Discovery of regulatory noncoding variants in individual cancer genomes by using cis-X

Yu Liu1,2,11✉, Chunliang Li3,11, Shuhong Shen1,4,11, Xiaolong Chen2, Karol Szlachtka2, Michael N. Edmonson2, Ying Shao2, Xiaotu Ma2, Judith Hyle3, Shaela Wright3, Bensheng Ju2, Michael C. Rusch2, Yanling Liu2, Benshang Li1,4, Michael Macias3,2, Liqing Tian2, John Easton2, Xiaotu Ma2, Ying Shao2, Xiaotu Ma2, Michael Macias3,2, Liqing Tian2, John Easton2, Maoxiang Qian5, Jun J. Yang5,6,7, Shaoyan Hu8, A. Thomas Look5,6,7 and Jinghui Zhang1,2,11

We developed cis-X, a computational method for discovering regulatory noncoding variants in cancer by integrating whole-genome and transcriptome sequencing data from a single cancer sample. cis-X first finds aberrantly cis-activated genes that exhibit allele-specific expression accompanied by an elevated outlier expression. It then searches for causal noncoding variants that may introduce aberrant transcription factor binding motifs or enhancer hijacking by structural variations. Analysis of 13 T-lineage acute lymphoblastic leukemias identified a recurrent intronic variant predicted to cis-activate the TAL1 oncogene, a finding validated in vivo by chromatin immunoprecipitation sequencing of a patient-derived xenograft. Candidate oncogenes include the prolactin receptor PRLR activated by a focal deletion that removes a CTCF-insulated neighborhood boundary. cis-X may be applied to pediatric and adult solid tumors that are aneuploid and heterogeneous. In contrast to existing approaches, which require large sample cohorts, cis-X enables the discovery of regulatory noncoding variants in individual cancer genomes.

Even though noncoding regions constitute over 98% of the human genome, the role of noncoding variants in human cancers is poorly understood. Recent epigenetic profiling studies have shown that more than 80% of the human genome is potentially functional and that the noncoding genome is enriched for disease-associated germline variants discovered by genome-wide association studies (GWAS). The oncogenic potential of somatically acquired noncoding variants is also becoming increasingly apparent. Studies have shown that different types of noncoding variants can activate neighboring proto-oncogene transcription in cis, such as genomic rearrangements, local genome duplication and sequence mutations. At present, whole-genome sequencing (WGS) analysis alone has limited power for discovering noncoding regulatory variants; interpretation of the functional impact of noncoding variants is inherently challenging, given the lack of a direct readout analogous to amino acid changes in coding regions. Nonetheless, alteration of gene transcription is an expected outcome of regulatory noncoding variants. Therefore, adaptation of expression quantitative trait locus-like approaches, which test for association of expression level and mutation status, have been used to discover regulatory noncoding variants in several pan-cancer studies. However, these approaches require both recurrence of somatic variants and availability of a large cohort of cancer samples.

Important features shared by regulatory noncoding variants causing aberrant transcriptional activation in cis in cancer include allelic bias with high levels of expression of the affected allele compared to the other allele, and expression levels of the gene qualifying as statistical outliers within a particular tumor type. These features have not been evaluated systematically by robust computational pipelines as the key features of noncoding regulatory variants in cancer. Because many cancer genomes are now analyzed by both WGS and whole-transcriptome sequencing (RNA sequencing (RNA-seq)), it has become feasible to systematically identify genes with both aberrant allele-specific expression and outlier high expression. Thus, we have developed cis-X (cis-expression), a computational tool that analyzes individual cancer genomes by systematically identifying candidate oncogenes with these two features. The computational pipeline then searches for noncoding genomic abnormalities, including copy number variations, structural variations and point mutations (single-nucleotide variants (SNVs) and indels) occurring within the same topologically associating domain (TAD).

As a demonstration of cis-X’s ability to identify oncogenes overexpressed due to noncoding variants, we applied cis-X to 13 pediatric T-lineage acute lymphoblastic leukemias (T-ALLs) from the Shanghai Children’s Medical Center (SCMC), which were analyzed using both WGS and RNA-seq. T-ALL is an ideal model for this demonstration because it is one of the best-studied cancers genomically, with a wealth of existing studies connecting various types of somatic noncoding variants with oncogenic activation of transcription factors. In addition, RNA-seq data generated from more than 260 T-ALLs by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project provide an ideal reference dataset to establish statistically significant outlier high
expression patterns for candidate genes identified in our study. We further show that cis-X can be used to find regulatory noncoding variants in aneuploid solid tumors such as neuroblastoma. Our results conclusively demonstrate that cis-X cannot only identify known and previously unknown regulatory noncoding variants targeting known oncogenes, but also implicate putative oncogenes for subsequent studies to establish selective dependency.

Results

Design of cis-X. cis-X was developed to discover cis-activating somatic variants in individual cancer genomes through integrative variant analysis in WGS and RNA-seq (Fig. 1 and Methods). Its core components identify candidate cis-activated genes that exhibit combined allele-specific expression (ASE) and outlier high expression to implicate regulatory noncoding variants. Allelic imbalance in each tumor was examined for heterozygous markers in DNA, including both germline SNPs and somatic SNVs across the whole genome (Fig. 1).

ASE was determined by measuring statistically significant over-representation of one variant allele in RNA-seq for heterozygous markers in tumor DNA. To accomplish this, we first compared allelic imbalance in empirical RNA-seq data with the theoretical binomial distribution and discovered a constant shift toward imbalance in regions expected to have balanced expression. To correct this, we applied a RNA-seq coverage-dependent Gaussian distribution to measure this deviation and test the ASE for each marker under the balanced transcription model characterized by the convolution of the two distributions (Gaussian and binomial). The convolution model provides a better description of experimentally observed gene transcription than other published ASE detection models, such as MBASED24, across a wide range of RNA-seq coverages (Extended Data Fig. 1 and Supplementary Note). A multi-marker ASE-run analysis was implemented in addition to a single-marker test (Methods). Genes would be identified as ASE candidates if they contained exonic or intronic ASE variants or overlapped with an ASE run. A simulation analysis showed that the power to detect allelic imbalance with convoluted model in cis-X is dependent on an imbalanced transcription ratio, aneuploidy, a number of heterozygous markers and the expression level of the target gene (Extended Data Fig. 2 and Supplementary Note).

For each gene, outlier high expression of a cancer sample of interest was determined by comparing its expression level to those of reference samples with the same tissue type (Extended Data Fig. 3 and Methods). A null distribution of ‘leave-one-out (LOO)’ t-statistic score27 was established using the reference samples. This was then used to determine the false discovery rate (FDR) of the LOO t-statistic score of a cancer sample of interest; those with an FDR < 0.05 were retained as having significant outlier high expression.

For the candidate genes with ASE and outlier high expression, somatically acquired variants within or across their TAD boundaries were considered candidate regulatory noncoding variants. These candidate variants were subsequently annotated for their potential to alter transcription factor binding motifs and other related sequence features (Methods).

Identification of cis-activated genes in pediatric T-ALL. We applied cis-X to analyze candidate cis-activated genes in 13 patients diagnosed with T-ALL at the SCMC (Supplementary Table 1) using a reference expression data matrix generated from 264 T-ALL RNA-seq from the National Cancer Institute (NCI) TARGET project.

On average, T-ALLs had 6,459 expressed protein-coding genes per case, each of which had at least 1 heterozygous variant in DNA (range: 5,321–7,091) that could be used to assess ASE (Supplementary Tables 2 and 3). From these, an average of 416 genes per sample (ranging from 216 to 977) exhibited significantly imbalanced expression from 1 of the 2 alleles; thus, they were ASE genes. By intersecting ASE genes with those showing outlier high expression (Fig. 2), we found a total of 222 candidate cis-activated genes in these 13 patients with T-ALL, ranging from 10 to 31 predicted cis-activated genes per sample (Supplementary Table 4a,b).

To assess the sensitivity of cis-X, we compared the predicted cis-activated genes with 15 T-ALL oncogenes known to be cis-activated by chromosomal translocations, upstream deletions or small noncoding insertions or SNVs20 identified by paired tumor-normal WGS analysis. This benchmark gene set includes TAL1/2 (n = 7), LMO1/2/3 (n = 7) and TLX3 (n = 1). cis-X identified 14 of the 15 (93.3%) of the benchmark genes (Fig. 2). Only one cis-regulated LMO2 gene was missed because LMO2 expression in the sample did not reach statistical significance as an overexpressed outlier.

Identification of somatic noncoding regulatory variants. We next searched for somatic genomic aberrations that could cause cis-activation of the 222 candidate cis-activated genes identified by cis-X (Methods). Of the 222 candidate genes, 33 had somatic alterations located within their TADs that cis-X predicted as candidate causal variants for cis-activation of the candidates (Fig. 2). These included 18 structural alterations affecting 14 unique genes and 12 point mutations (SNVs/indels) targeting 11 unique genes (Supplementary Table 5a,b). As expected, we detected the known noncoding variants for well-recognized T-ALL oncogenes, including those activating TAL1/2, LMO1/2 and TLX3. When point mutations generally have a single regulatory target, structural variants can potentially affect the expression levels of multiple targets in the neighborhood. For example, cis-X identified both ABTB2 and TMEM38B as cis-activated by structural variants; however, both genes were adjacent to known proto-oncogenes in the same TAD (LMO2 and TAL2, respectively); thus, they were potentially coregulated by an aberrant enhancer. In addition to frequently disregulated transcription factors, LMO3 exhibited aberrantly high levels of mono-allelic expression due to a translocation between LMO3 and the TCRB locus in a single T-ALL (Extended Data Figs. 4a,b). LMO3 activation is rare in T-ALL, having been previously identified in only a single T-ALL by chromatin conformation capture on chip (4C) technology27. Similarly, this was the only tumor with LMO3

Fig. 1 | cis-X workflow. cis-X was designed to perform integrated analysis of WGS and RNA-seq data generated from an individual tumor genome. It integrates ASE and outlier high expression as key signatures of cis-activated genes to seed the discovery of regulatory noncoding variants in the context of 3D architecture of the genome. Functional genomics data such as ChIP-seq generated from samples with matching tissue of origin and variant context can be provided by the user to enhance candidate variant annotation.
cis-activation of PRLR in T-ALL by upstream deletion. PRLR, which encodes the proline receptor, was one of the cis-activated candidate genes identified by cis-X and showed both ASE and outlier high expression in the T-ALL sample SJALL043558_D1 (Fig. 4a,b). A matching 546.7-kilobase (kb) focal deletion located 63.5 kb distal to the transcription start site of PRLR was detected in this tumor. This large focal deletion spanned the entire locus of PRLR, the adjacent insulated neighborhood to the PRLR hood between LMO2 and IL7R beyond the breakpoint of the deletion furthest upstream of PRLR (Fig. 4c). An active enhancer could be observed in tumors with positive IL7R expression (Fig. 4b), also harbored somatic focal deletions in the same region (Fig. 4c). The deletions included the nodal CTCF binding sites forming the boundary of the insulated neighborhood containing the PRLR gene, as defined by chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) interactions in Jurkat cells, a T-ALL cell line with no somatic alteration in this region (Fig. 4c). An active enhancer could be observed beyond the breakpoint of the deletion furthest upstream of PRLR in tumors with positive IL7R transcription (Extended Data Fig. 6c). The deletion disrupted the boundaries of the insulated neighborhood between PRLR and IL7R, bringing the active enhancer from the adjacent insulated neighborhood to the PRLR promoter, thus cis-activating PRLR.

cis-activating noncoding variants in neuroblastoma (NBL). We applied cis-X in NBL to test the broad applicability of cis-X on transcriptional activity of the mutant allele was enhanced 1.6-fold ($P = 0.003$, two-sided $t$-test) compared with the reference allele (Fig. 3c). The same mutation was also evident in three additional T-ALLs in an independent cohort (data not shown). The recurrent presence of this somatic mutation in concert with aberrantly high $TAL1$ expression levels from one allele, along with enhancer activity demonstrated by the luciferase assays, provide multiple lines of evidence supporting this as a new pathogenic noncoding sequence mutation in T-ALL.

We further investigated this noncoding mutation in a patient-derived xenograft (PDX) mouse model generated from one of the patient samples (SJALL018373) containing this mutation. We performed H3K27Ac and YY1 chromatin immunoprecipitation sequencing (ChIP–seq) analysis on genomic DNA from this PDX tumor to assay the enhancer activity and transcription factor binding caused by this mutation in the native three-dimensional (3D) genomic architecture of these leukemia cells. An active enhancer defined by H3K27Ac was observed around the noncoding mutation (Fig. 3d) but was absent in both normal T cells and those T-ALL cell lines neither carrying this mutation nor expressing $TAL1$ (Extended Data Fig. 5b). Notably, significant allelic imbalance was observed within this active enhancer since all H3K27Ac ChIP–seq reads contained the mutant allele (Fig. 3d; the mutant to wild-type allele ratio is 11.0, $P = 9.8 \times 10^{-4}$, two-sided binomial test), indicating that the enhancer was only present on the mutant allele but not on the wild-type allele. Importantly, YY1 ChIP–seq analysis of the PDX showed a sharp peak at the mutant site with the mutant allele present in nearly all aligned reads (Fig. 3d; $P = 1.5 \times 10^{-4}$, two-sided binomial test), validating YY1 binding at this mutant enhancer site.

---

**Fig. 2** Candidate cis-regulated genes identified by cis-X in 13 pediatric patients with T-ALL. Each dot represents a cis-activated candidate gene identified in a single tumor that passes the ASE test (y axis) and outlier high expression test (x axis). The significance of genes showing outlier high expression and higher transcription from one allele compared to the other are shown on the x and y axes separately as the $-\log_{10}(q)$ after multiple testing correction (Methods). The cis-activated candidates nominated from ASE runs alone are shown at the bottom of the plot. The horizontal and vertical lines in blue represent a $q$ value of 0.05. The candidates are color-coded into different groups based on the cis-activating genomic aberrations and candidate gene status.

A recurrent intron mutation activates $TAL1$ expression. cis-X identified mono-allelic high expression of $TAL1$ in two T-ALLs; both harbored a recurrent heterozygous C to T mutation located in intron 1 of $TAL1$ (chromosome 1, g.47696311C>T, hg19; Extended Data Fig. 5a). $TAL1$ expression levels in these two T-ALLs were comparable to those carrying known regulatory insertions or structural variations (Fig. 3a,b). The mutation was predicted to introduce a de novo binding motif for YY1, a transcription factor recently found to contribute to enhancer–promoter interactions similar to those mediated by CTCF. Luciferase assays revealed that
solid tumors with complex genomes. First, we analyzed matched WGS and RNA-seq data generated from four NBL cell lines, with a focus on finding regulatory copy number aberrations and structural variants in aneuploid genomes (Methods and Extended Data Fig. 7a). The gene-specific reference expression matrix was built with a cohort of 123 NBL samples from the NCI TARGET project18,32. Using the same parameters applied in T-ALL analysis, cis-X identified a total of 342 cis-activated candidates in the four NBL cell lines.

| SJALL018373_D1 | Reference allele | Nonreference allele |
|----------------|-----------------|--------------------|
| WGS RNA-seq    | C>T mutation    | SJALL043861_D1     |
| SJALL018373_D1 | Reference allele | Nonreference allele |
| WGS RNA-seq    | C>T mutation    |

---

Fig. 3 | Discovery and validation of a recurrent intronic noncoding mutation activating the TAL1 oncogene. a, ASE of TAL1 in the two T-ALLs carrying a recurrent somatic mutation in intron 1 of TAL1. ASE status was determined by the mono-allelic expression of the heterozygous germline/somatic variants labeled above the wiggle plot, with blue and red representing the reference and alternative alleles, respectively. The arrow indicates the position of the new noncoding C>T mutation. b, Expression of TAL1 in the two T-ALLs (shown in red) carrying the new mutation is comparable to those carrying known cis-activating regulatory variants of TAL1, including the insertion upstream of TAL1, STIL-TAL1 deletion and other rearrangements in the TAL1 locus, as shown in different colors. c, Luciferase assay showing increased transcription activity in cells that transformed with the mutant T allele compared to the wild-type C allele. Data are presented as the mean value ± s.d. (n = 3 independent experiments). A two-sided t-test was performed. d, The mutation introduced transcription factor YY1 binding and an active enhancer at the TAL1 locus in vivo, defined by YY1 and H3K27Ac ChIP-seq profiling generated from the patient-derived xenograft of the patient carrying this mutation. YY1 binding and the enhancer are only detected on the mutant allele since all reads under the peaks from the ChIP-seq experiment carry the mono-allelic mutant T allele (red), as shown at the bottom. By contrast, the tumor DNA harbors a heterozygous C to T mutation (Extended Data Fig. 5).
cell lines, ranging from 25 to 117 per sample. Detected candidate noncoding variants consisted of 22 structural variants affecting 17 unique genes (Supplementary Table 6 and Extended Data Fig. 7b). Notably, TERT was identified to be cis-activated by interchromosomal translocations in three cell lines, replicating previous findings of TERT activation in NBL (refs. 18,19) and other tumors (ref. 20).

To assess the performance of cis-X on heterogeneous tumor samples, we analyzed 90 NBL primary tumor samples with matching RNA-seq and WGS from the TARGET cohort (Supplementary Table 7), 42 of which had a positive immune cell infiltration signature based on previous analysis (ref. 21). With a focus on TERT, we found that in the ten samples identified to have TERT cis-activation by structural rearrangements, six had immune cell infiltration signatures and six had amplifications at the TERT region (Extended Data Fig. 8). These results demonstrate that cis-X can identify regulatory variants in cancer genomes of aneuploidy and tumor heterogeneity.

cis-X analysis of adult melanoma. To evaluate the performance of cis-X on adult cancer, we ran the pipeline on 38 The Cancer Genome Atlas (TCGA) melanomas that have both paired tumor-normal WGS and tumor RNA-seq from the TARGET cohort (Supplementary Table 9 and Methods). Among the 14 candidate noncoding structural variants was a deletion predicted to cis-activate CDKN2A in sample TCGA-DA-A1HY-06, which appeared paradoxical given that CDKN2A is a known tumor suppressor gene. The CDKN2A locus encodes two proteins translated in different reading frames: p16INK4A, a negative regulator of CDK4/6 of the retinoblastoma pathway; and ARF, a negative regulator of MDM2 of the TP53 pathway. The deletion in TCGA-DA-A1HY-06 was distinct from the other CDKN2A deletions in that it removed only exon 1(p) and the flanking intronic region unique to ARF but retained the full reading frame of p16INK4A (Fig. 5a). Notably, the RNA-seq data showed close-to-null expression of ARF but very high expression of p16INK4A (Fig. 5b,c). This is consistent with previous studies of mouse models, which found upregulation of p16INK4A in pre-B lymphocytes and keratinocytes lacking exon 1(p) (refs. 22–24). TCGA-DA-A1HY-06 also had the highest expression of the CDKN2A loci of the entire cohort, suggesting that the elevated expression of p16INK4A was not caused by loss of competition for alternative usage of exon 1(p). It is possible that p16INK4A cis-activation in TCGA-DA-A1HY-06 was caused by the loss of a repressive element around exon 1(p) identified recently by CRISPR screening (ref. 25); further functional experiments are needed to investigate this possibility. Importantly, mono-allelic expression of the somatic mutation p.Pro811Leu in exon 2 (Fig. 5c), which is known to disable p16INK4A binding to CDK4/6 and cause cytoplasmic mislocalization of p16INK4A (refs. 25,26), indicates that cis-activation of

**Fig. 4 | A putative oncogene PRLR in T-ALL is identified by cis-X.** a, ASE of PRLR in T-ALL SJALL043558_D1. All 12 heterozygous germline/somatic variants in this locus exhibit mono-allelic expression as indicated by their reference (blue) and nonreference allele (red) fraction in WGS and RNA-seq displayed above the RNA-seq wiggle plot. b, Outlier high expression of PRLR found in sample SJALL043558_D1 (red dot) together with two additional T-ALLs (orange dots) from the NCI TARGET T-ALL cohort (n = 264 samples). PRLR expression level (FPKM) was plotted on the y axis. c, Focal deletions identified in all 3 T-ALLs exhibiting the outlier PRLR expression shown in b. The cohesin ChIA-PET interactions and CTCF binding profile (blue) collected from Jurkat cell lines show the insulated neighborhood structure of the wild-type PRLR locus since no somatic alterations were detected in this region in Jurkat cells. The candidate boundaries of the insulated neighborhood affected by the deletions detected in three patient with T-ALL are indicated by the boxes.
the mutant p16INK4A allele probably contributed to the tumorigenesis of this melanoma sample.

Discussion

We have developed the computational method cis-X, which systematically identifies genes in individual cancer genomes that exhibit both unusually (outlier) high levels of expression and allelic bias such that one allele is expressed at significantly higher levels than the other. cis-X enables the discovery of somatically acquired noncoding regulatory DNA alterations that operate in cis to activate the expression of neighboring proto-oncogenes on the same allele in individual human cancers. This design overcomes two major limitations of existing approaches (that is, hotspot analysis or mutation-expression association): the requirement for a large cohort of samples, which is usually only attainable in a pan-cancer study and the requirement that noncoding variants are recurrent to achieve statistical significance. By contrast, cis-X can be applied to identify activated oncogenes in individual cancer samples. This was demonstrated in our discovery of LMO3 and PRLR in a single case in the 13 T-ALLs from the SCMC; the finding of PRLR cis-activation was subsequently replicated in two TARGET T-ALLs by using the copy number variation data generated from the SNP arrays. cis-X analysis can also unveil heterogeneity in the disruption of normal gene regulation by different variant types (that is, point mutations or structural alterations) dispersed in multiple regions. For example, our finding of the TAL1 intronic mutation is located 8.7 kb distal to the previously identified somatic insertion hotspot upstream of transcription initiation and the two regulatory mutations we found that activate LMO2 are 53 kb apart. This capability becomes especially important in precision medicine where genome analysis is applied in a case-specific way.

By effectively using the genetic variation data in WGS and interrogating associated expression status in RNA-seq, cis-X can identify the cis-regulated candidate genes within a single cancer genome and has the flexibility to analyze RNA-seq data mapped by different algorithms (Supplementary Note and Supplementary Fig. 1). One potential limitation is posed by a lack of heterozygous variants in a gene; this can be ameliorated to some extent by the ASE-run analysis implemented in cis-X, which enables use of informative markers flanking the genes. Meanwhile, broad adoption of total RNA-seq could further increase the power of cis-X since transcribed intronic variants from unspliced RNA can provide additional informative information.
markers for ASE analysis. In addition to the number of heterozygous variants, imbalanced transcription ratio (effect size), aneuploidy and the expression level of the target gene can all affect the power to detect ASE (Supplementary Note). Currently cis-X is designed only to identify cis-activating candidates that exhibit outlier high expression and will miss regulatory variants causing downregulation since reduced expression of one allele can be caused by many alternative mechanisms, such as imprinting or nonsense mediated decay, due to the presence of truncation variants. Additionally, it may not be effective in detecting regulatory noncoding variants on sex chromosomes due to scarcity of biallelic expression in normal cells, or in detecting variants that do not cause outlier high expression (Supplementary Note). Despite these limitations, cis-X provides an approach that complements the existing recurrence-based methods in detecting cis-activating somatic variants and can be applied to analyze solid tumors that may have extensive aneuploidy and heterogeneity (Fig. 5 and Extended Data Fig. 8).

Functional genomics data including 3D genome architecture and epigenetic profiling of regulatory elements could provide valuable information to noncoding variant prediction. We have incorporated a TAD structure to define the regulatory territory of a given cis-activated candidate in noncoding variant discovery since this is relatively stable across different tissue types41. In the current implementation, the TAD structure of the human embryonic stem cells (H1) is used as the default because a good fraction of proto-oncogenes are actively transcribed during the embryonic stage42. This can be replaced by custom TAD data generated from a matching cellular context provided by a user. However, epigenetic data such as the H3K27Ac ChIP–seq used for enhancer profiling can be highly specific to a tissue type, a cell lineage or presence of chromatin structure-based prediction of recurrent noncoding variants. Ideally, such experiments should be carried out in a PDX model as demonstrated in our validation of the new cis-X, which complements the recurrence-based methods employed by the Pan-Cancer Analysis of Whole Genomes Consortium in a recent study43, will lead to new insights on the noncoding driver variants that cause oncogenic activation. The publicly available cis-X software provides a powerful approach to investigate the functional role of noncoding variants. We expect it will spur extensive laboratory investigations and functional studies, both to assess the level of tumor cell dependency on the activated oncogene and to establish experimentally the mechanisms underlying aberrant transcriptional activation.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0659-5.

Received: 16 July 2019; Accepted: 5 June 2020; Published online: 6 July 2020

References
1. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
2. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195 (2012).
3. Khurana, E. et al. Role of non-coding sequence variants in cancer. Nat. Rev. Genet. 17, 93–108 (2016).
4. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 351, 1454–1458 (2016).
5. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. 49, 65–74 (2017).
6. Northcott, P. A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511, 428–434 (2014).
7. Zhang, J. et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat. Genet. 48, 1481–1489 (2016).
8. Zhang, X. et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nat. Genet. 48, 176–182 (2016).
9. Mansour, M. R. et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science 346, 1373–1377 (2014).
10. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. Science 339, 957–959 (2013).
11. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. Science 339, 959–961 (2013).
12. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 578, 102–111 (2020).
13. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).
14. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat. Genet. 46, 1258–1263 (2014).
15. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. 46, 1160–1165 (2014).
16. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat. Genet. 47, 710–716 (2015).
17. Kim, K. et al. Chromatin structure-based prediction of recurrent noncoding mutations in cancer. Nat. Genet. 48, 1321–1326 (2016).
18. Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature 555, 371–376 (2018).
19. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 487, 376–380 (2012).
20. Balzer, L. & Ferrando, A. The genetics and mechanisms of T cell acute lymphoblastic leukaemia. Nat. Rev. Cancer 16, 494–507 (2016).
21. Liu, Y. et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat. Genet. 49, 1211–1218 (2017).
22. Li, Z. et al. APOBEC signature mutation generates an oncogenic enhancer that drives LMO1 expression in T-ALL. Leukemia 31, 2057–2064 (2017).
23. Hsu, S. et al. Whole-genome noncoding sequence analysis in T-cell acute lymphoblastic leukaemia identifies oncogene enhancer mutations. Blood 129, 3264–3268 (2017).
24. Abraham, B. J. et al. Small genomic insertions form enhancers that misregulate oncogenes. *Nat. Commun.* **8**, 14385 (2017).
25. Rahman, S. et al. Activation of the *LMO2* oncogene through a somatically acquired neomorphic promoter in T-cell acute lymphoblastic leukemia. *Blood* **129**, 3221–3226 (2017).
26. Mayba, O. et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* **15**, 405 (2014).
27. Pawlikowska, I. et al. The most informative spacing test effectively discovers biologically relevant outliers or multiple modes in expression. *Bioinformatics* **30**, 1400–1408 (2014).
28. Simonis, M. et al. High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nat. Methods* **6**, 837–842 (2009).
29. Weintraub, A. S. et al. YY1 Is a structural regulator of enhancer-promoter loops. *Cell* **171**, 1573–1588.e28 (2017).
30. Ali, S. & Ali, S. Prolactin receptor regulates Stat5 tyrosine phosphorylation and nuclear translocation by two separate pathways. *J. Biol. Chem.* **273**, 7709–7716 (1998).
31. Goffin, V. Prolactin receptor targeting in breast and prostate cancers: New insights into an old challenge. *Pharmacol. Ther.* **179**, 111–126 (2017).
32. Pugh, T. J. et al. The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013).
33. Peifer, M. et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).
34. Valentijn, L. J. et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat. Genet.* **47**, 1411–1414 (2015).
35. Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
36. Zhang, Y. et al. High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nat. Commun.* **11**, 736 (2020).
37. Akbani, R. et al. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
38. Strub, T. et al. SIRT6 haploinsufficiency induces BRAFV600E melanoma cell resistance to MAPK inhibitors via IGF signalling. *Nat. Commun.* **9**, 3440 (2018).
39. Zhou, B. et al. INO80 governs superenhancer-mediated oncogenic transcription and tumor growth in melanoma. *Genes Dev.* **30**, 1440–1453 (2016).
40. Fontanals-Girera, B. et al. Harnessing BET inhibitor sensitivity reveals AMIGO2 as a melanoma survival gene. *Mol. Cell* **68**, 731–744.e9 (2018).
41. Kaufman, C. K. et al. A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* **351**, aad2197 (2016).
42. Lin, A. W. & Lowe, S. W. Oncogenic ras activates the ARF-p53 pathway to suppress epithelial cell transformation. *Proc. Natl Acad. Sci. USA* **98**, 5025–5030 (2001).
43. Kamiijo, T. et al. Tumor suppression at the mouse INK4a locus mediated by the alternative reading frame product p19 ARF. *Cell* **91**, 649–659 (1997).
44. Zhang, Y. et al. A cis-element within the ARF locus mediates repression of p16INK4A expression via long-range chromatin interactions. *Proc. Natl Acad. Sci. USA* **116**, 26644–26652 (2019).
45. Zhang, B. & Peng, Z. Defective folding of mutant p16INK4A proteins encoded by tumor-derived alleles. *J. Biol. Chem.* **271**, 28734–28737 (1996).
46. Walker, G. J., Gabrielli, B. G., Castellano, M. & Hayward, N. K. Functional reassessment of P16 variants using a transfection-based assay. *Int. J. Cancer* **82**, 305–312 (1999).
47. Yu, M. & Ren, B. The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.* **33**, 265–289 (2017).
48. Monk, M. & Holding, C. Human embryonic genes re-expressed in cancer cells. *Oncogene* **33**, 265–289 (2017).
49. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
50. Hidalgo, M. et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.* **4**, 998–1013 (2014).
Methods

cis-X analysis. The cis-X computational pipeline is designed to analyze somatically acquired regulatory noncoding variants from individual tumor genomes. The workflow for cis-X includes the following steps.

Modeling balanced transcription. To query heterozygous genetic variations in DNA that exhibits allelic imbalance in RNA expression we first define $i$, a value of imbalance, to quantitatively measure ASE:

$$i = \frac{|R_i - R_i|}{R_i + R_i}$$

where $R_i$ and $R_i$ denote the RNA-seq read count of alleles a and b of a heterozygous variant in tumor DNA. This value can range from 0, that is, balanced expression of two alleles ($R_i = R_i$) to 1, that is, mono-allelic expression ($R_i = 0$ or $R_i = 0$).

First, we compared the distribution of $i$ collected from the empirical data with the theoretical binomial distribution assuming balanced expression, which implies the probability of success is equal to the probability of observing one of the alleles ($P = 0.5$) and a number of tries is equal to the coverage in RNA-seq. Our preliminary analysis, which compared the empirical WGS and RNA-seq data from 10 T-ALL cell lines (including Jurkat, Loucy, MOLT-3, CCRF-CEM, DND-41, DUS28, KOPT-K1, P12-ICHIKAWA, PF-382 and RPMI-8402; unpublished data) with the theoretical binomial distribution, revealed a constant shift of empirically measured $i$ toward less balanced transcription (Extended Data Fig. 1). We reasoned that the shift in empirical data is caused by a combination of sampling bias of the two alleles and allelic transcription fluctuations. As a result, the balanced transcription can be modeled by convolution of corresponding distributions:

$$P(k, N) = B(k, N, P = 0.5) \times G(\mu, \sigma = \sigma(N))$$

where $B$ denotes the binomial distribution, $G$ denotes Gaussian distribution, $N$ denotes coverage in RNA-seq for a given marker, $\sigma$ denotes the number of reads for one allele observed in RNA-seq and $P$ for the possibility of transcription of one allele over the other, which is equal to the variant allele frequency in DNA for balanced transcription (for example, 0.5 for diploid regions, 0.3 or 0.67 for regions with three copy numbers). The asterisk denotes the operation of convolution.

To estimate the sigma in the Gaussian distribution for balanced transcription, we analyzed the WGS and RNA-seq data generated from 10 T-ALL cell lines. To select the balanced transcribed variants, we excluded genes with heterozygous variants within a single gene ranged from 1 to 20 and an imbalance transcription in-between were tested, including 2:1, 3:1, 4:1, 5:1, 10:1, and 20:1. The asterisk denotes the operation of convolution.

ASE candidate analysis. Heterozygous variants extracted from an individual tumor genome were used for allelic imbalance analysis in RNA-seq data. A coverage threshold of ten in both WGS and RNA-seq was applied in the current study. cis-X first identifies markers with an ASE signal by comparing the variant allele frequency (VAF) between genomic DNA and RNA transcripts with the convolution model. $P$ values for individual markers showing imbalanced expression were further combined based on gene structure and corrected for multiple testing. Genes with an FDR < 0.05 and an average VAF$_{\text{RNA}}$ - VAF$_{\text{DNA}}$ ≥ 0.3 (a threshold of 0.2 would be used if over 30% of markers in the gene fall into copy number aberration (CNA)/loss of heterozygosity (LOH) regions) were considered as ASE candidates. A multi-marker analysis was implemented in parallel to identify regions with consecutive markers that exhibited ASE as ASE runs. An ASE run requires a minimum of four markers showing significant ASE or mono-allelic expression and terminates if two or more markers are not qualified in these criteria sequentially. Genes overlapping with the ASE run were considered as allele-specific transcribed.

Outlier high expression candidate analysis. The expression of a given gene in the tumor sample under analysis was compared to a precalculated gene-specific reference expression matrix with a LOO test. Tumors of the same histotype were collected from the empirical data and were compared to a precalculated gene-specific reference expression matrix with a LOO test. We required a minimum of ten cases to build the reference expression matrix for each gene. For genes that could not meet the criteria, the expression value from the cohort as a whole was used. A null distribution of the t-statistic from the LOO test was established and used to estimate the FDR in outlier high expression analysis. A gene with an FDR < 0.05 was considered outlier high expression. Genes with both aberrant ASE and significant outlier high expression signature with an FPKM greater than user-specified threshold (FPKM > 5 was used in the T-ALL analysis presented in this study) were considered as cis-activated candidates. Potential oncogenes presented in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database were nominated if they showed BAF$_{\text{RNA}}$ - BAF$_{\text{DNA}}$ ≥ 0.4 for over 90% of markers inside the gene and significant outlier high expression with FPKM ≥ 1.

Transcription factor binding site analysis for somatic SNV/indels. Mutations were evaluated for potential regulatory function with a combined approach that takes into account a P value of motif prediction, expression level of the transcription factor in the tumor sample and absence of prediction in the reference genome. First, sequences were extracted from human regulatory regions (https://genome.ucsc.edu/goldenPath/help/twobit.html) and transcription factor motif analysis was carried out with the FIMO (v.4.9.0) package, with a P value threshold of 10^-3. A total of 614 human transcription factor binding motifs from the HOCOMOCO53 database were included in this analysis. Only mutations that could introduce a transcription factor binding motif that was absent from the reference sequence were kept for downstream analysis. We further required that a set of predicted transcription factors be highly expressed (FPKM > 10) in the individual tumor. The combined approach enables the detection of well-documented noncoding regulatory variants that can be missed by using a stringent P value cutoff alone. For example, in the T-ALL analysis presented in this study, the known validated TAL1 super-enhancer mutation that introduces an MYB binding motif would have been missed if a standard FDR were to be applied since the FDR of motif prediction was 0.236. Experimental validation of the predicted transcription factor binding motif using an in vivo model may be required for further investigation.

Predicting potential regulatory variants for cis-activated candidate genes. Somatic variants were assigned to dysregulated genes in the same TADS defined by human embryonic stem cell (H1) Hi-C data as the default. This is because TAD boundaries are relatively stable across different tissue types and independent of gene transcription status; alternatively, the cellular lineage of H1 is a good model for representing reactivation of proto-oncogenes normally expressed only during embryonic development. Alternatively, a TAD structure generated from a matching cellular context could be provided by the user. For sequence mutations, we limited the distance between mutation and promoter (>1-2kb) to 200 base pairs (bp) of the RefSeq transcription start site) of the potential cis-activated candidate to 200 kb, mimicking the median length of contact domains in the human genome. The candidate sequence mutations were further annotated with the DNA sequence of accessible regulatory regions defined by the Roadmap Epigenomics Project. In addition, user-provided functional genomics data, such as ChIP-seq or ATAC-seq generated from samples with matching tissue and variant context, can be incorporated for annotation. In situations where multiple variants were present in the same regulatory neighborhood of cis-activated gene, gross alterations (for example, structural variants or CNAs) were rated as having higher impact than sequence mutations due to their potential for causing enhancer hijacking.

cis-X simulation. A simulation was carried out to emulate the performance of cis-X in identifying the tumor regions with different ploidy levels. Variables that could affect ASE detection were included in this simulation: (1) copy number alterations, including simulation of both diploid and CNA regions with 1 or 2 copy gains and (2) coverage for each marker in RNA-seq. A variety of coverages were tested, ranging from 10 to 500, to represent different levels of gene expression. However, DNA coverage > 30 was used as a constant in this analysis for diploid regions. DNA coverage increases along with copy number gains, with 45 when copy number = 3 and 60 for copy number = 4. The number of heterozygous markers within a single gene ranged from 1 to 20 and an imbalance transcription ratio between alleles existed. An imbalance ratio of 1:1 was used to represent the balanced transcription needed to estimate the false positive rate, while a complete ASE was represented with a ratio of 100:1. Meanwhile, a list of different allele imbalance ratios in-between were tested, including 2:1, 3:1, 4:1, 5:1, 10:1 and 100:1. 2,000 simulations were carried out for each combination. Random sampling was performed for the alleles with different probabilities controlled by the allelic imbalance ratio. The imbalance between DNA and RNA was tested with the convoluted balanced transcription model as described earlier.
Patients and samples. Tumor and paired remission samples were collected from a total of 13 patients diagnosed with T-ALL from the SCMC. The study was approved by the SCMC institutional review board. Written informed consent was obtained from parents for all patients.

WG5 and analysis. DNA was extracted with the QIAamp DNA Blood Mini Kit (QIAGEN) according to the manufacturer’s instructions and quantified by agarose gel electrophoresis and Qubit dsDNA BR Assay Kit (catalog no. Q32850; Thermo Fisher Scientific). Then, 1 μg of genomic DNA (OD260/280 ratio ranging from 1.8 to 2.0) was diluted with 1× low Tris-EDTA buffer to a total volume of 50 μl and fragmented by sonication on the Covaris S2 (Covaris). Fragmented DNA was repaired, ligated with Illumina adapter and size-selected, aiming for a 250–300 bp product. The size-selected DNA library was then PCR-amplified for 15 cycles and validated using the Agilent 2100 Bioanalyzer (Agilent Technologies). The library was sequenced on an Illumina HiSeq X Ten sequencer according to Illumina–provided protocols for 2 × 150 paired-end sequencing.

WG5 data were mapped to reference human genome assembly GRCh37-lite with Burrows–Wheeler Aligner5. Somatic variants in each tumor were analyzed by Bambino (v.1.6.6) (SNV/indels), CONSOERTING5 (in both paired and tumor-only mode for somatic and germline CNA) and CRIST (v.1.0.8) (structural variants). The structural variants discovered in the NBL cell lines were further filtered against the Database of Genomic Variants5 with AnnovT5 (v.2.2.4), as well as a local germline rearrangements database constructed from analyzing WG5 data in the Pediatric Cancer Genome Project to remove common structural variants in the human genome. The detected somatic aberrations were manually curated to further rule out false discoveries.

Transcriptome sequencing (RNA-seq) and analysis. RNA purification, reverse transcription, library construction and sequencing were performed at WuXi NextCODE (Shanghai) according to the manufacturer’s instructions (Illumina). Ribosomal RNA (rRNA)-depleted sequencing libraries from total RNA were prepared using the Illumina TruSeq Stranded Total RNA Gold Library Preparation Kit according to the vendor’s recommendations. Approximately 1 μg of total RNA was used as the input material; the Ribo-Zero Gold Kit was used to remove both cytoplasmic and mitochondrial RNA. Sequencing was performed using a HiSeq 2500 system according to the manufacturer’s protocol. RNA-seq data were mapped with STRandless Alignment Tool (STAR)6, version 2.6.0. R scripts (in preparation), as described previously8. The gene-level read count was generated with HTSeq-count9; the number of FPKM was calculated on the basis of the transcript models in GENCODE v.19. Cluster analysis was carried out with the Ward’s minimum variance method with the same list of genes as used in a previous study10.

PDX. Mononuclear cells isolated from the patient’s bone marrow samples were engrafted into 5 irradiated NOG mice (3–4 week-old female mice; Beijing Vital River Laboratory Animal Technology) via intravenous injection. Weekly monitoring of peripheral appearance of leukemic cells (CD45+) was performed via flow cytometry with retro-orbital bleeding. Mice were bled in an animal facility with a 12 h/12 h light–dark cycle, 20–26 °C temperature and 40–70% humidity. Experiments were conducted under sterile conditions at the Crown Bioscience SPF facility and performed strictly according to the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocol was approved by the Institutional Ethical Committee of the Bambino (v.1.4.0) (SNV/indels, structural variants and copy number variants, and the germline SNPs characterized in our previous pan-cancer analysis13 were analyzed for candidate noncoding driver variants.

Analysis of adult melanoma by cis-X. A total of 38 adult melanoma cases with paired tumor-normal WGS and tumor RNA-seq data generated in the TCGA project14 were analyzed. Raw data were downloaded from the Genomic Data Commons (https://portal.gdc.cancer.gov/legacy-archive/tdac/home). For RNA-seq, FASTQ files were mapped to GRCh37-lite as described earlier. For the WGS data, the downloaded aligned BAM files were first transformed to FASTQ with BEDTools67 v.2.25.0 and processed as described earlier. The epigenetic dataset included 5 ATAC-seq data from 5 melanoma cell lines and 12 H3K27Ac ChiP–seq data from 5 melanoma cell lines and 2 normal human melanocyte cell lines. The ChiP–seq data were processed as described earlier. The ATAC-seq data were mapped to GRCh37-lite with the Burrows–Wheeler Aligner5 v.0.7.15. Ambiguously mapped and duplicated reads were removed subsequently with Picard v.2.6.0 (http://broadinstitute.github.io/picard). Aligned reads were transformed to BEDPE format with BEDTools67 and only nucleosome-free fragments (fragment length < 100) were kept for peak calling using MACS2 with default parameters. The resulting peaks from the ChiP–seq and ATAC-seq of melanoma cell lines were combined to annotate the cis-activated candidates detected by cis-X. Candidate variants were manually curated to remove false positives. Variants with hard-to-define target genes were excluded, such as structural variants involved in chromosome with many-to-many mapping of structural variants and target genes. SNVs with an ultraviolet-like signature (that is, COSMIC signature 7) were identified and labeled accordingly.

Cell culture. Jurkat cell culture were cultured in a suspension system at 37°C with 5% CO2 in complete Roswell Park Memorial Institute (RPMI) medium containing 10% FCS, 2 mM of glutamine, 100 μM of penicillin and 100μg/ml streptomycin.

 Luciferase assay. DNA fragments were synthesized by Integrated DNA Technologies and loaded into the pGL3-Promoter Vector (catalog no. E1761; Promega Corporation). A total of 1×10^5 Jurkat cells were resuspended in 100 μl of nucleofector solution mixture (Nucleofector Cell Line Nucleofector Kit V; Lonza); then, 1.5 μg of each reporter vector and 0.5 μg of pR-L-TK Renilla control luciferase reporter vector (Promega Corporation) were added. Cells with reporter plasmid DNA were electroporated into Jurkat cells using program X-065 on a Lonza Nucleofector 2b and resuspended in 2 ml of RPMI medium supplemented with 10% FCS and penicillin-streptomycin. After being incubated at 37 °C/5% CO2, 48h, cells were collected by centrifugation and luciferase activities were measured using the Dual-Luciferase Reporter Assay System (catalog no. E1960; Promega Corporation). For each putative enhancer, experiments were performed in triplicate and repeated independently three or four times (Supplementary Note). Cell numbers and transfection efficiency were normalized to Renilla luciferase activity.

Statistics and reproducibility. The statistical difference between mutant and reference sequences in the luciferase assay was performed by two-sided t-test. Each experiment was replicated at least twice independently. Overrepresentation of mutat rather than reference alleles in H3K27Ac and YY1 ChIP–seq data was performed using a two-sided binomial test. Analyses were performed with Prism 8 (GraphPad Software) and R v.3.10.0 (R Foundation for Statistical Computing).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. WGS and RNA-seq data for the SCMC cohort analyzed in this study can be accessed from the Genome Sequence Archive for Human under the National Genomics Data Center of China (http://bigd.big.ac.cn/gsa-human), under
accession nos. HRA000097 and HRA000096 for WGS and RNA-seq, respectively. The data are publicly available to users following a standard access application process for human genomic and associated phenotypic data. The ChIP-seq data generated in this study can be accessed from the Gene Expression Omnibus under accession nos. GSE113565 and GSE145549, for H3K27Ac and YY1, respectively, with the called peaks (in BED format) available upon request. Whole-exome sequencing and RNA-seq data for the TARGET T-ALL and NBL cohorts have been deposited in the database of Genotypes and Phenotypes (http://www.ncbi.nlm.nih.gov/gap) as part of previous projects under accession nos. phs000464 and phs000467, respectively. The WGS and RNA-seq data for the TCGA melanoma were downloaded from Genomic Data Commons data portal (https://portal.gdc.cancer.gov/legacy-archive/search/). The complete list of somatic variant calls for the 13 T-ALLs used as the input of the cis-X analysis presented in the manuscript can be accessed from our research laboratory page at http://www.stjuderesearch.org/site/lab/zhang/cis-x. Source data are provided with this paper.

Code availability

The cis-X package, together with detailed instructions and demo data, is available at https://www.stjuderesearch.org/site/lab/zhang/cis-x, https://platform.stjude.cloud/workflows/cis-x and https://github.com/stjude/cis-x. In addition to the source code, we have provided a Dockerfile along with the package to run cis-X in a container via Docker, to minimize the difficulty of running cis-X on different computing platforms.

References

51. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45, D777–D783 (2017).
52. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011).
53. Kulakovskiy, I. V. et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. 44, D116–D125 (2016).
54. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
56. Edmonson, M. N. et al. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics 27, 865–866 (2011).
57. Chen, X. et al. CONserting: integrating copy-number analysis with structural-variation detection. Nat. Methods 12, 527–530 (2015).
58. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods 8, 652–654 (2011).
59. MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 42, D986–D992 (2014).
60. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics 34, 3572–3574 (2018).
61. Parker, M. et al. C11orf95-RELA fusions drive oncogenic NF-κB signalling in ependymoma. Nature 506, 451–455 (2014).
62. Anders, S., Pyl, P. T. & Huber, W. HTSeq: a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
63. Zhang, X.-L. et al. Integrative epigenomic analysis reveals unique epigenetic signatures involved in unipotency of mouse female germline stem cells. Genome Biol. 17, 162 (2016).
64. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26, 1351–1359 (2008).
65. Zhang, Y. et al. Model-based analysis of ChIP–Seq (MACS). Genome Biol. 9, R137 (2008).
66. Cheng, Y. et al. Principles of regulatory information conservation between mouse and human. Nature 515, 371–375 (2014).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

Acknowledgements

This work was funded in part by National Institutes of Health grant nos. 1R35 CA210064-01 (to A.T.L.) and 1R01 CA216391-01A1 (to J.Z.), and Cancer Center Support Grant no. P30 CA021765 from the National Cancer Institute and the American Lebanese Syrian Associated Charities of St. Jude Children’s Research Hospital. We thank B. Abraham, M. Zimmerman, A. Durbin, D. Wheeler and D. Flasch for critically reviewing the manuscript, and C. Sherr for providing the literature relevant to p16 activation.

Author contributions

Yu Liu and J.Z. designed the cis-X software. Yu Liu, M.N.E., M.M. implemented the software. Yu Liu and J.Z. analyzed the data with help from X.C., M.N.E., K.S., X.M., Yanling Liu and M.C.R. Yu Liu, C.L., S.S. and J.Z. designed the experiments. C.L., S.S., Y.S., J.H., S.W., B.J., B.L. and J.E. performed the experiments. M.M., X.C. and L.T. tested the cis-X software. M.Q., J.J.T. and S.H. provided independent cohort validation. Yu Liu, A.T.L. and J.Z. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0659-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0659-5.
Correspondence and requests for materials should be addressed to Yu Liu or J.Z.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Transcription imbalance modeling. Cumulative distribution of transcription imbalance under binomial transcription model (dotted line), beta-binomial model as implemented in MBASED (solid line), balanced transcription model (dashed line) and experimentally observed data (dots). Different RNA-seq coverages (N=10, 50 and 100) are shown separately.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Simulation analysis of allele-specific expression detection in cis-X. Each panel represents a simulation of allelic imbalance ranging from 1:1 (no allele-specific expression) to 10000:1 (complete mono-allelic expression). Percentage of simulations identified as allele-specific expression from a group of 2,000 simulations are shown on y-axis, with plots on each panel representing simulation results with different imbalanced transcription ratio between two alleles. The imbalanced ratio of 1:1 represents the false positive rate was showed on the top, while plots in the other lines represent false negative rates of detecting transcription imbalance at various allelic ratio. Coverage for the markers in RNA-seq is shown on the x-axis. Each column, labeled by a distinct color, represents a distinct ploidy group (that is copy number alterations), while shape of each plot represents the number of markers within a gene for assessing allele-specific expression.
Extended Data Fig. 3 | Workflow for constructing the gene-specific reference expression matrix.
Extended Data Fig. 4 | LMO3 activation in T-ALL. (a) Allele specific expression of LMO3 in T-ALL SJTALL013797_D1. Eight heterozygous variants are present in LMO3 locus in this tumor, with the B-allele fractions from WGS and RNA-seq plotted on the top of the wiggle plot. (b) Outlier high expression of LMO3 was observed in this sample compared to the NCI TARGET T-ALL cohort (n = 264 samples). (c) Gene expression based clustering of the combined cohort of 13 SCMC T-ALLs and 264 NCI TARGET T-ALLs showed that SJTALL013797_D1 is clustered with other T-ALLs driven by TAL/LMO activation. The same genes from the previous study (Liu et al. Nature Genetics, 2017) were used in clustering the combined cohort. Colors on the top track represent different T-ALL subtypes.
Extended Data Fig. 5 | Somatically acquired noncoding mutation activating TAL1 in T-ALL sample SJALL018373. (a) The heterozygous C to T mutation (indicated by arrow, with mutant allele T shown in red) was only present in the tumor DNA but not in the remission sample from whole genome sequencing data. (b) H3K27Ac profile from ChIP-seq at TAL1 locus. The active enhancer present in the mutation positive PDX sample (as shown in Fig. 3d) was absent from normal T cells (CD3, CD4 and CD8) or from the T-ALL cell line (LOUCY) with no TAL1 expression.
Extended Data Fig. 6 | Activating deletion upstream PRLR. Expression (FPKM on y-axis) of SPEF2 (a) and IL7R (b) in the T-ALLs. The 3 tumors carrying the focal deletions (SJALL043558_D1, PATFYZ, and PATRUN) are labeled. (c) H3K27Ac profiles from ChIP-seq show active enhancer upstream of IL7R in the PDX (derived from patient SJALL018373) and a T-ALL cell line (KOPT-K1) having high IL7R transcription; both samples have the wild-type allele at this locus.
Extended Data Fig. 7 | Analysis of pediatric neuroblastoma with cis-X. (a) Copy number variations identified in the four neuroblastoma cell lines. The blue and red colors represent the deletion and amplifications, respectively, identified in these cell lines. (b) Circos plot showing the cis-activating structural rearrangements identified in NBL cell lines by cis-X. The copy number alterations in each genome are shown in the inner track, with blue lines representing a copy number of 1 and red a copy number of three. The cis-activating structural variants are shown as links in the middle of the plot, with purple links representing inter-chromosome translocations and green for intra-chromosome translocations. The target genes activated by these rearrangements are labeled on the outer track of each plot.
Extended Data Fig. 8 | TERT cis-activation by somatic non-coding variants in neuroblastoma. The analysis was based on 90 NBL primary tumor samples with matching RNA-seq and WGS from TARGET, 42 of which had positive immune cell infiltration signature based on prior analysis (Ma et al, Nature, 2018). (a) Samples with somatic copy number alterations (CNA, marked by red or blue blocks) or/and structural variations (SVs, marked by circles) at TERT locus. All except for one (PARAMT, marked #) were detected by cis-X as cis-activated candidates. Samples marked with * have immune cell infiltration signature. Samples highlighted in gray are used to illustrate allele-specific expression (ASE) below. (b) Examples of ASE detected in neuroblastoma with or without infiltrating immune cells. Variant allele fraction in DNA (by WGS) and RNA (by RNA-seq) of SNPs, depicted as bar graph, demonstrates that ASE analysis is not affected by the presence of immune cell infiltration signature in tumor samples.
Extended Data Fig. 9 | TERT expression in melanoma and neuroblastoma. TERT expression in adult TCGA melanoma (MEL) samples (n = 38), pediatric neuroblastoma (NBL) patient samples from TARGET project (n = 90) and cell lines (n = 4) analyzed in this study. The MEL samples were color-coded by TERT promoter mutation status while the NBL samples were marked by the status of cis-activation, infiltrating immune cells and cell-lines as depicted in figure legend.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No softwares were used for data collection.

Data analysis

We developed cis-X which is available through GitHub (https://github.com/stjude/cis-x). Softwares used in cis-X pipeline include perl (ver 5.10.1), R (ver 3.1.0), Java (ver 1.8.0_66), twoBitToFa, bedtools (v2.25.0), and FIMO. Other softwares used in the analysis include bwa (v0.7.12), SPP (v1.10.1), MACS(v2.1.1.20160309), Picard, Bambino, CREST, CONserting, StrongArm, HTseq-count (0.11.2), AnnotSV (v2.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Whole genome sequencing and RNA-seq data for the SCMC cohort analyzed in this study can be accessed from Genome Sequence Archive for Human under National Genomics Data Center of China (http://bigd.big.ac.cn/gsa-human), with accession number HRA000097 and HRA000096 for WGS and RNA-seq respectively. Chip-seq data generated in this study can be accessed from GEO under accession number GSE113565 and GSE145549, for H3K27Ac and YY1 respectively, with the called peaks (in bed format) available upon request. WES and RNA-seq data for TARGET T-ALL and NBL cohort are in the database of genotypes and phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) under accession number phs000464 (T-ALL) and phs000467 (NBL). The WGS and RNAseq data for TCGA melanoma are collected from GDC (https://portal.gdc.cancer.gov/legacy-archive/search/f). The complete list of somatic variant calls for the 13 T-ALLs used as input of cis-X analysis...
presented in the manuscript can be accessed from our research lab page at http://www.stjuderesearch.org/site/lab/zhang/cis-x.

Databases used include: COSMIC (https://cancer.sanger.ac.uk/cosmic), ENCODE (https://www.encodeproject.org/), HOCOMOCO (https://hocomoco.autosome.ru/), DGV (http://dgv.tcag.ca/dgv/app/home), Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/data/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Sample size**
We analyzed data from 3 cohorts in this study. (1) 13 T-ALL diagnosed and treated in Shanghai Children’s Medical Center; (2) 90 Neuroblastoma samples from TARGET project; (3) 38 Melanoma samples from TCGA project. We included all patients in the analysis with matched whole genome sequencing data for both tumor and normal samples and transcriptome sequencing data for the tumor samples available.

**Data exclusions**
No data was excluded from this analysis.

**Replication**
The luciferase assay experiments were replicated by at least two independent experiments. Two different mapping methods in analyzing RNA-seq data were applied in the allelic specific expression analysis. Results from the two methods are highly consistent, in terms of both the heterozygous markers detected (with a median of 97.7% markers could be detected from both methods) and the variant allele frequency from these analysis as described in Supplementary Note.

**Randomization**
No randomization was performed in this study. Patients are grouped by disease types in noncoding variant analysis. ChIP-seq and luciferase assay experiments were grouped by biological controls.

**Blinding**
Investigators were not blinding to the tumor types during the analysis. Experiment and input groups in ChIP-seq were blinded before analysis. The variants including negative controls were blinded in luciferase assay.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

**Materials & experimental systems**

| n/a | Involved in the study |
|-----|------------------------|
| ☐   | Antibodies             |
| ☒   | Eukaryotic cell lines  |
| ☐   | Palaeontology and archaeology |
| ☐   | Animals and other organisms |
| ☐   | Human research participants |
| ☒   | Clinical data          |
| ☐   | Dual use research of concern |

**Methods**

| n/a | Involved in the study |
|-----|------------------------|
| ☐   | ChIP-seq               |
| ☒   | Flow cytometry         |
| ☒   | MRI-based neuroimaging |

**Antibodies**

**Antibodies used**
Human H3K27Ac antibody: ab4729, Abcam, Human YY1 antibody: #61779, active motif.

**Validation**
Validation of H3K27Ac and YY1 antibody for the application and species are provided by the manufacturer. H3K27Ac: https://www.abcam.com/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.pdf YY1: https://www.activemotif.com/documents/tds/61779.pdf

**Eukaryotic cell lines**

Policy information about cell lines

**Cell line source(s)**
Jurkat cell is purchased from ATCC. GIMEN, NB16, NB1643 and KELLY cells are provided by Dr. Thomas Look at Dana-Farber Cancer Institute.
Authentication: Jurkat cell is authenticated by ATCC with no further authentication. GIMEN, NB16, NB1643 and KELLY were authenticated with STR in Dana-Farber Cancer Institute.

Mycoplasma contamination: All cell lines were tested negative for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register): No commonly misidentified cell line was used in this study.

Animals and other organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

Laboratory animals: NOG mice (3-4 weeks old, female) were used for generating patient derived xenograft model.

Wild animals: No wild animal was included in this study.

Field-collected samples: No filed-collected sample was included in this study.

Ethics oversight: The protocol was approved by the Committee on the Ethics of Animal Experiments of Crown Bioscience (Crown Bioscience IACUC Committee).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics: 13 patients diagnosed as T-ALL in Shanghai Children's Medical Center and treated following standard procedure were included. Patients included 12 male and 1 female with median age of 7.15 year-old (ranging from 3 to 14.87 year-old). All 13 patients are Asian.

Recruitment: Patients involved in this study were treated and recruited in Shanghai Children's Medical Center, Shanghai, China. No selection applied to patient recruitment.

Ethics oversight: The study was approved by the Shanghai Children's Medical Center Institutional Review Boards (IRB).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

Confirm that both raw and final processed data have been deposited in a public database such as GEO.

Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication. Data are deposited in GEO, with accession GSE113565 and GSE145549.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113565
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145549

Files in database submission

H3K27Ac ChIP-seq: AL5570ChIP_R1.Merge.bw, AL5570Input.bw, AL5570ChIP_R1.Merge.fq.gz, AL5570Input_R1.fq.gz
YY1 ChIP-seq: YY1_ChIPseq_treat_pileup.bw, YY1_ChIPseq_control_lambda.bw, 5570_P2_CHIP_S9_R1_001.fq.gz, 5570_P2_INPUT_S10_R1_001.fq.gz

Genome browser session (e.g. UCSC): Not applicable.

Methodology

Replicates: No replicate for ChIP-seq experiment in this study.

Sequencing depth:

H3K27Ac ChIP-seq: Single-end read with 50bp length. Total number of reads are 80764692 (experiment) and 11330542 (input); uniquely mapped reads: 58771424 (experiment) and 7820256 (input).
YY1 ChIP-seq: Single-end read with 75bp length. Total number of reads are 36174891 (experiment) and 22439779 (input); uniquely mapped reads: 30952493 (experiment) and 18888645 (input).

Antibodies: Human H3K27Ac antibody: ab4729, Abcam, Human YY1 antibody: #61779, active motif.

Peak calling parameters: MACS2 parameters: -q 0.05 --nomodel --extsize 180

Data quality: Input control was used to evaluate peak enrichment in MACS2 with a q cutoff value of 0.05. A total of 25775 and 19617 peaks were called in H3K27Ac and YY1 ChIP-seq experiment, respectively.
Software

Analysis was carried out with ChIP-seq analysis pipeline on St. Jude Cloud (https://platform.stjude.cloud/tools/chip-seq).