Cell Reports Methods, Volume *4*

Supplemental information

Anchored-fusion enables targeted fusion search in bulk and single-cell RNA sequencing data

Xilu Yuan, Haishuai Wang, Zhongquan Sun, Chunpeng Zhou, Simon Chong Chu, Jiajun Bu, and Ning Shen

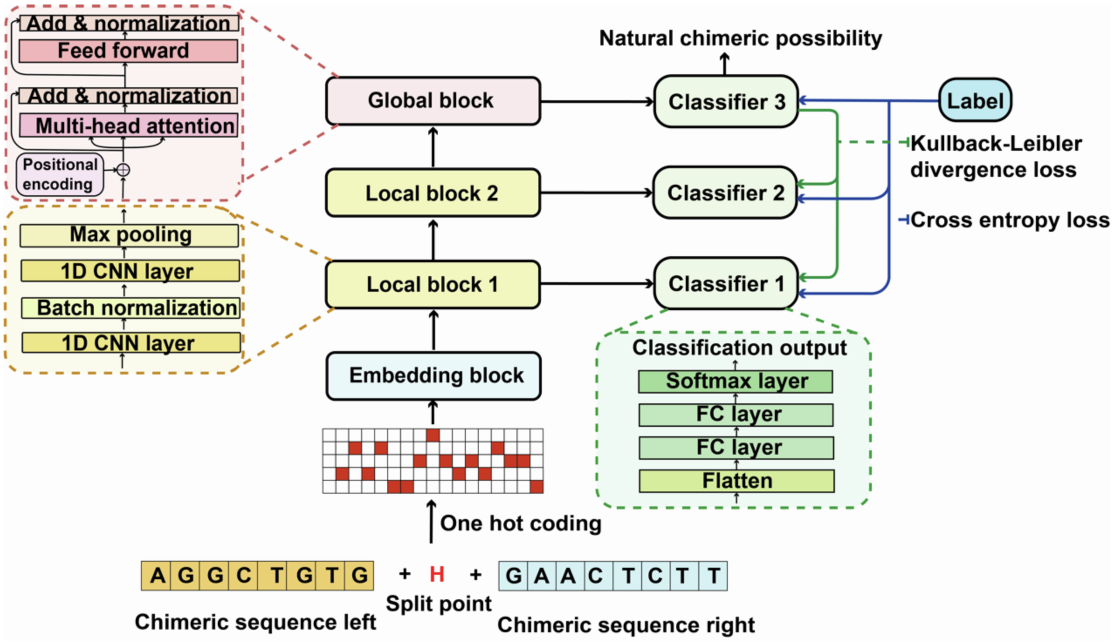

Figure S1. [Details of Hierarchical View Learning and Distillation Model] Related to Figure 1

Our HVLD model takes chimeric sequences as input, with chimeric breakpoints marked by a special character 'H'. These sequences are encoded into one-hot form through an embedding layer. The output of the embedding layer is then processed by two local blocks based on 1D convolution to extract local sequence information, followed by a global block based on multi-head attention to extract global information. After each block, a classifier is connected to output the classification probability of the sequence. Self-distillation learning is utilized to make the output of shallow classifiers closer to that of the deep one, thereby improving the classification ability of the model. Only the output of the deepest classifier is used as the final classification result of the model.

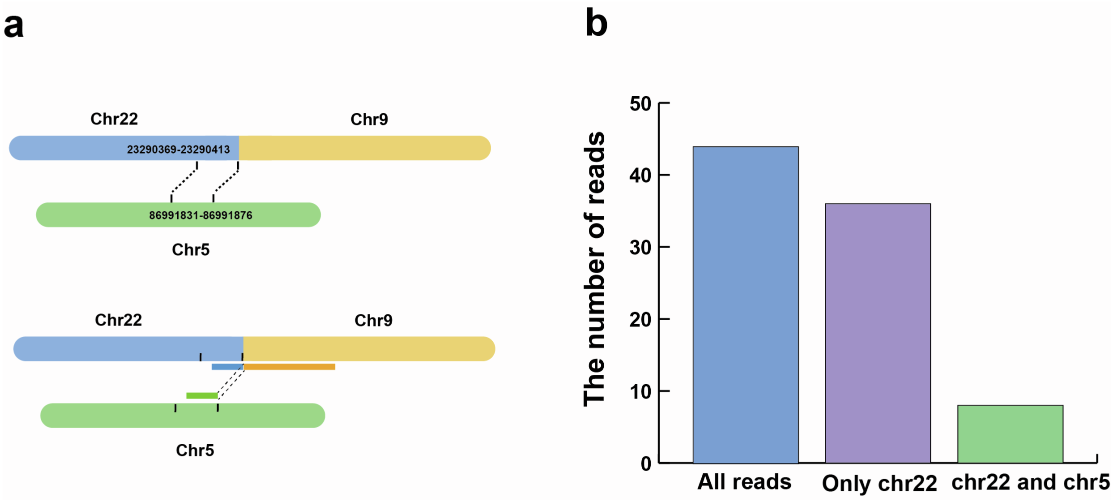

Figure S2. [The BCR-ABL1 breakpoint on chromosome 22 is located in a short homologous sequence of chromosome 5] Related to Figure 1

(a). The presence of a short homologous sequence of chromosome 22 at the BCR-ABL1 breakpoint on chromosome 5 results in some BCR-ABL1 fusion fragments having their BCR portion aligned to both the homologous regions of chromosome 22 and chromosome 5, leading to their exclusion. (b). We selected 250bp regions upstream and downstream of the BCR-ABL1 fusion transcript site and simulated the generation of 100 pairs of 100bp reads. Among them, 44 reads contained the fusion site. Out of these, only 36 reads were aligned solely to chromosome 22 and considered support reads, while 8 reads that matched both chromosome 22 and chromosome 5 were filtered out.

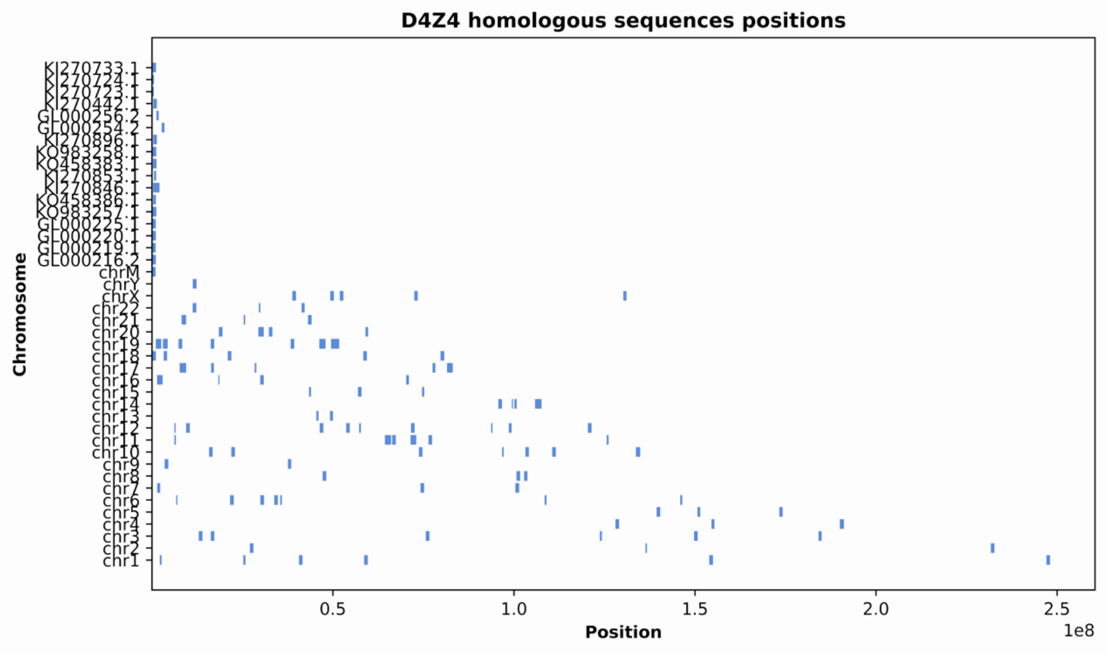

Figure S3. [The distribution of the D4Z4 segment of the DUX4 gene in the genome] Related to Figure 1

To facilitate observation, we have magnified the length of each D4Z4 copy by 1000 times. The DUX4 gene has hundreds of copies across the chromosome set. In order to address the issue of multiple copies of the DUX4 gene, fusion gene detection tools such as STAR-Fusion based on RNA-seq specifically align reads that can map to the DUX4 gene to a region on chromosome 4, specifically 190173774-190175845, while excluding any homologous segments from alignment.

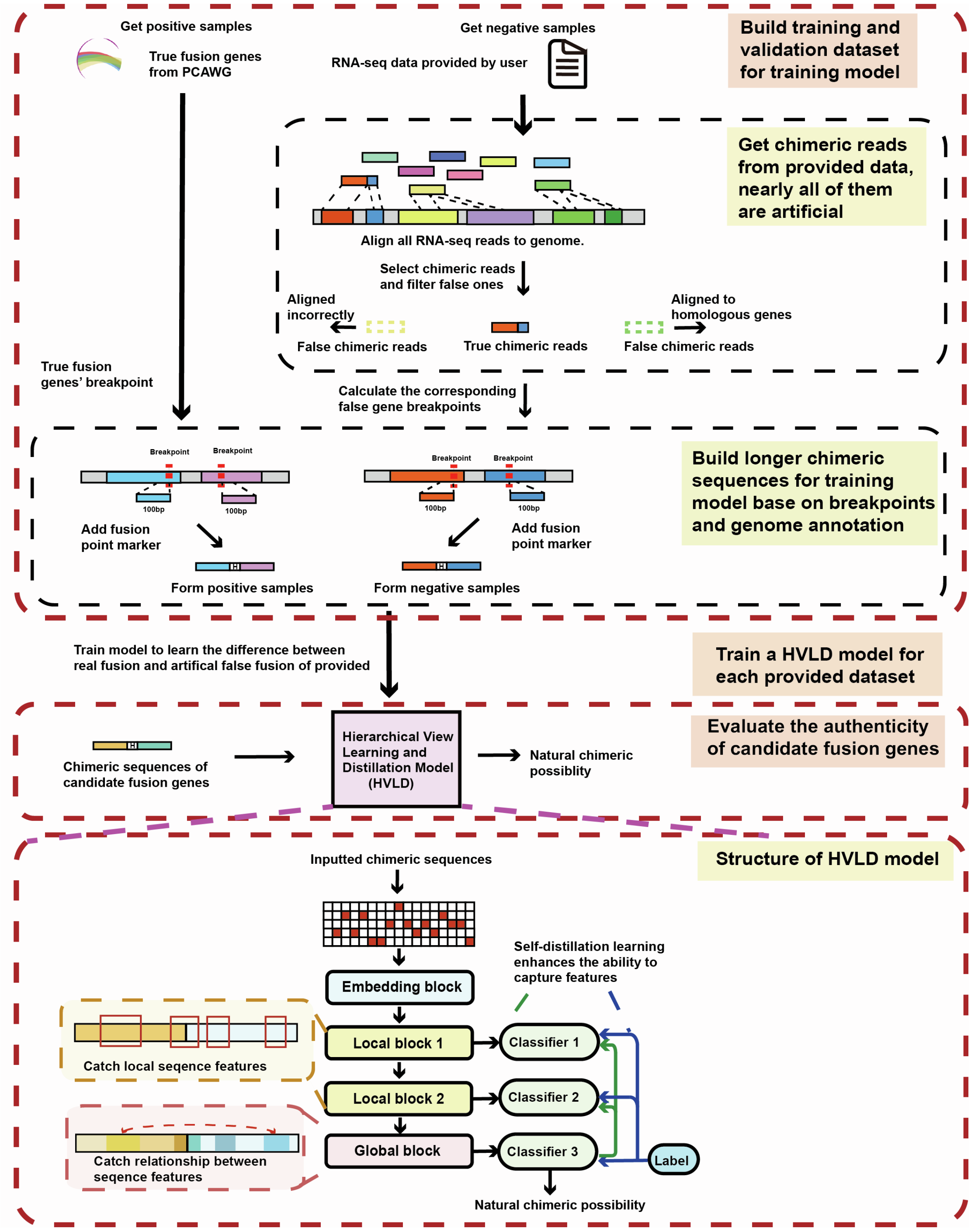

Figure S4. [The process of obtaining positive and negative samples for training and testing the HVLD model] Related to Figure 2 and STAR Methods

This process includes getting chimeric reads from the dataset inputted by user, building positive and negative input sequences with fusion breakpoints and genome annotation. Then these samples are used to train and test HVLD model.

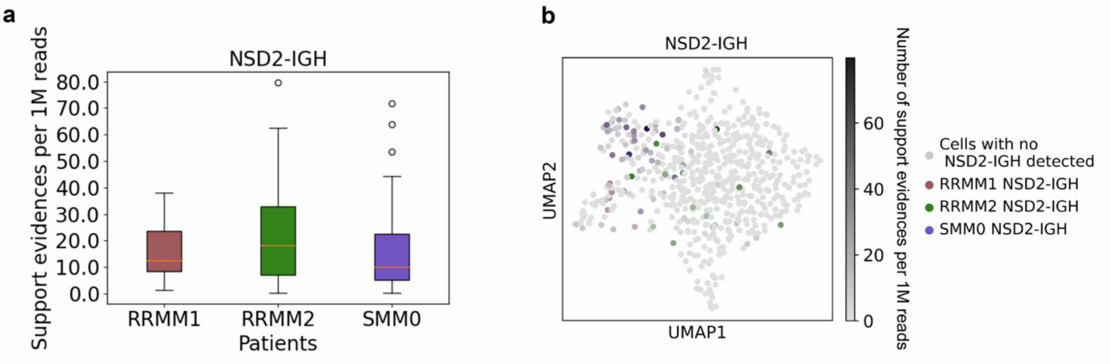

Figure S5. [Distribution of NSD2-IGH fusion gene supporting evidences in three positive patients] Related to Figure 6

(a) A box plot is used to represent the distribution of the number of supporting evidences per one million reads among three NSD2-IGH positive patients. The result showed that the average number of supporting evidences in patients RRMM1 and SMM0 was lower than that in patient RRMM2, and the variance was also smaller. According to the research conducted by Jang et al., patient RRMM2 has both NSD2-IGH fusion and trisomy 11,15, while patients RRMM1 and SMM0 only have NSD2-IGH. Additionally, the gene expression of single cells from patients RRMM1 and SMM0 is more similar to each other, while RRMM2 is distinct from them[1]. The difference in NSD2-IGH expression among these patients may explain their other varying manifestations. (b) The shade of color represents the number of support evidences per 1M reads in every single cell. The three different colors represent the three patients with the NSD2-IGH fusion gene detected by Anchored-fusion.

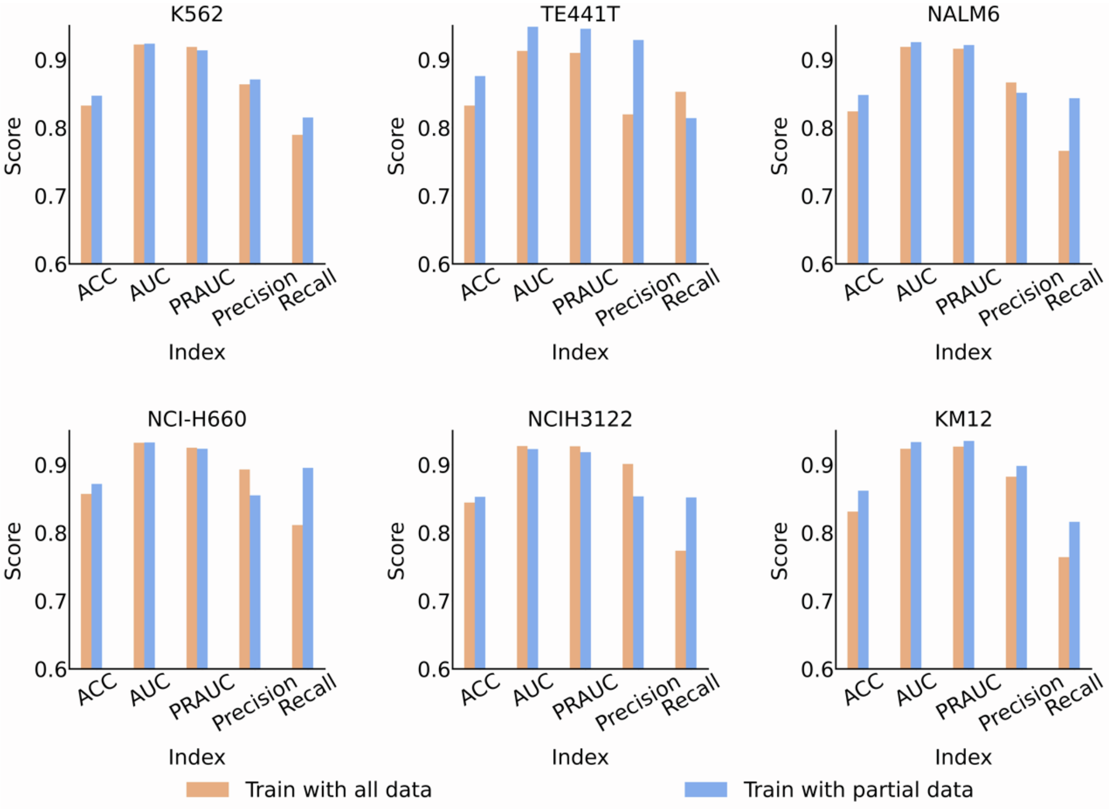

Figure S6. [Performance trained with all artificial fusion samples and randomly choosing a subset as negative examples respectively] Figure 2 and STAR Methods

First, we separated a portion of the negative examples from all artificial fusion samples as the test set. Next, we selected a subset from all artificial fusion samples, which was equal in number to the positive examples, to serve as the training set. On the other hand, we also chose all negative examples except those in the test set as additional negative examples. Before each training epoch, we randomly selected the same number of negative examples as the positive examples to participate in the training. We train for 200 epochs and retain the parameters that achieve the highest AUC metric on the test set. The results indicated that the performance achieved by randomly selecting negative examples slightly surpasses that of using all negative examples in terms of accuracy and AUC metrics.

Table S1. [Summary of fusion breakpoints of the target fusion gene in each cell line found by each tool] Related to Figure 4

Gene	Fusion position supported by reference	Anchored-fusion	STAR-Fusion	FusionInspector	Arriba	FusionCatcher
BCR-ABL[2]	BCR:chr22:23290413 ABL1:chr9:130854064	BCR:chr22:23290413 ABL1:chr9:130854064	BCR:chr22:23290413 ABL1:chr9:130854064	BCR:chr22:23290413 ABL1:chr9:130854064	BCR:chr22:23290413 ALB1:chr9:130854064 BCR:chr22:23290413 ABL1:chr9:130854067	BCR:chr22:23290413 ABL1:chr9:130854064 BCR:chr22:23290413 ABL1:chr9:130854067 BCR:chr22:23290413 ABL1:chr9:13080369
CIC-DUX4[3]	CIC:chr19:42295124 DUX4:chr4:190174727	CIC:chr19:42295124 DUX4:954	CIC:chr19:42295125 DUX4:chr4:190174728	CIC:chr19:42295124 DUX4:chr4:190174727	CIC:chr19:42295124 DUX4:chr4:190072185	CIC:chr19:42295124 DUX4:chr4:190174727
DUX4-IGH[4]	DUX4:chr4:190174997	DUX4:190174997 IGH:chr14:105896680 DUX4:190175284 IGH:chr14:106638525	DUX4:190175284 IGH:chr14:106638525			DUX4:190174997 IGH:chr14:190287409 DUX4:190174997 IGH:chr14:105779420 IGH:chr14:105864256 DUX4:105782863 IGH:chr14:105863898 DUX4:105782863 IGH:chr14:105864257 DUX4:190173638 IGH:chr14:105835818 DUX4:190173116
TMPRSS2-ERG[5]	TMPRSS2:chr21:41498119 ERG:chr21:38445621 TMPRSS2:chr21:41508081 ERG:chr21:38445621	TMPRSS2:chr21:41498119 ERG:chr21:38445621 TMPRSS2:chr21:41508081 ERG:chr21:38445621	TMPRSS2:chr21:41498119 ERG:chr21:38445621 TMPRSS2:chr21:41507950 ERG:chr21:38445621	TMPRSS2:chr21:41498119 ERG:chr21:38445621 TMPRSS2:chr21:41508081 ERG:chr21:38445621 TMPRSS2:chr21:41507950 ERG:chr21:38445621	TMPRSS2:chr21:41498119 ERG:chr21:38445621 TMPRSS2:chr21:41508081 ERG:chr21:38445621 TMPRSS2:chr21:41507950 ERG:chr21:38445621	TMPRSS2:chr21:41498119 ERG:chr21:38445621 TMPRSS2:chr21:41508081 ERG:chr21:38445621 TMPRSS2:chr21:41498119 ERG:chr21:38493946 TMPRSS2:chr21:41497979 ERG:chr21:38494438
EML4-ALK[6]	EML4:chr2:42295516 ALK:chr2:29223528	EML4:chr2:42295516 ALK:chr2:29223528	EML4:chr2:42295516 ALK:chr2:29223528	EML4:chr2:42295516 ALK:chr2:29223528	EML4:chr2:42295516 ALK:chr2:29223528	EML4:chr2:42295516 ALK:chr2:29223528 EML4:chr2:42299749 ALK:chr2:29223741
TPM3-NTRK1[7]	TPM3:chr1:154170400 NTRK1:chr1:156874571 TPM3:chr1:154169305 NTRK1:chr1:156874571	TPM3:chr1:154170400 NTRK1:chr1:156874571 TPM3:chr1:154169305 NTRK1:chr1:156874571 TPM3:chr1:154165016 NTRK1:ch1:156873751	TPM3:chr1:154170400 NTRK1:chr1:156874571 TPM3:chr1:154169305 NTRK1:chr1:156874571 TPM3:chr1:154165016 NTRK1:ch1:156873751	TPM3:chr1:154170400 NTRK1:chr1:156874571 TPM3:chr1:154169305 NTRK1:chr1:156874571 TPM3:chr1:154165016 NTRK1:ch1:156873751	TPM3:chr1:154170400 NTRK1:chr1:156874571 TPM3:chr1:154169305 NTRK1:chr1:156874571	TPM3:chr1:154170400 NTRK1:chr1:156874571 TPM3:chr1:154169305 NTRK1:chr1:156874571 TPM3:chr1:154165016 NTRK1:ch1:156873751

References for Supplementary Information

S1. Jang, J.S., Li, Y., Mitra, A.K., Bi, L., Abyzov, A., van Wijnen, A.J., Baughn, L.B., Van Ness, B., Rajkumar, V., Kumar, S., et al. (2019). Molecular signatures of multiple myeloma progression through single cell RNA-Seq. Blood Cancer J. *9*, 1–10. 10.1038/s41408-018-0160-x.

S2. Massimino, M., Tirrò, E., Stella, S., Manzella, L., Pennisi, M.S., Romano, C., Vitale, S.R., Puma, A., Tomarchio, C., Di Gregorio, S., et al. (2021). Impact of the Breakpoint Region on the Leukemogenic Potential and the TKI Responsiveness of Atypical BCR-ABL1 Transcripts. Front. Pharmacol. *12*, 669469. 10.3389/fphar.2021.669469.

S3. Yoshimoto, T., Tanaka, M., Homme, M., Yamazaki, Y., Takazawa, Y., Antonescu, C.R., and Nakamura, T. (2017). CIC-DUX4 Induces Small Round Cell Sarcomas Distinct from Ewing Sarcoma. Cancer Res. *77*, 2927–2937. 10.1158/0008-5472.CAN-16-3351.

S4. Tian, L., Shao, Y., Nance, S., Dang, J., Xu, B., Ma, X., Li, Y., Ju, B., Dong, L., Newman, S., et al. (2019). Long-read sequencing unveils IGH-DUX4 translocation into the silenced IGH allele in B-cell acute lymphoblastic leukemia. Nat. Commun. *10*, 2789. 10.1038/s41467-019-10637-8.

S5. Mertz, K.D., Setlur, S.R., Dhanasekaran, S.M., Demichelis, F., Perner, S., Tomlins, S., Tchinda, J., Laxman, B., Vessella, R.L., Beroukhim, R., et al. (2007). Molecular Characterization of TMPRSS2-ERG Gene Fusion in the NCI-H660 Prostate Cancer Cell Line: A New Perspective for an Old Model. Neoplasia N. Y. N *9*, 200–206.

S6. Koivunen, J.P., Mermel, C., Zejnullahu, K., Murphy, C., Lifshits, E., Holmes, A.J., Choi, H.G., Kim, J., Chiang, D., Thomas, R., et al. (2008). EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *14*, 4275–4283. 10.1158/1078-0432.CCR-08-0168.

S7. Ardini, E., Bosotti, R., Borgia, A.L., De Ponti, C., Somaschini, A., Cammarota, R., Amboldi, N., Raddrizzani, L., Milani, A., Magnaghi, P., et al. (2014). The TPM3-NTRK1 rearrangement is a recurring event in colorectal carcinoma and is associated with tumor sensitivity to TRKA kinase inhibition. Mol. Oncol. *8*, 1495–1507. 10.1016/j.molonc.2014.06.001.