Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers

Bing He¹, Peng Gao¹, Yang-Yang Ding¹,², Chia-Hui Chen¹, Gregory Chen³, Changya Chen¹, Hannah Kim¹, Sarah K. Tasian¹,²,⁷, Stephen P. Hunger¹,²,⁷, Kai Tan¹,²,⁴,⁵,⁶,⁷

¹ Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
² Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
³ Medical Scientist Training Program, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
⁴ Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA
⁵ Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
⁶ Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
⁷ Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Correspondence: Kai Tan, tank1@email.chop.edu

Disclosure of potential conflicts of interest

The authors declare no competing interests.
Abstract

Interpreting the function of noncoding mutations in cancer genomes remains a major challenge. Here we developed a computational framework to identify risk noncoding mutations of all classes by joint analysis of mutation and gene expression data. We identified thousands of SNVs/small indels and structural variants as candidate risk mutations in five major pediatric cancers. We experimentally validated the oncogenic role of CHD4 overexpression via enhancer hijacking in B-ALL. We observed a general exclusivity of coding and noncoding mutations affecting the same genes and pathways. We showed that integrated mutation signatures can help define novel patient subtypes with different clinical outcomes. Our study introduces a general strategy to systematically identify and characterize the full spectrum of noncoding mutations in cancers.
Introduction

Extensive research efforts have identified recurrent mutated genes in multiple types of childhood cancer including several recent pan-cancer studies which have defined landscapes of coding mutations in common pediatric cancers (1-4). However, current cancer mutation landscapes are far from complete without systematic analysis of the noncoding portion of the genome. For instance, recurrent leukemia-associated genetic alterations cannot be identified in 10%~20% of children with acute leukemias (1,3-5), thereby making it challenging to design targeted therapies for such patients.

The vast majority of somatic mutations in cancer genomes occur in noncoding regions because 98% of the human genome is made up of noncoding sequences, and the somatic mutation rate of noncoding regions is similar to that of coding regions (1). The Therapeutically Applicable Research To Generate Effective Treatments (TARGET) project has sequenced over 1000 genomes from five common pediatric cancers. Similarly, The Cancer Genome Atlas (TCGA) project has molecularly characterized over 20,000 primary cancer genomes spanning 33 types of adult cancers. Despite this rapid accumulation of whole genome sequences (WGS) for both pediatric and adult cancers, identification and interpretation of the functional impact of noncoding mutations remains challenging.

Noncoding regulatory sequences, particularly enhancers and promoters, are key determinants of tissue-specific gene expression. Multiple mutation types have been reported to disrupt enhancers and promoters and expression of their target genes, including single nucleotide variants (SNVs), small insertion and deletions (indels), and large structural variants (SVs), including deletions, insertions, duplications, inversions, and translocations. Most prior studies of noncoding mutations have focused on SNVs and small indels and revealed a number of noncoding causal mutations. One example is the SNP rs2168101 G>T located in the enhancer of LMO1, which disrupts GATA3 binding and LMO1 expression in neuroblastoma patients (6). Another example is the heterozygous indels (2-18 bp) located at -7.5 kb from TAL1 transcription start site in T-acute lymphoblastic leukemia (T-ALL), which introduces new binding sites for MYB to create a super-enhancer upstream of TAL1 (7).

Only about 30% of SVs result in in-frame gene fusions that join the protein coding regions of two genes (8). The majority of SV break points are located in noncoding regions and do not change gene structure. Although less well studied compared to SNVs and small indels, several seminal studies have revealed oncogenic roles of noncoding SVs by redirecting enhancers/promoters to oncogenes or from tumor suppressor genes (9,10). Such enhancer rearrangement events have been identified in diverse cancer types, suggesting that this could be a common mechanism of oncogenesis. For instance, interstitial deletion in the pseudoautosomal region 1 (PAR1) of the X/Y chromosomes places the transcription of CRLF2 (cytokine receptor-like factor 2) under the control of the P2RY8 (purinergic 2 receptor Y 8) enhancer in Philadelphia chromosome-like and Down Syndrome-associated B-ALL (11). Other examples
include complex structural variants in childhood medulloblastoma, which rearranges the DDX31 (DEAD-Box Helicase 31) enhancer to GFI1B (growth factor independent 1B) (9), and t(3;8) in B-cell lymphoma that rearranges the BCL6 enhancer to MYC (10).

Given the prevalence of noncoding mutations and the drastic increase of whole genome sequencing data, novel computational methods are critically needed to systematically identify risk noncoding mutations. Here, we introduce the PANGEA method (Predictive Analysis of Noncoding Genomic Enhancer/promoter Alterations), a general computational framework for systematic analysis of noncoding mutations and their impact on gene expression. This method simultaneously identifies all classes of somatic mutations that are associated with gene expression changes, including SNVs, small indels, copy number variations, and structural variants. Using PANGEA, we have conducted a pan-cancer analysis of noncoding mutations in 501 pediatric cancer patients of five histotypes with matched WGS and RNA-Sequencing (RNA-Seq) data generated by the TARGET project. We identified a comprehensive list of recurrent noncoding mutations as candidate risk mutations in these cancers. An integrated analysis of both coding and noncoding mutations revealed distinct pathways affected by either coding and noncoding mutations. We also show that integrated mutation signatures can help define novel patient subtypes with different clinical outcomes.

Results

A comprehensive catalog of recurrent noncoding mutations

The TARGET data portal (https://ocg.cancer.gov/programs/target/data-matrix) contains matched WGS and RNA-Seq data for 501 patients, including 163 patients with B-cell acute lymphoblastic leukemia (B-ALL), 153 patients with acute myeloid leukemia (AML), 100 with neuroblastoma (NBL), 53 with Wilms tumor (WT), and 32 with osteosarcoma (OS). All patients have WGS data for both tumor and germline or remission samples, which defined alterations as germline or somatic.

We identified somatic SNVs and small indels (< 60bp) for the five cancer types using tumor and remission samples (Figure S1A). We evaluated the accuracy of our mutation calling pipeline using two approaches. First, we used a benchmark set generated by Zook et al. (12). It consists of a set of high-confidence SNVs and indels in subjects NA12878 and NA24631 from the 1000 Genome Project, identified by integrating 14 data sets using five sequencing technologies, seven read mappers and three variant callers. Using this benchmark set, we estimated the precision of our pipeline to be ~95% (Table S1). Second, the TARGET project provides a list of high-confidence SNV calls in B-ALL, AML, NBL, and WT that were validated using multiple experimental protocols including whole exome sequencing (WES), RNA-Seq and targeted...
sequencing. Of the 735 high-confidence calls, our analysis pipeline identified 681 (92%) of them, suggesting high sensitivity of our pipeline (Figure S1B).

Across the five cancer types, the average mutation rate ranges from 0.16 SNV/indel per million bases (Mb) in AML to 0.55 per Mb in OS (Figure S1C). The higher somatic mutation rate in OS is consistent with a previous WGS study, in which regional clusters of hypermutation, termed kataegis, were identified in 85% of pediatric OS patients (13). As expected, over 98% of the identified mutations are located in noncoding regions, while less than 1% of the mutations are located in coding regions (Figure S1D).

Next, we used Delly2 (14) and Lumpy to identify structural variants (SVs), including large deletions (DELs), tandem duplications (DUPs), inversions (INVs), and translocations (TRANs). We only kept SVs that were called by both methods and passed additional filtering criteria (Figure S2A) as the final set of identified SVs. In total, we identified 26,757 SVs across 501 patients. The average number of identified SVs per patient ranges from 18 in WT to 71 in OS (Figure S2B). These numbers are comparable with published data by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) consortium. Previously, Roberts et al. identified and experimentally validated 21 SVs that joined exons of two genes in frame in B-ALL patients (15). We identified 20 of those SVs, suggesting high sensitivity of our pipeline. Among the identified SVs, 9,831 are in-frame changes and potentially generate fusion genes. We evaluated the fusion gene predictions by comparing to matched RNA-Seq data from the patients. Of the 9,831 fusion genes predicted by our pipeline, 2,323 of them were supported by RNA-Seq data from the same patients. The TARGET consortium experimentally tested 12 known translocations in 212 leukemia patients. Based on this set of validated SVs, our SV calling pipeline has an accuracy of 98% (Figure S2C, Table S2). Taken together, using multiple benchmarking data sets, we show that our SV calling pipeline is highly accurate.

**Noncoding mutations disrupting enhancer/promoter sequences or enhancer-promoter interactions**

The functional consequence of noncoding mutations is difficult to interpret without comprehensive annotation of noncoding regulatory DNA sequences in cancer genomes. To this end, we used publicly available epigenomic data of disease-relevant cell types to construct an enhancer catalog for the five cancer types in this study (see Methods). Specifically, we used ChIP-Seq data of histone modification marks (H3K4Me1, H3K4Me3, H3K27Ac) to predict transcription enhancers using the chromatin signature identification by artificial neural network (CSI-ANN) algorithm (Table S3) (16). In total, we identified 282,021 enhancers for the five cancer types at the False Discovery Rate (FDR) of 0.05 (Figure S3A). Overall, 91% of predicted enhancers are supported either by ATAC-Seq data in cancer-relevant cell types or by sequence conservation across 20 mammalian genomes (Figure S3A) or both, suggesting high quality of the predicted
enhancers. Next, we predicted target gene(s) of each enhancer using the Integrated Method for Predicting Enhancer Targets (IM-PET) algorithm (17), using public histone modification ChIP-Seq and RNA-Seq data in disease-relevant cell types. In total, we predicted 635,096 enhancer-promoter (EP) pairs across the five cancer types (Figure S3A). We compared the predicted EP pairs with published high-resolution Hi-C/ChIA-PET data in human B cells, myeloid cells and kidney cells (Table S3, Methods). Seventy-four percent (317,698) of our EP predictions are supported by either Hi-C or ChIA-PET data (Figure S3B), suggesting high quality of our predictions.

To identify recurrent mutations that disrupt either enhancer/promoter sequences or enhancer-promoter interactions, we intersected the catalog of somatic mutations with the catalogs of enhancers/promoters and EP interactions. Across the five cancer types, 16% SNVs or indels overlap with enhancers/promoters in cell types relevant to a given pediatric cancer. In comparison, only 8% of all SNVs across 34 TCGA cancer types (as a control set) overlap with enhancers/promoters in the relevant cell types (average p-value = 5e-13) (Figure S4A). The identified CNVs also have significantly higher overlap with the enhancers in relevant cell types compared to all CNVs reported by the TCGA consortium (15% vs 7%, average p-value = 2e-11) (Figure S4B). The break points of inversions and translocations are significantly enriched between predicted EP pairs compared to all break points reported by the TCGA consortium (45% vs 36%, average p-value = 1e-18) (Figure S4C). Taken together, these results suggest that various types of somatic mutations can disrupt the cis-regulatory landscapes of pediatric cancers, especially cis-regulatory elements involved in regulating tissue-specific gene regulation of the specific tumor histotype.

Systematic identification of candidate risk noncoding mutations

To help prioritize noncoding mutations, we developed the PANGEA method to systematically identify recurrent noncoding mutations that disrupt the transcriptional regulation of a gene. Unlike previous methods, PANGEA uniquely considers all types of noncoding mutations, including SNVs, small indels, CNVs, and SVs. These mutation types can either disrupt enhancer/promoter sequences or disrupt the interactions between enhancers and promoters which is critical for transcription activation. After tabulating all such mutations, we use weighted elastic net to perform a regression analysis of gene expression on the sets of noncoding mutations across the patient cohort (Figure 1A). Because the EP interactions are predicted computationally, we use EP prediction scores as the weights for each predictor (mutation) in order to include a confidence measure of EP interactions in our regression model. Candidate risk mutations are predicted based on the statistical significance of the corresponding regression coefficients (see Methods for details). The PANGEA software package is available at https://github.com/tanlabcode/PANGEA.
Using multiple-testing adjusted p-value < 0.05 as the cutoff, we identified 1,405 genes whose expression changes can be predicted by SNVs/small indels in their enhancers/promoters, 55 genes whose expression changes can be predicted by CNVs in their enhancers, 1,082 genes whose expression changes can be predicted by SVs that disrupt their EP interactions (Table S4). In total, disruption of enhancer function by recurrent noncoding mutations were found in 477 out of 501 pediatric cancer patients (95%) (Figure 1B). The quantile–quantile plot shows large deviation of the observed P values for the predicted risk noncoding mutations compared to those of random expectations using an independent ICGC cohort (n=2715 donors), suggesting low false prediction rate (Figure S5, Supplementary Methods). Over half of the predicted risk noncoding mutations are SNVs and small indels (66%), followed by translocations (27%) and other types of SVs (Figure 1C). However, when adjusted by the overall frequency of each type of mutation, SVs become the most frequent type of risk noncoding mutations (Figure S6).

The enhancers that are affected by risk noncoding mutations are more cell-type-specific compared to all enhancers in our enhancer catalog (Figure 1D). In addition, these enhancers have more overlap with published super enhancers in cell types that are relevant to the given cancer type (18) (Figure 1E). Finally, these enhancers have high levels of sequence conservation across 20 mammalian species (Figure 1F). Taken together, these results provide additional support to the predicted risk noncoding mutations.

**Risk enhancer rearrangements**

Several seminal studies have reported ‘enhancer hijacking’, also known as oncogenic rearrangement of enhancers due to translocation/inversion (9,10,19-22). In this study, we performed a systematic analysis of enhancer hijacking events across the five cancer types. Amongst 501 total patients, 405 patients (81%) have at least one predicted enhancer hijacking event in their genomes. We identified several known enhancer hijacking events. For example, we found the t(14;X)(q32;p22) translocation in 12 B-ALL patients, which hijacks multiple enhancers of IGHV to the vicinity of CRLF2 resulting in significant overexpression of CRLF2 in these patients (Figure S7A). We also discovered enhancer hijacking from three different genomic loci to the common TERT gene locus (t(10;5)(p22;p15), t(5;5)(q34;p15), and t(5;5)(q12;p15) (Figure S7B). These TERT-related enhancer hijacking events were observed in 15 neuroblastoma patients and the resultant translocations led to significant overexpression of TERT in these patients (Figure S7B).

However, the majority of our predicted enhancer hijacking events have not been previously reported. One such event involves hijacking of enhancers to upregulate CHD4, which has three translocation partners (t(12;22)(p13;q13), t(12;19)(p13;p13), t(9;12)(p24;p13)). In total, 12 B-ALL patients had this translocation. Most translocations in this region result in fusion genes involving a nearby gene ZNF384. A previous study has focused on the function of ZNF384
fusion genes (23). However, in the TARGET cohort, we observed 2 patients in which the translocation break point is actually located downstream of ZNF384 and did not generate an ZNF384 fusion (Figure 2A). Furthermore, ZNF384 fusion is not correlated with patient prognosis. Together, these data suggests a driver of oncogenesis that is independent of ZNF384. In all patients with the translocation, we found that enhancers such as that of EP300, TCF3, and SMARCA2 are hijacked to chr12:p13. The expression level of CHD4, a gene adjacent to ZNF384, is significantly increased in these patients (Figure 2B). Similarly, in a recent published mixed phenotype acute leukemia (MPAL) cohort, we also observed CHD4 expression increase in the patients with ZNF384-involved rearrangement (3) (Figure S8). Moreover, B-ALL patients with the enhancer hijacking event have significantly shorter time to relapse (Figure 2C). We thus hypothesize that enhancer hijacking translocation events that bring potent enhancers to the promoter of CHD4 may be an independent oncogenic driver in B-ALL.

CHD4 (chromodomain-helicase-DNA-binding protein 4) is a component of the nucleosome remodeling and deacetylase (NuRD) complex, which plays an important role in B-cell development by regulating B-cell-specific transcription (24). In addition, CHD4 is known to function as a repressor of several tumor suppressor genes, and inhibition of CHD4 reduces the growth of AML and colon cancer cells (25,26). These data suggest a role of CHD4 in the oncogenesis of B-ALL. To investigate the potential role of CHD4, we first identified differential expressed genes in the patients with CHD4 overexpression. In total, there are 666 up-regulated and 922 down-regulated genes in patients with CHD4 rearrangements (Table S5). Using published TF ChIP-Seq data in human GM12878 cells (27), we found CHD4 binding sites are significantly enriched at enhancers or promoters of the down-regulated genes (Figure S9A). Among the 157 down-regulated genes with CHD4 binding sites, several of them encode well-known regulators of B cell development including PAX5, IRF4, TCF3 and EBF1 (Figure S9B, C). These results suggest a role of CHD4 in B-ALL through regulating the expression of key TFs in B-cell development.

To test the potential oncogenic role of CHD4 in B-ALL, we knocked out CHD4 in the NALM-6 and REH B-ALL cell lines. We performed growth competition assays to compare the growth phenotypes of CHD4 knockout with non-knockout leukemic cells. Both NALM-6 and REH cells showed impaired growth with CHD4 knockout (Figure 2D). Next, we introduced the translocation t(6;15)(qF2;qE1) in murine Ba/f3 cells using CRISPR-Cas9. The translocation break points were designed to be located in the intergenic regions near CHD4 and EP300, placing the EP300 enhancer (Figure 2E) upstream of the CHD4 promoter without creating a fusion gene (Figure S10A). In Ba/f3 cells with the translocation, we observed increased expression of CHD4 at both mRNA and protein levels (Figure 2G, H). In contrast, expression of nearby genes was not increased in cells with the translocation (Figure 2G). More importantly, the introduced translocation enables Ba/f3 cells to proliferate in the absence of IL-3 (Figure 2H), suggesting oncogenic transformation. Additionally, the expression levels of TCF3, PAX5, and EBF1 were significantly decreased in the cells with
CHD4 rearrangement, suggesting the role of CHD4 in regulating those key regulators (Figure S10B). In summary, these results strongly support an oncogenic role of CHD4 in B-ALL due to enhancer hijacking.

Risk enhancer copy number alterations

We found that 309 patients in the TARGET cohort (62%) had enhancer amplification/deletion and associated target gene expression change. MYCN is frequently amplified in neuroblastoma patients (28). In the NBL cohort, we observed that the body of the MYCN gene is amplified in 34 patients. However, we also observed 11 patients with amplification of the MYCN enhancer rather than the gene body (Figure 3A). MYCN expression level is significantly higher in those patients compared to the patients without MYCN amplification (Figure 3B). Moreover, patients with only MYCN enhancer amplification have shorter time to relapse compared with other patients (Figure 3C), including patients with MYCN gene body amplification. This result suggests that enhancer amplification alone is sufficient to up-regulate MYCN and drive aggressive neuroblastoma. Another example involves deletion of the ATG3 enhancer in 15 B-ALL patients (Figure S11A), resulting in decreased ATG3 expression in those patients (Figure S11A). Autophagy related 3 (ATG3) is a ubiquitin-like-conjugating enzyme and plays a role in the regulation of autophagy; downregulation of ATG3 has been reported in myelodysplastic syndrome (MDS) patients progressing to leukemia (29).

Risk enhancer/promoter SNVs and small indels

We found that 346 patients in the TARGET cohort (69%) had SNVs/small indels located in enhancer/promoter regions that caused target expression change. For instance, we found 6 AML patients who have SNVs in the GFI1B +11k enhancer (30). The GATA2 binding sites in the enhancer were disrupted and GFI1B expression decreased correspondingly (Figure 3D, E). Growth factor independence 1b (GFI1B) encodes a key transcription factor regulating dormancy and proliferation of hematopoietic stem cells (HSCs) and the development of erythroid and megakaryocytic cells (31,32). Recent studies had revealed its critical role as a tumor suppressor in AML, as low GFI1B expression is associated with poor patient survival (33). Consistent with previous studies, the 6 patients with GFI1B enhancer mutation have significantly shorter time to relapse (Figure 3F). Another example involves 9 AML patients who have mutations in the IDH2 +56k enhancer (Figure S11B). This enhancer is part of a known super enhancer of IDH2(34), and it is constitutively active in myeloid cell types. Previously, nonsynonymous mutations of IDH2 have been reported in 9%-19% of adult AML patients, but relatively rare in childhood AML (<4%) (35,36). Mutations in IDH2, as well as in several other genes involved in regulation of DNA methylation (such as DNMT3, TET2, IDH1), were detected to have higher variant allele frequency (VAF) in adult AML patients, suggesting they were early
mutational events in AML (37). We did not identify nonsynonymous IDH2 mutations, but did find recurrent mutations in the IDH2 enhancer. The mutations are predicted to disrupt ERG/FLI1 and E2F4 binding sites in the enhancer. The expression level of IDH2 is significantly lower in patients with these mutations. Taken together, these results suggest a different mechanism of IDH2 disruption in pediatric AML that was previously unappreciated. We also found 4 patients who have SNVs in the GATA2 +126k enhancer and 2 patients who have SNVs in the GATA2 promoter (Figure S11C). The enhancer was previously reported to be rearranged to vicinity of EVI1 in adult AML patients (38). The mutations were predicted to disrupt FUBP1 binding site, and the expression of GATA2 was significantly lower in the patients with mutations in enhancer/promoter regions.

Coding and noncoding mutations affect distinct sets of genes and pathways

In total, our analysis of five pediatric cancer types have identified 1,175 genes recurrently altered in their coding regions, and 2,162 genes recurrently altered in their noncoding regions. Surprisingly, the overlap between the two groups of genes is very small (62 genes, 2%), suggesting general exclusivity of coding and noncoding mutations affecting a given gene (Figure 4A). We investigated the genomic features of the genes affected by coding versus noncoding mutations. We found that genes affected by coding mutations are longer and have higher level of sequence conservation (Figure 4B, C). On the other hand, genes affected by noncoding mutations are linked to more regulating enhancers (Figure 4D), and their expression is more tissue-specific (Figure 4E). In addition, the regulating enhancers of those genes are more conserved (Figure 4F). Previous studies have suggested SNVs and small indels occur more frequently in genomic regions with late replication timing, while translocations and inversions occur more frequently in the genomic regions with early replication timing (39). Consistent with this trend, we found that genes affected by SNVs and small indels in coding regions are located in the relatively late replicating regions, while the genes in other groups have relatively earlier replication timing (Figure 4G). Pathway analysis reveals signaling pathways, including JAK/STAT, MAPK and Wnt signaling pathways, are mostly affected by coding mutations (Figure S12). In stark contrast, genes in metabolic pathways are mostly affected by noncoding mutations (Figure S12). Correspondingly, we found metabolic genes tend to be located in the early replicating region (Figure S13). Taken together, these data suggest the exclusivity of genes affected by coding versus noncoding mutations is likely due to the different genomic location and features of those genes (Figure 4H).

Level of TF regulon disruption correlates with patient prognosis
Genes encoding lineage-specific transcription factors (TFs) are frequently mutated in pediatric cancers. For example, TFs that regulate B-cell development including IKZF1, EBF1, PAX5, TCF3 are frequently altered in B-ALL patients (40-43). In addition, MYCN and ZNF281 are known prognostic markers for neuroblastomas (44,45), while RUNX1 and CBFB are also frequently altered in pediatric AML and are linked to unfavorable clinical outcomes (46).

In contrast to coding mutations of TFs, regulatory output of TFs can also be disrupted by mutations affecting individual enhancers/promoters/EP interactions regulating the target genes. Therefore, we compared the effects of coding and noncoding mutations on TF regulons (defined as the set of genes regulated by a TF) and patient outcome. For each cancer type, we report the top 10 most frequently affected regulons by combined coding and noncoding mutations. Most of the identified TFs have a known role in the specific cancer type (Figure 5A and Table S6). Furthermore, We found that regulon disruption by noncoding mutations occur more frequently than regulon disruption by coding mutations of the TF genes (Figure 5B). This is probably because mutations of a TF gene lead to a bigger disruption of the TF regulon, compared to mutations disrupting regulation of individual TF targets.

We hypothesized that the level of regulon disruption is correlated with patient disease outcome. To test this hypothesis, we plotted time to relapse stratified by the degree of regulon disruption (i.e., TF coding mutation versus disruption of individual TF targets by noncoding mutations). Indeed, we found a strong correlation between time to relapse and the degree of regulon disruption. For instance, in B-ALL, patients with PAX5 deletion have the shortest time to relapse, followed by patients with disruption of PAX5 EP interactions. Finally, patients without any PAX5 regulon mutation have the longest time to relapse (Figure 5C). We found the same correlations for RUNX1 regulon in AML (Figure 5D), and WT1 regulon in WT (Figure 5E). For PAX5, RUNX1, and WT1, the mutations are all loss-of-function of the TFs. We also found correlations involving gain-of-function mutations of the TFs, including TCF3 in B-ALL and MYCN in NBL. For TCF3, because the fusion event leads to gain-of-function of TCF3, patients with disruption of TCF3 EP interactions have longer time to relapse compared to the patients without any TCF3 regulon mutation (Figure 5F).

Presumably, the gain-of-function of TCF3 partially mitigate the effect of disruption of TCF target genes due to noncoding mutations. Similarly, NBL patients with disruption of MYCN EP interactions have longer time to relapse compared to the patients without any MYCN regulon mutation (Figure 5G). In summary, our data suggest that there is typically a wide spectrum of regulon disruption for key TFs in pediatric cancers. And the level of TF regulon disruption is associated with patient prognosis.

**Integrated mutation profiles suggest novel disease subtypes**

Mutational profiles of coding regions have been widely used for patient stratification and prognosis purposes. This is not the case for noncoding
mutational profiles. We therefore constructed mutational profiles of the five pediatric cancers using both coding and noncoding mutations. Clustering analysis of these combined mutational profiles enabled us to discover novel patient subtypes.

For B-ALL, 7 patient clusters are identified (Figure 6A). Of the 7 clusters, four are characterized by known translocation events in B-ALL including ZNF384 rearrangement, ETV6-RUNX1, TCF3-PBX1, and IGHV translocation. All these are previously reported to be recurrent SVs in B-ALL (47,48). However, the mechanism by which these fusion proteins contribute to leukemogenesis has not been fully elucidated. Here, we found that in addition to creating fusion genes, the translocations can alter the expression of nearby genes by re-arranging locations of distal enhancers (Figure 6A, affected genes listed below heatmap).

Two other clusters (C5 and C6) are characterized by novel inversion or translocation. Patients in C5 have Inv(2) (n=14). The inversion is predicted to hijack an enhancer to near SUPT7L and ATRAID and up-regulate the expression of these two genes (Figure S14A, B). SUPT7L is a subunit of the SPT3-TAFII31-GCN5L acetylase (STAGA) complex, which is known to regulate the stability of the TCF3-PBX1 oncoprotein in ALL(49). Interestingly, we found Inv(2) significantly co-occurs with TCF3-PBX1 in 6 patients (Figure S14C). Both TCF3-PBX1 and Inv(2) are associated with aggressive clinical outcome in B-ALL. Patients who have both mutations show even shorter time to relapse compared to patients with either type of mutation (Figure S14D), suggesting synergy of these two pathways contributing to disease outcome.

Patient cluster C6 is characterized by the inter-chromosomal translocations at chr2 (t(2;7)(q21;q11), t(2;11)(q21;q11)). The translocations occur in 15 patients, and are predicted to disrupt the EP interaction involving the ANKDR30BL gene (Figure S15A), resulting in significant decrease of ANKDR30BL expression in patients with this translocation. Interestingly, MIR663B, a microRNA located in the intron of ANKDR30BL, is also down-regulated in those patients (Figure S15B). MIR663B is known to regulate the expression of CCL17, CD40, and PIK3CD in chronic lymphocytic leukemia (CLL) (50). Consistent with the previous finding, we observed significant expression increase of those genes in patients with MIR663B down-regulation (Figure S15B).

We also identified two potential novel cancer subtypes in NBL (C6 and C7, Figure 6C). Cluster 7 consists of 9 NBL patients with translocation on chr17 that results in enhancer rearrangement and down-regulation of ERBB2 (t(1;17)(p33;q12), t(9;17)(p21;q12), t(11;17)(q13;q12)) (Figure S16A, B). ERBB2 is essential for normal embryonic development and has a critical function in oncogenesis and progression of several cancer types including breast cancer, lung cancer, and leukemias (51-53). In addition, low ERBB2 expression is associated with poor patient survival in NBL (54). Consistently, patients in cluster 7 have shorter time to relapse (Figure S16C). The other novel NBL subtype (C6) is characterized by amplification of the TGM6 enhancer in 11 NBL patients and consequently increased TGM6 expression in these patients (Figure S16D, E). Transglutaminase 6 (TGM6) is a protein associated with nervous system
development (55). Transglutaminases, particularly TGM2, is known to play important roles in neurite outgrowth and modulation of neuronal cell survival (56). Our result suggests a potential oncogenic role of TGM6 in NBL.

A novel AML subtype (cluster C5) is characterized by enhancer deletion of ZNF37A in 16 patients (Figure S17A). ZNF37A is involved in fusion events in breast cancer (57) and adult AML (58). Notably, AML patients with ZNF37A enhancer deletion have lower ZNF37A expression and shorter time to relapse (Figure S17B, C), suggesting the prognostic significance of this noncoding mutation.

A novel WT subtype (cluster C4) is characterized by enhancer rearrangement of GAS6 in 6 patients (Figure S17D). Growth arrest specific 6 (GAS6) is a ligand for receptor tyrosine kinases AXL, TYRO3 and MER whose signaling is implicated in cell growth and survival (59,60). Patients with the translocation have lower GAS6 expression and shorter time to relapse (Fig S17E, F).

Discussion

The landscape of noncoding mutations in pediatric cancers has not been comprehensively characterized. Here, we developed PANGEA, a general regression-based method to identify all classes of candidate risk noncoding mutations by joint analysis of patients’ mutations and gene expression profiles. Application of PANGEA led to a comprehensive and prioritized list of candidate risk noncoding mutations in five major pediatric cancers.

Previous studies on noncoding mutations have been focused on SNVs and small indels. In contrast, systemic analysis of SVs has been lacking. Due to their much larger sizes, SVs may have a bigger impact on shaping the cis-regulatory landscape than SNVs (61,62). In support of this notion, we indeed found that SVs are the most frequent class of risk noncoding mutations when adjusted for background occurrence frequency (Supplementary Fig S5). In total, our analysis has revealed 1,137 risk SVs affecting the expression of over 2,000 genes across five pediatric cancer types.

Previous studies have also been focused on fusion genes generated by SVs since they are intuitive candidates of driver events. However, only ~30% of SVs generate fusion genes according to the current deposited SVs at ICGC. Moreover, ~35% of the gene fusions in the TARGET cohort are generated via microhomology-mediated end jointing (MMEJ). Since microhomology tends to be located at the break points of nonpathogenic SVs (63,64), these data suggest the fusion genes may not be oncogenic drivers in cases of MMEJ. Instead, in our analysis, we found that 55% SVs altered regulatory landscape and expression of nearby genes of the break points. These genes warrant careful investigation for a role in oncogenesis in future studies.

We found that coding and noncoding mutations affect distinct sets of genes and pathways. This mutual exclusivity is likely due to the different genomic locations of these two classes of genes. We found that genes affected by
noncoding mutations tend to be located in the region with early replication timing. This trend is consistent with previous reports that the occurrence of genomic rearrangements tends to occur in regions of early replication (65,66). Whether this correlation indicates a novel oncogenic mechanism needs to be further investigated. For instance, we found metabolic genes tend to be located in the early replicating region, and are more frequently affected by noncoding mutations. Rewiring of metabolism is a hallmark of cancer (67,68). Recent systematic analysis of TCGA data for 8 cancer types have reported that over 75% of metabolic genes are differentially expressed in each cancer type (69). However, it is unclear that to what degree the metabolism rewiring is mediated by noncoding mutations in those cancer types. Here our analysis suggests metabolic genes may be preferentially affected by noncoding mutation. In summary, our results highlight the need for comparative analysis of both coding and noncoding since novel cancer-related genes and pathways may be unveiled with comprehensive noncoding mutation analysis.

Many lineage-specific TFs are frequently perturbed in various cancers. However, the underlying oncogenic mechanism is challenging to understand. With comprehensive noncoding analysis, our approach provides a means to understand the detailed molecular mechanism underlying TF perturbation in cancers. For instance, recurrently de-regulated target genes of a TF suggest they can be important mediators of the disrupted TFs. Clinically, we found that the level of disruption of a transcription factor regulon can be used to stratify patient survival. More sophisticated computational models can be developed to prioritize target genes of perturbed TFs that contribute to oncogenesis.

Methods

Identification of single nucleotide variants (SNVs) and small indels

We used GATK (v3.8) and Freebayes (v1.0.2) (Table S9) to call SNVs and small indels. We first generated a set of initial SNV and indel calls with the default parameters of each software. Several filters were applied during the post-processing of the initial calls as suggested by the previous paper (70). First, we filtered mutations that overlap with low complexity regions. Second, we excluded regions with excessive read depth, as those regions are probably associated with spurious mappings. Third, we required the mutations to have multiple observations of the alternate (non-reference) allele in reads from both DNA strands. Last, we used the p-value cutoff of 0.01. SNVs passing these filters were intersected with annotations in dbSNP build 149. Calls matching both the position and allele of known dbSNP entries were removed (Figure S1A).

Identification of structural variants (SVs)
Delly (v0.7.2) and Lumpy (v0.2.13) (Table S9) were used to call structural variants. We used the default parameter settings of both software with the exception of setting minimum mapping quality threshold to zero as advised by the Complete Genomics data analysis pipeline (https://target-data.nci.nih.gov/Public/Resources/WGS/CGI/READMEs/). The initial SVs called by both software were retained for further filtering. First, we removed SVs in which break points were located in repetitive regions. Second, we removed SVs that are also identified in the baseline genomes which consist of 261 WGS samples from the 1000 Genome Project (www.internationalgenome.org/data). Finally, we selected SVs with at least 7 supporting reads as the final set of SVs (Figure S2A).

**RNA-Seq data analysis**

Raw reads from RNA-Seq were mapped to the reference human genome (release GRCh37) using STAR with default parameter setting. Transcripts were assembled using Cufflinks using mapped fragments outputted by STAR. Refseq (GRCh37) was used for the annotation of known transcripts. Normalized transcript abundance was computed using Cufflinks and expressed as FPKM (Fragments Per Kilobase of transcripts per Million mapped reads).

**Weighted elastic net model as a general framework for predicting risk mutations disrupting enhancer function**

For a given gene promoter, we consider all types of mutations that could potentially disrupt its regulation, including SNVs and small indels, copy number variations, inversions, and translocations. These mutations could either disrupt the function of the cis-regulatory sequences per se or disrupt the interactions between the enhancer and the promoter. For the latter category of mutations, SVs could hijack enhancers for a given promoter. We define potential enhancer hijacking as existing enhancer relocated to new region with a nearby promoter (<200k bp). For each promoter, its potential enhancers are detected based on patient SV data. We developed a regression-based approach to identify specific mutations that are associated with gene expression change. We used elastic net to implement the regression analysis. Elastic net combines the strength of ridge regression and least absolute shrinkage and selection operator (LASSO). It can enforce sparsity, has no limitation on the number of selected variables, and encourages grouping effect in the presence of highly correlated predictors.

For each gene, let us consider a regression model with \( n \) observations (i.e. \( n \) patients). Suppose that \( x_i = (x_{ij}, \ldots x_{nj})^T, \, j = 1, \ldots, p \) are the predictors (mutations disrupting promoter regulation) and \( y = (y_1, \ldots y_p)^T \) is the gene expression of the target promoters. \( X = [x_1, \ldots, x_n] \) denotes the predictor matrix. The total number of predictors are the number of disrupting mutations summed
over all promoters/enhancers regulating a given gene. The regression model can be expressed as $y = X\beta + \epsilon$ where $\beta = (\beta_1, \ldots, \beta_p)^T$ and the noise term $\epsilon \sim N(0, \sigma^2 I_n)$. A model fitting procedure produces the estimate of $\beta$, $\hat{\beta}$. The elastic net model is defined as:

$$\arg\min_{\beta} \frac{1}{2} \|y - \Sigma_{j=1}^p x_j \beta_j\|^2_2 + \lambda_1 \Sigma_{j=1}^p |\beta_j| + \lambda_2 \Sigma_{j=1}^p |\beta_j|^2$$

where $\lambda_1$ and $\lambda_2$ are tuning parameters that balance the goodness-of-fit and complexity of the model.

Because of the enhancer-promoter links are computationally predicted by the IM-PET algorithm, we address the uncertainty of the prediction by proposing a weighted elastic net approach. Specifically, we associate the probability score of an enhancer-promoter prediction (computed by IM-PET) to the coefficient of the predictor in the model, to enforce penalty on predictors caused by uncertainty in EP predictions. The weighted elastic net model is as following:

$$\arg\min_{\beta} \frac{1}{2} \|y - \Sigma_{j=1}^p x_j \beta_j\|^2_2 + \lambda_1 \Sigma_{j=1}^p w_j |\beta_j| + \lambda_2 \Sigma_{j=1}^p w_j |\beta_j|^2$$

where $w_i > 0, j = 1, \ldots, p$ are weight based on the EP prediction score computed by IM-PET.

CRISPR/Cas9-mediated translocation

sgRNAs targeting the break points were designed using CRISPOR (http://crispor.tefor.net) and cloned into the CRISPR vector pX459 (Addgene plasmid # 448139). CRISPR/Cas9-mediated translocation in Ba/F3 cells was performed as described in the previous study (71) with some modifications. Plasmids were co-transfected to Ba/F3 cells by electroporation. Dead cells were removed by centrifugation at 300x g and RT for 5 min 72 h post electroporation. Cell concentration and viability were measured using Countess II (Life Technologies). Live cells were re-suspended with Ba/F3 conditional medium to a concentration of 5 cells/mL. 100 μL of cell suspension was transferred to a 96-well plate and cultured for 2-3 weeks for selection of single cell clones. Genomic DNA was isolated using Quick-DNA 96 kit (ZYMO) and used for screening for clones with translocation by PCR (Table S7). Clones with translocation were further confirmed by Sanger sequencing.

Generation of cell lines with stably expressed Cas9 endonuclease

The B-cell acute lymphoblastic leukemia cell lines NALM-6 and REH were lentivirally transduced with plasmid expressing Cas9 nuclease (LentiCRISPR V2; Addgene Plasmid #52961) as described below at a MOI of 0.5. Cells then underwent 14 days of antibiotic selection with media containing puromycin 0.25 μg/mL to generate cell lines that stably express Cas9. Both cell lines were cultured in RPMI media supplemented with 10% FBS and 100 unit/mL.
penicillin/streptomycin. All cell lines were validated by ATCC STR profiling and confirmed to be Mycoplasma free.

**sgRNA design**

The sgRNAs targeting CHD4 were designed by Feng Zhang’s laboratory (72) (Table S8). The sequences for non-targeting control sgRNAs were chosen from previously described literature (73).

**Lentivirus production and transduction**

Lentivirus was produced by transfecting HEK293FT cells with packaging plasmids pMD2.G and psPAX2, and the individual CRISPR components (Cas9 expressing plasmid or sgRNA expressing plasmid: MCB306, Addgene Plasmid #89360 or MCB320, Addgene Plasmid #89359). Lentivirus-laden supernatant was harvested at 48 and 72 hrs after transfection, and viral supernatant was filtered through 0.45 μm polyvinylidene difluoride filter (Millipore) and concentrated using ultracentrifugation at 25,000 rpm for 2 hrs at 4 °C. Virus was then tittered via transduction of target cell line (REH or NALM-6).

Cells were lentivirally transduced by spinfection with centrifugation of cells with virus in the presence of 8μg/ml polybrene (Millipore) at 1,000xg for two hrs. Wildtype cells were transduced with virus containing plasmid expressing Cas9 at a MOI of 0.5. Cells stably expressing Cas9 were then transduced with plasmid containing sgRNA targeting CHD4 or non-targeting control at a MOI of 0.4. Three days after virus transduction, GFP-positive (transduced) cells were sorted by FACS, then seeded into 6-well plates for recovery before the following analyses.

**Data and Software availability**

Data and software used in this study are noted in the Method Details section above and Table S9. In this study, we developed a new software, PANGEA which is deposited at https://github.com/tanlabcode/PANGEA.

**Acknowledgments**

This work was supported by National Institutes of Health grants GM104369, GM108716, HG006130, CA226187, CA233285 (to KT), K08CA184418 (to SKT), and a pilot grant from the Children’s Hospital of Philadelphia Research Institute (to KT and SKT).

**Author contributions**

B.H. and K.T. designed the overall study, analyzed data, and wrote paper. P.G. performed most of the experiments. C.H.C and Y.D. performed competition
growth assays. C.C., Y.D, G.C., H.K, S.K.T, S.P.H helped with experimental design, data interpretation, and manuscript writing.
Figure 1. The PANGEA algorithm for simultaneous identification of candidate risk noncoding mutations of all classes.

A) Overview of the PANGEA algorithm. Given a catalog of enhancer-promoter (EP) interactions, a catalog of somatic mutations, and transcriptome data, the algorithm first constructs a count matrix of various classes of mutations disrupting the function of enhancer(s)/promoter(s) of each gene, including small mutations and structural variants. Weighted elastic net regression is then used to identify mutations that are significantly correlated with changes in target gene expression. B) Fraction of patients stratified by different classes of noncoding mutations affecting enhancers/promoters. C) Proportion of genes whose expression is disrupted by different classes of predicted risk noncoding mutations. SNV, single nucleotide variant; DEL, deletion; DUP, duplication; TRAN, translocation. D) Mutations predicted by PANGEA affect enhancers that have higher tissue specificity. Enhancer specificity is measured by the number of cell types in which the enhancer is observed to be active. P-value of one-sided t-test is shown (n=282,021); E) Mutations predicted by PANGEA affect higher percentage of super enhancers. p-value of hypergeometric test is shown (n=282,021); F) Mutations predicted by PANGEA affect enhancers with higher sequence conservation across 20 mammalian species. Enhancer conservation is measured by the average PhastCons score of the enhancer sequence. P-value of one-sided t-test is shown (n=282,021).
Figure 2. Enhancer hijacking to CHD4 by translocation in B-ALL patients.

A) Translocations result in enhancer hijacking to CHD4 in B-ALL. Three different translocation partners are identified in the B-ALL patient cohort. Shown tracks are histone modification ChIP-Seq data of human CD19+ B cells (identifying enhancers), SNVs and SV break points (BPs). For break point track, each vertical line represents a patient. The hijacked enhancers predicted to regulate CHD4 are highlighted in brown. B) Expression levels of CHD4 in patients with and without translocations. P-value of one-sided t-test is shown (n=163). C) B-ALL patients with CHD4 translocation have shorter time to relapse. P-value of log-rank test is shown (n=163). D) CHD4 knockout impairs the growth of B-ALL cell lines, NAML-6 and REH. CHD4 knockout was done using CRISPR-Cas9. E) Luciferase reporter assay of hijacked enhancer. BG, no DNA vector control; EV, vector containing no enhancer; NC, negative control, genomic region without enhancer histone mark H3K4me1 and H3K27ac; Enh, test EP300 enhancer. F) Western blots of CHD4 protein level in Ba/F3 cells with and without introduced translocation. P-value of one-sided t-test is shown (n=2). G) Relative mRNA levels of CHD4, NOP2, ZNF384, and EP300 in Ba/F3 cells with and without introduced translocations. P-values were calculated using one-sided t-test (n=2 for all tests). H) Murine Ba/F3 cells with induced translocation undergoes oncogenic transformation. Enhancer hijacking translocation was induced by CRISPR-Cas9. Error bars represent one standard deviation.
Figure 3. Examples of disruption of enhancer function by copy number alteration and point mutation.

A) Segmental duplications result in amplification of MYC enhancer in neuroblastoma. Shown tracks are histone modification ChIP-Seq data in human neurocrest cells and identified CNVs. Color bars in CNV track indicate MYCN amplification types. Orange, amplification of MYCN enhancers; Green, amplification of both MYCN enhancers and gene body; Blue, amplification of MYCN gene body only. B) Expression level of MYC in patients with and without copy number change. C) Kaplan-Meier plots of time to relapse for NBL patients with MYC gene amplification, MYC enhancer amplification, and without any copy number change in either gene or enhancer. P-value of one-sided log-rank test is shown (n=100). D) Point mutations result in disruption of GATA2 binding sites which regulate GFI1B expression in AML. Shown tracks are histone modification ChIP-Seq data in human neurocrest cells and identified SNVs. E) Expression level of GFI1B in patients with and without enhancer mutations. P-value of one-sided t-test is shown (n=153). F) Kaplan-Meier plots of time to relapse for AML patients with and without GFI1B enhancer point mutations P-value of one-sided log-rank test is shown (n=153).
Figure 4. Coding and noncoding mutations affect distinct sets of genes.

A) Venn diagrams of genes affected by recurrent coding and noncoding mutations in five cancer types. B-G) Features of genes affected by coding and noncoding mutations and genes without any mutation: gene exon length (B), exon conservation level measured by Phastcons score (C), enhancer degree (number of regulating enhancers) (D), gene expression specificity measured by fold change of average expression in a given cancer type compared to that of all five pediatric cancer types (E), enhancer conservation level measured by Phastcons score (F) and replication timing (G). H) Summary of different genomic features for the genes affected by coding and noncoding mutations. P-values of one-sided t-test are shown (n=21,841).
Figure 5. Degree of regulon disruption of key TFs is correlated with clinical outcome.

A) Top 10 most frequently disrupted regulons in each cancer type. Bar plots show the percentage of patients with regulon disruption of the listed TFs; TFs with known role in the given cancer are highlighted in red. B) Number of patients affected by regulon disruption (either TF coding mutations or target gene noncoding mutations). Asterisks indicate TFs whose regulon disruptions are correlated with patient time to relapse (log-rank test p < 0.05, n=501). C) Kaplan-Meier plots of time to relapse for B-ALL patients with PAX5 deletion, EP disruption involving PAX5, and without any mutation of the PAX5 regulon; D) Kaplan-Meier plots of time to relapse for AML patients with RUNX1 fusion, EP disruption involving RUNX1, and without any mutation of the RUNX1 regulon. E) Kaplan-Meier plots of time to relapse for WT patients with WT1 mutation, EP disruption involving WT1, and without any mutation of the WT1 regulon. F) Kaplan-Meier plots of time to relapse for B-ALL patients with TCF3 fusion, EP disruption involving TCF3, and without any mutation of the TCF3 regulon; G) Kaplan-Meier plots of time to relapse for NBL patients with MYC amplification, EP disruption involving MYC, and without any mutation of the MYC regulon. P-values of one-sided log-rank test are shown. The numbers of samples are indicated in brackets.
Figure 6. Integrative mutation signatures of pediatric cancers.

A)- E) Mutation signatures of B-ALL, AML, NBL, WT, and OS patients. Heatmaps were generated using recurrently mutated genes (due to either coding or noncoding mutations) in at least 5 patients. Patients are clustered according to the combined coding and noncoding mutation profiles. The color of each cell indicates the class of mutations. Rows, genes. Columns, patients. For presentation purpose, the heatmaps are divided into submaps based on noncoding (top) or coding mutations (bottom). The tables below the heatmaps list up- or down-regulated genes due to noncoding mutations in all patients in a given cluster.
References

1. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature 2018;555(7696):371-6 doi 10.1038/nature25795.

2. Grobner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature 2018;555(7696):321-7 doi 10.1038/nature25480.

3. Alexander TB, Gu Z, Iacobucci I, Dickerson K, Choi JK, Xu B, et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. Nature 2018;562(7727):373-9 doi 10.1038/s41586-018-0436-0.

4. Bolouri H, Farrar JE, Triche T, Jr., Ries RE, Lim EL, Alonzo TA, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat Med 2018;24(1):103-12 doi 10.1038/nm.4439.

5. Mullighan CG. Genomic characterization of childhood acute lymphoblastic leukemia. Seminars in hematology 2013;50(4):314-24 doi 10.1053/j.seminhematol.2013.10.001.

6. Oldridge DA, Wood AC, Weichert-Leahey N, Crimmins I, Sussman R, Winter C, et al. Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. Nature 2015;528(7582):418-21 doi 10.1038/nature15540.

7. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science 2014;346(6215):1373-7 doi 10.1126/science.1259037.

8. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature 2010;464(7291):993-8 doi 10.1038/nature08987.

9. Northcott PA, Lee C, Zichner T, Stutz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GF1 family oncogenes in medulloblastoma. Nature 2014;511(7510):428-34 doi 10.1038/nature13379.

10. Ryan RJ, Drier Y, Whitton H, Cotton MJ, Kaur J, Issner R, et al. Detection of Enhancer-Associated Rearrangements Reveals Mechanisms of Oncogene Dysregulation in B-cell Lymphoma. Cancer Discov 2015;5(10):1058-71 doi 10.1158/2159-8290.CD-15-0370.

11. Tasian SK, Loh ML. Understanding the biology of CRLF2-overexpressing acute lymphoblastic leukemia. Critical reviews in oncogenesis 2011;16(1-2):13-24.

12. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature biotechnology 2014;32(3):246-51 doi 10.1038/nbt.2835.
13. Perry JA, Kiezun A, Tonzi P, Van Allen EM, Carter SL, Baca SC, et al. Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma. Proceedings of the National Academy of Sciences of the United States of America 2014;111(51):E5564-73 doi 10.1073/pnas.1419260111.

14. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012;28(18):i333-i9 doi 10.1093/bioinformatics/bts378.

15. Roberts KG, Li Y, Payne-Turner D, Harvey RC, Yang YL, Pei D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N Engl J Med 2014;371(11):1005-15 doi 10.1056/NEJMoa1403088.

16. Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 2010;26(13):1579-86 doi 10.1093/bioinformatics/btq248.

17. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proceedings of the National Academy of Sciences of the United States of America 2014;111(21):E2191-9 doi 10.1073/pnas.1320308111.

18. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research 2016;44(D1):D164-71 doi 10.1093/nar/gkv1002.

19. Drier Y, Cotton MJ, Williamson KE, Gillespie SM, Ryan RJ, Kluk MJ, et al. An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. Nat Genet 2016;48(3):265-72 doi 10.1038/ng.3502.

20. Valentijn LJ, Koster J, Zwijnenburg DA, Hasselt NE, van Sluis P, Volckmann R, et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. Nat Genet 2015;47(12):1411-4 doi 10.1038/ng.3438.

21. Peifer M, Hertwig F, Roels F, Driedax D, Gartlgruber M, Menon R, et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. Nature 2015;526(7575):700-4 doi 10.1038/nature14980.

22. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell 2014;26(3):319-30 doi 10.1016/j.ccr.2014.07.014.

23. Qian M, Zhang H, Kham SK, Liu S, Jiang C, Zhao X, et al. Whole-transcriptome sequencing identifies a distinct subtype of acute lymphoblastic leukemia with predominant genomic abnormalities of EP300 and CREBBP. Genome Res 2017;27(2):185-95 doi 10.1101/gr.209163.116.

24. Dege C, Hagman J. Mi-2/NuRD chromatin remodeling complexes regulate B and T-lymphocyte development and function. Immunol Rev 2014;261(1):126-40 doi 10.1111/imr.12209.

25. Cai Y, Geutjes EJ, de Lint K, Roopman P, Bruurs L, Yu LR, et al. The NuRD complex cooperates with DNMTs to maintain silencing of key
26. Sperlazza J, Rahmani M, Beckta J, Aust M, Hawkins E, Wang SZ, et al. Depletion of the chromatin remodeler CHD4 sensitizes AML blasts to genotoxic agents and reduces tumor formation. Blood 2015;126(12):1462-72 doi 10.1182/blood-2015-03-631606.

27. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res 2012;22(9):1658-67 doi 10.1101/gr.136838.111.

28. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The genetic landscape of high-risk neuroblastoma. Nat Genet 2013;45(3):279-84 doi 10.1038/ng.2529.

29. Zhuang L, Ma Y, Wang Q, Zhang J, Zhu C, Zhang L, et al. Atg3 Overexpression Enhances Bortezomib-Induced Cell Death in SKM-1 Cell. PloS one 2016;11(7):e0158761 doi 10.1371/journal.pone.0158761.

30. Moignard V, Macaulay IC, Swiers G, Buettner F, Schutte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nature cell biology 2013;15(4):363-72 doi 10.1038/ncb2709.

31. Laurent B, Randrianarison-Huetz V, Marechal V, Mayeux P, Dusanter-Fourt I, Dumenil D. High-mobility group protein HMGB2 regulates human erythroid differentiation through trans-activation of GFI1B transcription. Blood 2010;115(3):687-95 doi 10.1182/blood-2009-06-230094.

32. Vassen L, Beauchemin H, Lemsaddek W, Krongold J, Trudel M, Moroy T. Growth factor independence 1b (gfi1b) is important for the maturation of erythroid cells and the regulation of embryonic globin expression. PloS one 2014;9(5):e96636 doi 10.1371/journal.pone.0096636.

33. Thivakaran A, Botezatu L, Hones JM, Schutte J, Vassen L, Al-Matary YS, et al. Gfi1b: a key player in the genesis and maintenance of acute myeloid leukemia and myelodysplastic syndrome. Haematologica 2018;103(4):614-25 doi 10.3324/haematol.2017.167288.

34. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell 2013;155(4):934-47 doi 10.1016/j.cell.2013.09.053.

35. Damm F, Thol F, Hollink I, Zimmermann M, Reinhardt K, van den Heuvel-Eibrink MM, et al. Prevalence and prognostic value of IDH1 and IDH2 mutations in childhood AML: a study of the AML-BFM and DCOG study groups. Leukemia 2011;25(11):1704-10 doi 10.1038/leu.2011.142.

36. Green CL, Evans CM, Zhao L, Hills RK, Burnett AK, Linch DC, et al. The prognostic significance of IDH2 mutations in AML depends on the location of the mutation. Blood 2011;118(2):409-12 doi 10.1182/blood-2010-12-322479.

37. Patel JL, Schumacher JA, Frizzell K, Sorrells S, Shen W, Clayton A, et al. Coexisting and cooperating mutations in NPM1-mutated acute myeloid
leukemia. Leukemia research 2017;56:7-12 doi 10.1016/j.leukres.2017.01.027.

38. Groschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, Erpelinck C, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell 2014;157(2):369-81 doi 10.1016/j.cell.2014.02.019.

39. Sima J, Gilbert DM. Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. Current opinion in genetics & development 2014;25:93-100 doi 10.1016/j.gde.2013.11.022.

40. van der Veer A, Waanders E, Pieters R, Willemse ME, Van Reijmersdal SV, Russell LJ, et al. Independent prognostic value of BCR-ABL1-like signature and IKZF1 deletion, but not high CRLF2 expression, in children with B-cell precursor ALL. Blood 2013;122(15):2622-9 doi 10.1182/blood-2012-10-62358.

41. Nebral K, Denk D, Attarbaschi A, Konig M, Mann G, Haas OA, et al. Incidence and diversity of PAX5 fusion genes in childhood acute lymphoblastic leukemia. Leukemia 2009;23(1):134-43 doi 10.1038/leu.2008.306.

42. Schwab C, Ryan SL, Chilton L, Elliott A, Murray J, Richardson S, et al. EBF1-PDGFRB fusion in pediatric B-cell precursor acute lymphoblastic leukemia (BCP-ALL): genetic profile and clinical implications. Blood 2016;127(18):2214-8 doi 10.1182/blood-2015-09-670166.

43. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LA, Miller CB, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med 2009;360(5):470-80 doi 10.1056/NEJMoa0808253.

44. Pieraccioli M, Nicolai S, Pitolli C, Agostini M, Antonov A, Malewicz M, et al. ZNF281 inhibits neuronal differentiation and is a prognostic marker for neuroblastoma. Proceedings of the National Academy of Sciences of the United States of America 2018;115(28):7356-61 doi 10.1073/pnas.1801435115.

45. Dzieran J, Rodriguez Garcia A, Westermark UK, Henley AB, Eyre Sanchez E, Trager C, et al. MYCN-amplified neuroblastoma maintains an aggressive and undifferentiated phenotype by deregulation of estrogen and NGF signaling. Proceedings of the National Academy of Sciences of the United States of America 2018;115(6):E1229-E38 doi 10.1073/pnas.1710901115.

46. Faber ZJ, Chen X, Gedman AL, Boggs K, Cheng J, Ma J, et al. The genomic landscape of core-binding factor acute myeloid leukemias. Nat Genet 2016;48(12):1551-6 doi 10.1038/ng.3709.

47. Gocho Y, Kiyokawa N, Ichikawa H, Nakabayashi K, Osumi T, Ishibashi T, et al. A novel recurrent EP300-ZNF384 gene fusion in B-cell precursor acute lymphoblastic leukemia. Leukemia 2015;29(12):2445-8 doi 10.1038/leu.2015.111.
48. Makar AB, McMartin KE, Palese M, Tephly TR. Formate assay in body fluids: application in methanol poisoning. Biochemical medicine 1975;13(2):117-26.

49. Holmlund T, Lindberg MJ, Grander D, Wallberg AE. GCN5 acetylates and regulates the stability of the oncoprotein E2A-PBX1 in acute lymphoblastic leukemia. Leukemia 2013;27(3):578-85 doi 10.1038/leu.2012.265.

50. De Cecco L, Capaia M, Zupi S, Cutrona G, Matis S, Brizzolara A, et al. Interleukin 21 Controls mRNA and MicroRNA Expression in CD40-Activated Chronic Lymphocytic Leukemia Cells. PloS one 2015;10(8):e0134706 doi 10.1371/journal.pone.0134706.

51. Rodriguez-Rodriguez S, Pomerantz A, Demichelis-Gomez R, Barrera-Lumbreras G, Barrales-Benitez O, Aguayo-Gonzalez A. Expression of HER2/Neu in B-Cell Acute Lymphoblastic Leukemia. Revista de investigacion clinica; organo del Hospital de Enfermedades de la Nutricion 2016;68(4):171-5.

52. Shao X, Liu Y, Li Y, Xian M, Zhou Q, Yang B, et al. The HER2 inhibitor TAK165 Sensitizes Human Acute Myeloid Leukemia Cells to Retinoic Acid-Induced Myeloid Differentiation by activating MEK/ERK mediated RARalpha/STAT1 axis. Scientific reports 2016;6:24589 doi 10.1038/srep24589.

53. Szelachowska J, Jelen M, Kornafel J. Prognostic significance of intracellular laminin and Her2/neu overexpression in non-small cell lung cancer. Anticancer research 2006;26(5B):3871-6.

54. Izycka-Swieszewska E, Wozniak A, Kot J, Grajkowska W, Balcerska A, Perek D, et al. Prognostic significance of HER2 expression in neuroblastoma tumors. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2010;23(9):1261-8 doi 10.1038/modpathol.2010.115.

55. Thomas H, Beck K, Adamczyk M, Aeschlimmann P, Langley M, Oita RC, et al. Transglutaminase 6: a protein associated with central nervous system development and motor function. Amino acids 2013;44(1):161-77 doi 10.1007/s00726-011-1091-z.

56. Algarni AS, Hargreaves AJ, Dickenson JM. Activation of transglutaminase 2 by nerve growth factor in differentiating neuroblastoma cells: A role in cell survival and neurite outgrowth. European journal of pharmacology 2018;820:113-29 doi 10.1016/j.ejphar.2017.12.023.

57. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. Nature communications 2014;5:4846 doi 10.1038/ncomms5846.

58. Wang L, Sun Y, Sun Y, Meng L, Xu X. First case of AML with rare chromosome translocations: a case report of twins. BMC cancer 2018;18(1):458 doi 10.1186/s12885-018-4396-4.

59. Shiozawa Y, Pedersen EA, Patel LR, Ziegler AM, Havens AM, Jung Y, et al. GAS6/AXL axis regulates prostate cancer invasion, proliferation, and survival in the bone marrow niche. Neoplasia 2010;12(2):116-27.
60. Wu G, Ma Z, Hu W, Wang D, Gong B, Fan C, et al. Molecular insights of Gas6/TAM in cancer development and therapy. Cell death & disease 2017;8(3):e2700 doi 10.1038/cddis.2017.113.

61. Onishi-Seebacher M, Korbel JO. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. BioEssays : news and reviews in molecular, cellular and developmental biology 2011;33(11):840-50 doi 10.1002/bies.201100075.

62. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nature reviews Genetics 2013;14(2):125-38 doi 10.1038/nrg3373.

63. Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet 2012;44(4):390-7, S1 doi 10.1038/ng.2202.

64. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseleu PS, Hsieh CH, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 2013;153(4):919-29 doi 10.1016/j.cell.2013.04.010.

65. Janoueix-Lerosey I, Hupe P, Maciorowski Z, La Rosa P, Schleiermacher G, Pierron G, et al. Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma. Cell Cycle 2005;4(12):1842-6 doi 10.4161/cc.4.12.2257.

66. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nature communications 2013;4:1502 doi 10.1038/ncomms2502.

67. Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab 2016;23(1):27-47 doi 10.1016/j.cmet.2015.12.006.

68. Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. Cell 2008;134(5):703-7 doi 10.1016/j.cell.2008.08.021.

69. Kim P, Cheng F, Zhao J, Zhao Z. ccmGDB: a database for cancer cell metabolism genes. Nucleic acids research 2016;44(D1):D959-68 doi 10.1093/nar/gkv1128.

70. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 2014;30(20):2843-51 doi 10.1093/bioinformatics/btu356.

71. Torres R, Martin MC, Garcia A, Cigudosa JC, Ramirez JC, Rodriguez-Perales S. Engineering human tumour-associated chromosomal translocations with the RNA-guided CRISPR-Cas9 system. Nature communications 2014;5:3964 doi 10.1038/ncomms4964.

72. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nature methods 2014;11(8):783-4 doi 10.1038/nmeth.3047.

73. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-
target effects of CRISPR-Cas9. Nature biotechnology 2016;34(2):184-91
doi 10.1038/nbt.3437.