation on 2,200 fusion genes from 520 cancer cell-lines

We downloaded the fusion gene information of the 2,269 validated in-frame fusion genes from 529 cancer cell-lines by Klijn et al. (2015). Out of these, 2,162 fusion genes had the breakpoints at the exon junction-junction positions. For these cases, we made a 20kb long DNA sequence as the input of FusionAI and ran it. To test STAR-fusion and Arriba, we also made the simulation RNA-seq data the same way we did for evaluation on ChiTaRS-3.1 data as the input data above.

Comparison with existing fusion gene prediction tools for three cell-lines

K562 is a myelogenous leukemia cell-line with the most famous fusion gene, *BCR-ABL1*. MCF7 is the most studied breast cancer cell-line with multiple identified fusion genes. H2228 is the non-small cell lung cancer cell-line with the *EML4-ALK* fusion gene. We ran STAR-fusion and Arriba for these cell-lines' RNA-seq data which were downloaded from the Sequence Read Archive (SRA) of NCBI (Leinonen et al., 2011) with SRA accession of SRR521460, SRR064286, and DRR016705.1 for K562, MCF7, and H2228, respectively. To run FusionAI, we made 20kb long DNA sequences based on the breakpoints that were predicted by STAR-fusion and Arriba. For validation, we also made FusionAI input data for the experimentally validated fusion genes among these three cells from the work by Klijn et al.

Identification of DNA sequence motif

First, we assembled the sequence of the top 1% feature importance scored regions into the merged sequence contigs and saved them fasta format file was after checking a continuous alignment. Then, to identify the most frequent DNA sequence motifs, we used MEME by MEME Suite (Bailey et al., 2009) with the 'any number of repetitions' option and identified enriched DNA sequence motifs around the broken regions by fusion genes for individual interested fusion gene groups such as all fusion-positives, intra-chromosomal fusion-positives, kinase fusion genes, and transcription factor fusion genes. We used the default threshold of p-value, 1E-4 as used in MEME Suite. After identifying the most frequent motifs, we ran GOMO to identify the enriched GO biological process of the genes that have the binding sites of our finding motifs in their promoters.

Making FusionAI input data

FusionAI runs based on the DNA sequence composed of two genes involved in a fusion gene with $\pm$ 5 kb flanking sequence from each breakpoint. The pre-processing python script of the FusionAI package provides for the user to make the input sequence data of FusionAI from the fusion breakpoint information. FusionAI package makes the input sequence data using the nibFrag based on the human genome sequence of the hg19 version, which can be downloaded from the UCSC Genome Browser. The tab-delimited data format with fusion gene pairs, chromosome, breakpoint, strand, and input sequence is read by the FusionAI model and FusionAI gives the prediction score. In the FusionAI package, the user can make the input data using Run_FusionAI.py based on the interested breakpoint information.

Human genomic features information

We integrated loci information of different types of human genomic features across five big categories including virus integration sites, repeats, structural variants, chromatin states, and gene expression regulation. First, we downloaded the virus integration site information from the Viral Integration Site DataBase (VISDB) (Tang et al., 2020) and we lifted it over to the hg19 version using the liftover tool from the UCSC Genome Browser since FusionAI's training was done based on the sequence of the hg19 version (Navarro Gonzalez et al., 2021). Except for this virus information, all genomic coordinates are based on hg19 version. We integrated 13 types of repeats (Alu repeats, A-Phased repeats, Directed repeats, DNA transposons, "G-Quadruplex, forming repeats", Inverted repeats, L1 repeats, L2 repeats, "Low_complexity, A/T rich regions", Microsatellites, MIR repeats, Mirror repeats, and Z-DNA motifs) from RepeatMasker (Bao et al., 2015) and MicroSatellite DataBase (MSDB) (Avvaru et al., 2020). For the diverse types of structural variants including the copy number variants, we downloaded the arranged breakpoint information of the structural variants from dbVar (Lappalainen et al., 2013). The chromatin states category include the loci of 15 different types of chromatin states such as 1_TssA, 2_TssAFlnk, 3_TxFlnk, 4_Tx, 5_TxWk, 6_EnhG, 7_Enh, 8_ZNF_Rpts, 9_Het, 10_TssBiv, 11_BivFlnk, 12_EnhBiv, 13_ReprPC, 14_ReprPCWk, and 15_Quies, from the previous study on the chromatin state calls using a 15-state model for 12 cell lines, were obtained from the Roadmap Epigenomics Mapping Consortium (Ernst and Kellis, 2017; Roadmap Epigenomics et al., 2015). The gene expression regulatory category includes five types of features as CPGisland, Methylation, Promoters, ReplicationTiming, and TAD boundaries. The information of the first three feature categories was downloaded from the FANTOM5 collection (Lizio et al., 2019). We downloaded the replication timing-specific peak regions from the ENCODE portal site by selecting the assay type of the replication timing (Davis et al., 2018). We used 2,477 loci of common TAD boundaries from a previous study that made high-resolution chromosome conformation (Hi-C) datasets from five human cell lines based on the (Akdemir et al., 2020). The detailed statistics for individual feature categories with their significance results by Chi-square test are in Table S5.

Gene ontology enrichment analysis

To identify the enriched biological processes in the overlapped genes between the top 1% FI scored regions and individual human genomic features of 44 categories, we used ToppFun of the ToppGene Suite (Chen et al., 2009). We limited the results by the Benjamini and Hochberg false discovery rate < 0.05 and the number of genes in the gene group < 500. We showed the top enriched GO biological process of each feature category in Figure 4B.