Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children

Xu, Heng; Zhang, Hui; Yang, Wenjian; Yadav, Rachita; Morrison, Alanna C.; Qian, Maoxiang; Devidas, Meenakshi; Liu, Yu; Perez-Andreu, Virginia; Zhao, Xujie; Hansen, Torben; Holm, Jens; Gupta, Ramneek

Published in:
Nature Communications

Link to article, DOI:
10.1038/ncomms8553

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Xu, H., Zhang, H., Yang, W., Yadav, R., Morrison, A. C., Qian, M., ... Gupta, R. (2015). Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. Nature Communications, 6, [7553]. DOI: 10.1038/ncomms8553
Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children

Heng Xu1,2,*, Hui Zhang1,3,*, Wenjian Yang1, Rachita Yadav4, Alanna C. Morrison5, Maoxiang Qian1, Meenakshi Devidas6, Yu Liu7, Virginia Perez-Andreu1, Xujie Zhao1, Julie M. Gastier-Foster8, Philip J. Lupo9, Geoff Neale10, Elizabeth Raetz11, Eric Larsen12, W. Paul Bowman13, William L. Carroll14, Naomi Winick15, Richard Williams16, Torben Hansen17, Jens-Christian Holm18, Elaine Mardis19, Robert Fulton19, Ching-Hon Pu20,21, Jinghui Zhang7, Charles G. Mullighan20,22, William E. Evans1,20, Stephen P. Hunger23, Ramneek Gupta4, Kjeld Schmiegelow24, Mignon L. Loh25, Mary V. Relling1,20 & Jun J. Yang1,20

There is increasing evidence from genome-wide association studies for a strong inherited genetic basis of susceptibility to acute lymphoblastic leukaemia (ALL) in children, yet the effects of protein-coding variants on ALL risk have not been systematically evaluated. Here we show a missense variant in CDKN2A associated with the development of ALL at genome-wide significance (rs3731249, \( P = 9.4 \times 10^{-23} \), odds ratio = 2.23). Functional studies indicate that this hypomorphic variant results in reduced tumour suppressor function of p16INK4A, increases the susceptibility to leukaemic transformation of haematopoietic progenitor cells, and is preferentially retained in ALL tumour cells. Resequencing the CDKN2A–CDKN2B locus in 2,407 childhood ALL cases reveals 19 additional putative functional germline variants. These results provide direct functional evidence for the influence of inherited genetic variation on ALL risk, highlighting the important and complex roles of CDKN2A–CDKN2B tumour suppressors in leukaemogenesis.
The risk of developing acute lymphoblastic leukaemia (ALL) is highest between 2 and 5 years after birth, with initiating sentinel somatic genomic lesions (for example, chromosomal translocations) detectable at the time of birth in many cases. This early disease onset suggests a strong inherited genetic basis for ALL susceptibility, and recent genome-wide association studies (GWAS) have discovered at least six risk loci: ARID5B, IKZF1, CEBPE, PIP4K2A–BM11, GATA3 and CDKN2A–CDKN2B. These ALL risk genes are directly involved in haematopoietic stem cell function, lymphocyte differentiation and development, and cell cycle regulation, several of which are also commonly targeted by somatic genomic lesions. In particular, the CDKN2A–CDKN2B locus is one of the most frequently deleted genomic regions in childhood ALL with focal copy number loss in both B- and T-cell ALL.

The vast majority of variants examined in previous ALL GWAS are intronic or intergenic. Although it is now evident that non-coding variants related to disease traits are significantly over-represented in regulatory DNA and often function as modulators of local or distal gene transcription, questions also arise whether coding variants within ALL susceptibility genes might confer even greater effects on disease development. Moreover, a large number of low-frequency and rare-coding germline variants have been discovered by exome-sequencing efforts, but their contributions to ALL pathogenesis have yet to be examined systematically.

In the present study, we perform an exome-focused GWAS to systematically examine the impact of germline-coding variants on the development of ALL in children of European descent, experimentally explore the functional consequences of the genome-wide significant variant in the CDKN2A gene, and comprehensively characterize coding variation at this locus by targeted resequencing.

**Results**

**Exome-focused GWAS of ALL susceptibility.** In the discovery GWAS, we genotyped 1,773 children with B-ALL and 10,448 non-ALL controls of European descent for 247,505 variants using the Illumina Infinium HumanExome array. Three loci with genome-wide significant association signals were observed: ARID5B (10q21.2), IKZF1 (7q12.2) and CDKN2A (9p21.3) (Fig. 1). Non-coding variants rs10821936 in ARID5B and rs4132601 in IKZF1 showed the strongest association (P = 9.9 × 10^{-40} and 4.3 × 10^{-37}), the logistic regression test, respectively; Fig. 1 and Supplementary Table 1), confirming previous GWAS findings from our group and others. Whether coding variants within ALL susceptibility genes might be strongly associated with susceptibility to ALL in populations of European descent. Genotype correlation between the coding variant rs3731249 and the intronic rs3731217 is exceedingly low (P < 0.01 in Europeans, Supplementary Fig. 1), and multivariate analyses including both SNPs indicated their independent contributions to ALL risk (Supplementary Table 2). In the replication cohort of 409 childhood ALL cases and 1,599 non-ALL controls of European descent, the association signal at rs3731249 was validated (P = 5.2 × 10^{-4}, odds ratio = 1.73 (1.27–2.36), the logistic regression test, Table 1) and this variant also remained significant after adjusting for rs3731217.

**Functional characterization of the rs3731249 variant.** To experimentally evaluate the effects of rs3731249 on ALL leukaemogenesis, we directly compared the effect of wildtype versus variant allele p16^{INK4A} (p.148T variant) on BCR–ABL1-mediated leukaemic transformation in vitro. We chose a pair of haematopoietic progenitor Ba/F3 cell line because it is inherently p16^{INK4A}-defective due to methylation at the Ink4a-Arf locus, and ectopic expression of BCR–ABL1 in Ba/F3 cells efficiently induces exogenous cytokine (interleukin 3 (IL3))-independent proliferation. Over-expression of wild-type p16^{INK4A} significantly inhibited leukaemic transformation by BCR–ABL1 (Fig. 2a, Supplementary Fig. 2), consistent with its role as a critical tumour suppressor in ALL. In contrast, Ba/F3 cells overexpressing variant p16^{INK4A} were significantly more susceptible to BCR–ABL1 transformation measured by IL3-independent growth, suggesting that the p.148T variant is likely hypomorphic with reduced tumour suppressor function. In Ba/F3 cells transfected with both variant and wild-type p16^{INK4A}, the relative ratio of the p.148T (variant) to p.148A ratio was highest between 2 and 5 years after birth, with initiating sentinel somatic genomic lesions (for example, chromosomal translocations) detectable at the time of birth in many cases. This early disease onset suggests a strong inherited genetic basis for ALL susceptibility, and recent genome-wide association studies (GWAS) have discovered at least six risk loci: ARID5B, IKZF1, CEBPE, PIP4K2A–BM11, GATA3 and CDKN2A–CDKN2B. These ALL risk genes are directly involved in haematopoietic stem cell function, lymphocyte differentiation and development, and cell cycle regulation, several of which are also commonly targeted by somatic genomic lesions. In particular, the CDKN2A–CDKN2B locus is one of the most frequently deleted genomic regions in childhood ALL with focal copy number loss in both B- and T-cell ALL.

In the present study, we perform an exome-focused GWAS to systematically examine the impact of germline-coding variants on the development of ALL in children of European descent, experimentally explore the functional consequences of the genome-wide significant variant in the CDKN2A gene, and comprehensively characterize coding variation at this locus by targeted resequencing.

**Results**

**Exome-focused GWAS of ALL susceptibility.** In the discovery GWAS, we genotyped 1,773 children with B-ALL and 10,448 non-ALL controls of European descent for 247,505 variants using the Illumina Infinium HumanExome array. Three loci with genome-wide significant association signals were observed: ARID5B (10q21.2), IKZF1 (7q12.2) and CDKN2A (9p21.3) (Fig. 1). Non-coding variants rs10821936 in ARID5B and rs4132601 in IKZF1 showed the strongest association (P = 9.9 × 10^{-40} and 4.3 × 10^{-37}), the logistic regression test, respectively; Fig. 1 and Supplementary Table 1), confirming previous GWAS findings from our group and others. Whether coding variants within ALL susceptibility genes might be strongly associated with susceptibility to ALL in populations of European descent. Genotype correlation between the coding variant rs3731249 and the intronic rs3731217 is exceedingly low (P < 0.01 in Europeans, Supplementary Fig. 1), and multivariate analyses including both SNPs indicated their independent contributions to ALL risk (Supplementary Table 2). In the replication cohort of 409 childhood ALL cases and 1,599 non-ALL controls of European descent, the association signal at rs3731249 was validated (P = 5.2 × 10^{-4}, odds ratio = 1.73 (1.27–2.36), the logistic regression test, Table 1) and this variant also remained significant after adjusting for rs3731217.

**Functional characterization of the rs3731249 variant.** To experimentally evaluate the effects of rs3731249 on ALL leukaemogenesis, we directly compared the effect of wildtype versus variant allele p16^{INK4A} (p.148A versus p.148T) on BCR–ABL1-mediated leukaemic transformation in vitro. We chose a pair of haematopoietic progenitor Ba/F3 cell line because it is inherently p16^{INK4A}-defective due to methylation at the Ink4a-Arf locus, and ectopic expression of BCR–ABL1 in Ba/F3 cells efficiently induces exogenous cytokine (interleukin 3 (IL3))-independent proliferation. Over-expression of wild-type p16^{INK4A} significantly inhibited leukaemic transformation by BCR–ABL1 (Fig. 2a, Supplementary Fig. 2), consistent with its role as a critical tumour suppressor in ALL. In contrast, Ba/F3 cells overexpressing variant p16^{INK4A} were significantly more susceptible to BCR–ABL1 transformation measured by IL3-independent growth, suggesting that the p.148T variant is likely hypomorphic with reduced tumour suppressor function. In Ba/F3 cells transfected with both variant and wild-type p16^{INK4A}, the relative ratio of the p.148T (variant) to p.148A
gene (p16INK4A and p14ARF) in children with ALL a higher burden of rare missense variants in relative to controls. Comparing with 4,300 European American individuals from the ESP cohort, we resequenced the coding region of the CDKN2A and CDKN2B genes in germline DNA from 2,407 childhood ALL cases (1,450 of which were also included in the discovery GWAS). In addition to rs3731249, we observed another 13 germline exonic variants in tumour suppressors p16INK4A and p14ARF encoded by the CDKN2A gene, 12 of which result in amino-acid sequence changes (Fig. 3). These missense variants were all singletons, except for the p.D125H variant in p16INK4A and the p.A121T variant in p14ARF observed in two and five cases, respectively. Five variants were predicted to be damaging based on combined annotation dependent depletion23 (CADD score >13, Supplementary Table 3), and we did not observe germline insertions or deletions in CDKN2A in our ALL cohort. Comparing with 4,300 European American individuals from the NHLBI GO Exome Sequencing Project (ESP), there was a trend for a higher burden of rare missense variants in relative to controls the CDKN2A gene (p16INK4A and p14ARF) in children with ALL (0.71% versus 0.23%, P=0.0045; Fisher’s exact test, Fig. 3). In addition, we identified six germline-coding variants in the adjacent CDKN2B gene in this cohort of children with ALL, although there was no significant over-representation compared with European controls in the ESP cohort (0.83% versus 0.79%, Fig. 3).

Figure 2 | Functional characterization of ALL risk variant at rs3731249. (a) Mouse haematopoietic progenitor cell Ba/f3 overexpressing wildtype, variant p16INK4A, or transfected with control vector was transduced with leukaemia oncogenic BCR–ABL1 fusion gene. Cell proliferation in the absence of cytokine IL3 was measured daily as an indicator of leukaemic transformation. Ectopic expression of p16INK4A (p.148A, green) significantly potentiated leukaemic transformation by BCR–ABL1, compared with cells expressing wild-type p16INK4A (p.148A, blue), consistent with the association of this allele with susceptibility to ALL. Data represent the mean of three replicates ± s.e.m. (b) Allele-specific expression of p16INK4A in ALL blasts was determined by comparing the number of sequence reads for transcripts containing C or T alleles at rs3731249 (p16INK4A[p.148A] versus p16INK4A[p.148T]), in 15 childhood ALL cases with heterogeneous genotype in the germline DNA at this SNP. Each dot represents an ALL case (red indicates cases with somatic deletion (loss of heterozygosity) and blue indicates cases without copy number change in tumour) and the line of identity indicates equal expression of both alleles. P-value was estimated by paired t-test based on the number of sequence reads for each allele.

Table 1 | Genome-wide significant association and replication of novel coding B-ALL susceptibility variant at CDKN2A locus.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Chr</th>
<th>Position</th>
<th>Gene</th>
<th>Alleles</th>
<th>Genotype</th>
<th>GWAS series</th>
<th>Replication series</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3731249</td>
<td>9</td>
<td>21970916</td>
<td>CDKN2A</td>
<td>C/T</td>
<td>Case (N=1,773)</td>
<td>Control (N=10,441)</td>
<td>P-value</td>
</tr>
<tr>
<td>TT (0.45%)</td>
<td>8</td>
<td>224 (12.63%)</td>
<td>619 (5.93%)</td>
<td>1.541 (86.92%)</td>
<td>9.4 × 10^-23</td>
<td>2.23 (1.90-2.61)</td>
<td>7 (1.71%)</td>
</tr>
</tbody>
</table>

Chr, chromosome; CI, confidence interval; GWAS, genome-wide association studies; OR, odds ratio; SNP, single-nucleotide polymorphism.

*Chrosomal localisations are based on hg19

†Bold denotes the allele that had a significantly higher frequency in children with B-ALL than in the non-ALL controls (that is, risk allele for B-ALL)

‡OR represents the increase in the risk of developing ALL for each copy of the risk allele compared with subjects who do not carry the risk allele; P-values and ORs were estimated by the logistic regression test.

Discussion

Encoding three tumour suppressor proteins (p16INK4A, p14ARF and p15INK4B), the CDKN2A–CDKN2B locus at 9p21 is promiscuously associated with tumorigenesis and commonly mediated transformation (Supplementary Fig. 3), consistent with the increased leukaemia risk conferred by the variant allele at rs3731249. To further examine the potential susceptibility to ALL conferred by the rs3731249 in patients, we compared the genotype distribution in RNA and DNA from primary leukaemic blasts and matched germline samples from children with ALL (Fig. 2b). Of 15 cases with the heterozygous germline genotype at this SNP, six exhibited somatic deletion of one copy of CDKN2A, all of which retained the risk allele in tumour cells. Even in cases not affected by somatic copy number loss at this locus, the variant p16INK4A(c.442T) was preferentially transcribed relative to wildtype (c.442C), with allele-biased expression ranging from 61 to 100%, Fig. 2b). Altogether, these results pointed to the possibility that cells carrying the hypomorphic risk allele at rs3731249 might have been enriched during leukaemogenesis.
targeted by somatic mutation, deletion and/or hypermethylation in various cancers. p16INK4A and p15INK4B are highly homologous inhibitors of cyclin-dependent kinase and function mainly as master regulators of cell cycle entry via the Rb-E2F signalling axis. Although also encoded by the CDKN2A gene, p14ARF utilizes a completely different reading frame with distinct tumour suppression functions by inhibiting MDM2 and activating p53. Suppressed during normal haematopoiesis, p16INK4A and p14ARF expression is activated on oncogenic stimuli (for example, constitutive expression of BCR-ABL1 fusion) to trigger cell cycle exit (senescence) or apoptosis as a means of eliminating oncogene-stressed cells. In fact, the p14ARF, therefore, likely to have a more direct effect on the former. This hypothesis is supported by the fact that haematopoietic progenitor cells (Ba/f3) expressing variant p16INK4A were substantially more susceptible to BCR-ABL1-mediated leukaemic transformation compared with cells with the wild-type protein (Fig. 2a), pointing to rs3731249 as a possible functional variant directly contributing to the association with ALL risk. The structural basis of the hypomorphic effects of the p.A148T variant is unclear, since this residue is not directly involved in binding to CDK4 or CDK6. However, there was evidence that the variant p16INK4A (p.148T) is preferentially retained in the nucleus compared with the wild-type p16INK4A (p.148A), compromising its ability to inhibit CDKs in the cytoplasm. The relative contribution of p16INK4A versus p14ARF to ALL pathogenesis is not unequivocal because somatic deletions at this locus almost always lead to the loss of both genes. Although the rs3731249 variant also results in sequence changes of the 3′-UTR of the p14ARF transcript, bioinformatic prediction did not identify any potential effects on mRNA stability or microRNA binding and no difference was observed in reporter gene transcription under the influence of 3′-UTR containing either the wildtype or variant allele at rs3731249 (Supplementary Fig. 4), suggesting minimal effects of this variant on p14ARF transcription. Finally, rs3731249 is also observed in non-European populations, for example, there was a trend for a higher frequency of the risk allele in African American children with ALL than that in individuals from this racial background in the NHLBI ESP cohort (0.58% in 260 ALL cases versus 0.38% in 2,203 controls), although a much larger sample size is needed to rigorously examine the statistical significance of such differences. It should be noted that we and others previously showed that the
non-coding ALL risk variants (rs17756311 and rs3731217) at this locus had much stronger effects in European Americans than in other race/ethnic groups\(^{29,30}\), suggesting potential racial differences in genetic susceptibility to ALL.

We subsequently identified additional coding variants in p16\(^{INK4A}\), p14\(^{ARF}\), and p15\(^{INK4B}\) by sequencing, most of which were low frequency or rare. While there was a modest over-representation of potentially damaging coding variants in ALL cases compared with controls (Fig. 3), our data do not suggest that rare variants contribute substantially to the associations with ALL susceptibility observed at this locus. It should also be noted that the vast majority of coding variants within the CDKN2A gene affects only one of the two tumour suppressors (either p16\(^{INK4A}\) or p14\(^{ARF}\)). Interestingly, rs199888003 is the only variant that is located in the coding region of both p16\(^{INK4A}\) and p14\(^{ARF}\), resulting in an alanine-to-threonine change in p14\(^{ARF}\) (p.A121T) with synonymous effect on p16\(^{INK4A}\). This is also the most frequent germline missense variant in p14\(^{ARF}\) in our cohort and was over-represented in ALL compared with non-ALL controls (0.21% versus 0.046%, respectively, Fig. 3). This substitution of threonine in p14\(^{ARF}\) adds a possible phosphorylation and phosphorylation site and also introduces a phosphoprotein-binding FHA domain implicated in DNA damage response and cell cycling\(^{31}\). Future studies are warranted to determine the exact consequences of this variant on p14\(^{ARF}\) functions. To systemically evaluate the contribution of low frequency and rare-coding variants to ALL risk, we also performed genome-wide gene-level burden test but did not observe any genome-wide significant associations (Supplementary Table 4). Of the six known ALL risk loci, we noted two coding variants in CEBPE (rs141903485 and rs146580935, Supplementary Table 5) nominally associated with ALL susceptibility.

In conclusion, we comprehensively evaluated exonic genetic variations for association with ALL susceptibility and identified novel coding risk variants at the CDKN2A–CDKN2B locus that may directly affect tumour suppressor functions and potentiate leukemic transformation. These results provided functional evidence for the inference of inherited genetic variants on ALL leukemogenesis, further indicating that a continuum of genetic variations in both host and tumour genomes contribute to malignant transformation and cancer risk.

**Methods**

**Subjects and samples.** The discovery GWAS consisted of 1,773 childhood B-ALL cases and 10,448 non-ALL controls of European descent (>90% European genetic ancestry as estimated using STRUCTURE\(^{32,33}\)). All cases were from the Children’s Oncology Group (COG) AALL0232 study (N = 1,277)\(^{7}\), the COG 1999 protocol (N = 115)\(^{34}\) and St Jude Total Therapy XIIIB and XV protocols (N = 381)\(^{7}\). Unrelated individuals of European descent from the Aerohtherosclerosis Risk in Communities (ARIC) study\(^{20,21}\) were used as non-ALL controls because the prevalence of adult survivors of childhood ALL is less than 1 in 10,000 in the US. The replication series included 409 children with ALL from NOPHO ALL92, ALL2000 and ALL2008 protocols\(^{35,36}\) and 1,599 unrelated non-ALL controls from Danish registry (Danish Medical Biobank study 1000 INK4A in Holbæk and at random schools in Zealand, Denmark. ALL cases were selected only on the basis of sample availability, and we did not observe any statistically significant differences in demographic or clinical features of children included versus not included in this genetic study. We elected to focus on individuals of European descent to minimize population stratification\(^{36}\). Germline DNA for cases was extracted from peripheral blood or bone marrow samples obtained during clinical remission (<5% ALL blasts by morphology). This study was approved by the Institutional Review Board at St Jude Children’s Research Hospital and COG member institutions and the Ethics Committee at the Danish Data Protection Agency, Region Zealand and the University Hospital Righospitalet, Denmark. Informed consent was obtained from parents, guardians, or patients, as appropriate.

**Genotyping and quality control.** SNP genotyping was performed in germline DNA using the Illumina Infinium HumanExome Array v1.0 in the discovery GWAS, and using Illumina HumanCoreExome chip for the replication series. Genotype calls (coded as 0, 1, and 2 for AA, AB and BB genotypes) were determined using the Illumina Genomestudio Software. For the ALL cases, samples for which genotypes were not obtained at <98% SNP on the array were removed and were excluded from the analyses. Quality control procedures were performed for both samples and SNPs on the basis of call rate, minor allele frequency (MAF), and Hardy Weinberg equilibrium (Supplementary Fig. 5). Detailed quality control for the non-ALL controls from the ARIC study was performed at the University of Texas Health Science Center following established protocols\(^{37}\).

We performed principal component analysis of cases and controls in the discovery GWAS to characterize population substructure (Supplementary Fig. 6).

**Genome-wide analyses.** In the discovery GWAS, the association of each SNP individually with ALL susceptibility was tested by comparing the genotype frequency between ALL cases and non-ALL controls in logistic regression models, after adjusting for top 10 principal components to control for population stratification. A quantitative-quantitative (Q-Q) plot was constructed and the Manhattan plot was only minimal inflation at the upper tail of the distribution (λ = 1.08, Supplementary Fig. 7). In the replication studies, we evaluated the novel genome-wide significant variant rs3731249, using the same logistic regression models. Multivariate logistic regression model including both rs3731217 and rs3731249 were tested to determine independent association signals at the CDKN2A locus in both discovery and replication series.

We also performed gene-level analyses to evaluate the aggregated effects of low-frequency variants on ALL susceptibility, using the SKAT test\(^{38}\). Missense, stop codon-altering and splice-site variants with MAF <5% were included. In total, 12,679 genes with at least two variants were tested.

**CDKN2A–CDKN2B resequencing and rare variant analyses.** Germline DNA from 2,407 children with ALL was used to create individual Illumina dual-indexed libraries. These libraries were pooled in sets of 96 and hybridized with a custom version of the Roche NimbleGen SeqCap EZ custom probes to capture the C-myc, C–JAK2 region on 9p21. Quantitative PCR was used to determine the appropriate capture product titre necessary to efficiently populate an Illumina HiSeq 2000 flowcell for paired-end 2 × 101 bp sequencing. Each sequence pool of 96 samples was demultiplexed, with coverages of >20 × depth across >90% of the targeted regions for nearly all samples. Sequence reads in FASTQ format were streamed and aligned using the BWA algorithm\(^{39}\) and signed with the Genome Analysis Toolkit (GATK) pipeline\(^{40}\). In total, 20,933,984,037 SNP and 18,911,286,581 indel alignments were called using the GATK pipeline version 3.1 (ref. 38). We calculated the proportion of rare variant-carriers in ALL subjects (either homozygous or heterozygous) versus that in individuals of European descent in the ESP cohort (non-ALL controls), focusing on variants with MAF <1%. Statistical significance of the differences was estimated using Fisher’s exact test.

CDKN2A sequencing was also performed in matched germline and diagnostic ALL tumour DNA by Complete Genomics for all cases with available data, and in tumour RNA by RNA-seq. Details regarding sequencing, data analysis and coverage are available at ftp://caftpd.nci.nih.gov/pub/bcc_target/ALL/Phase_II/sequence/WGS/CGI_TARGET_Pipeline_README.pdf or as previously described\(^{39}\) (European Genome Phenome archive: EGAS00001000654).

**Leukemic transformation assay in Ba/f3 cells.** The full-length CDKN2A was purchased from GE Healthcare. The p.A141T variant (rs3731249) was introduced by site-directed mutagenesis (forward primer: 5’-TGCGCCGACGATGAGGGG AGTGTCGCTCCAG-3’; reverse primer: 5’-TCTGAGGGCACTTCCGTCATCGTCGGGCGCA-3’). Promoter region of p14\(^{ARF}\) was amplified using the T7-oligo primer and ligated into pGEM-T Easy vector (Promega). 293 T cells (American Type Culture Collection) were transiently transfected with 5 µg of the plasmid and harvested 48 h after transfection. Supernatants containing cl20c–p16\(^{INK4A}\) or cl20c–p16\(^{INK4A}\) were collected and analyzed using the GATK pipeline version 3.1 (ref. 38). We compared the proportion of rare variant-carriers in ALL subjects (either homozygous or heterozygous) versus that in individuals of European descent in the ESP cohort (non-ALL controls), focusing on variants with MAF <1%. Statistical significance of the differences was estimated using Fisher’s exact test.

CDKN2A sequencing was also performed in matched germline and diagnostic ALL tumour DNA by Complete Genomics for all cases with available data, and in tumour RNA by RNA-seq. Details regarding sequencing, data analysis and coverage are available at ftp://caftpd.nci.nih.gov/pub/bcc_target/ALL/Phase_II/sequence/WGS/CGI_TARGET_Pipeline_README.pdf or as previously described\(^{39}\) (European Genome Phenome archive: EGAS00001000654).

**Leukemic transformation assay in Ba/f3 cells.** The full-length CDKN2A was purchased from GE Healthcare. The p.A141T variant (rs3731249) was introduced by site-directed mutagenesis (forward primer: 5’-TGCGCCGACGATGAGGGG AGTGTCGCTCCAG-3’; reverse primer: 5’-TCTGAGGGCACTTCCGTCATCGTCGGGCGCA-3’). Promoter region of p14\(^{ARF}\) was amplified using the T7-oligo primer and ligated into pGEM-T Easy vector (Promega). 293 T cells (American Type Culture Collection) were transiently transfected with 5 µg of the plasmid and harvested 48 h after transfection. Supernatants containing cl20c–p16\(^{INK4A}\) or cl20c–p16\(^{INK4A}\) were collected and analyzed using the GATK pipeline version 3.1 (ref. 38). We compared the proportion of rare variant-carriers in ALL subjects (either homozygous or heterozygous) versus that in individuals of European descent in the ESP cohort (non-ALL controls), focusing on variants with MAF <1%. Statistical significance of the differences was estimated using Fisher’s exact test.

CDKN2A sequencing was also performed in matched germline and diagnostic ALL tumour DNA by Complete Genomics for all cases with available data, and in tumour RNA by RNA-seq. Details regarding sequencing, data analysis and coverage are available at ftp://caftpd.nci.nih.gov/pub/bcc_target/ALL/Phase_II/sequence/WGS/CGI_TARGET_Pipeline_README.pdf or as previously described\(^{39}\) (European Genome Phenome archive: EGAS00001000654).
For quantitative reverse transcription-PCR (qRT-PCR), total RNA was extracted using the RNAasy Micro kit (Qiagen) according to the manufacturer’s protocol. Total RNA (500 ng) was reverse-transcribed into cDNA using oligoT primers and the SuperScript III reverse transcriptase kit (Invitrogen). Quantitative real-time PCR was performed by using ABI Prism 7900HT detection system (Applied Biosystems) with Faststart SYBR Green master mix (Roche). Relative expression was calculated as a ratio of BCR-ABL1 to Hprt. Primer sequences of BCR-ABL1 and Hprt were as follows: BCR-ABL1 (forward: 5′-CTGGCCACAGATGGCAAC-3′; reverse: 5′-GACGATGTGTTCATTCTC-3′; Hprt (forward: 5′-TTCTCCTTGGATGGAAT-3′).

To co-express rs3731249 variant and wild-type p16INKA BA/Bc cells were transduced with equal molar cL20c-p16INKA-3′-IRES-GEFP and cL20c-p16INKA-3′-IRES-IPFP lentivirus and cells successfully transduced with both were selected by flow cytometry sorting for GFP/YFP double positivity.

Luciferase reporter assays. The p14INKA 3′-UTR vector (3′-UTR for Human NM_005189.2) was placed downstream of luciferase reporter gene on the pEXZ-MT01 backbone was purchased from GeneCopeia and the T variant at rs3731249 was introduced by site-directed mutagenesis (forward primer: 5′-CCATGGCCGCGATAGCCTGGCTCCTGACATCC-3′; reverse primer: 5′-GGATGTCCTGACGCTTCCACGACTTGGCAGTGG-3′). For reporter gene assay, 2.5 × 10⁶ 293 T cells cultured in 96-well plate were transiently transfectioned with 100 ng empty vector, or variant, or wild-type p14INKA 3′UTR constructs using Lipofectamine 2000 (Invitrogen). Firstfire luciferase activities were measured 24 h after transfection by the Luciferase Assay System (Promega). The results were normalized against Renilla luciferase. Each reporter construct transfection was replicated at least three times, and each sample was assayed in triplicate.

Acknowledgements
We thank the patients and parents who participated in the clinical protocols included in this study and the clinicians and research staff at participating institutions. J.J.Y. is supported by the American Society of Hematology Scholar Award and by the Order of St. Francis Foundation. H.Z. is a St Baldrick’s International Scholar. V.P.A is supported by the Spanish Ministry of Education Fellowship Grant and by the St Jude Children’s Research Hospital Academic Programs Special Fellowship. C.G.M. is a Pew Scholar in the Biomedical Sciences and by the Order of St. Francis Foundation. H.Z. is a St Baldrick’s International Scholar. C.G.M. is supported by the National Institutes of Health (grant numbers CA62926 and CA155878) and by the St. Jude Children’s Research Hospital Academic Programs Special Fellowship. C.G.M. is a Pew Scholar in the Biomedical Sciences and by the Order of St. Francis Foundation. H.Z. is a St Baldrick’s International Scholar. C.G.M. is supported by the National Institutes of Health (grant numbers CA155878 and CA155878). W.L. is a St. Jude Children’s Research Hospital Academic Programs Special Fellowship. C.G.M. is a Pew Scholar in the Biomedical Sciences and by the Order of St. Francis Foundation. H.Z. is a St Baldrick’s International Scholar. C.G.M. is supported by the National Institutes of Health (grant numbers CA62926 and CA155878). W.L. is a St. Jude Children’s Research Hospital Academic Programs Special Fellowship. C.G.M. is a Pew Scholar in the Biomedical Sciences and by the Order of St. Francis Foundation. H.Z. is a St Baldrick’s International Scholar. C.G.M. is supported by the National Institutes of Health (grant numbers CA62926 and CA155878).
The authors also thank the staff and participants of the ARIC study for their important contributions.

**Author contributions**


**Additional information**

**Accession codes.** The RNA-seq data have been deposited in European Genome Phenome archive under the accession codes EGAS00001000654.

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/


This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/