SUPPLEMENTARY METHODS

USA case control sets.

Patients with *de novo* AML included in both USA sets were enrolled onto Cancer and Leukemia Group B (CALGB) companion protocols CALGB 8461 (cytogenetic studies), CALGB 9665 (leukemia tissue bank) and CALGB 20202 (molecular studies), and were similarly treated on CALGB trials[1-9]. CALGB is now part of the Alliance for Clinical Trials in Oncology (Alliance). Patients with acute promyelocytic leukemia, AML secondary to myelodysplastic syndromes and patients with therapy-related AML were not included in the study. The 2369 USA control subjects were individuals of self-reported European descent recruited in Columbus, Ohio, and excluded individuals previously diagnosed with any cancer. For the USA1 sample set, 1572 of the controls were from a previously genotyped sample series and reanalyzed for this study[10]. The remaining 142 controls from USA1 and the 655 controls for the USA2 sample set were genotyped from blood DNA as previously reported[10], and briefly described below. DNA from USA AML patients was obtained from pre-treatment bone marrow or blood samples. Study protocols were in accordance with the Declaration of Helsinki and approved by the institutional review boards at each center within the Alliance, and all patients and healthy donors provided written informed consent. Statistical analyses were reviewed by the Alliance Statistics and Data Center.

German case control set.

Three hundred fifty German AML patient samples were obtained from the University of Leipzig. All patients were enrolled on *Ostdeutsche Studiengruppe für Hämatologie und Onkologie* (OSHO) protocols, OSHO 033, OSHO 061, OSHO 069, OSHO 075, and OSHO 083. DNA was extracted from pretreatment blood or bone marrow samples at the University of Leipzig, and genotyping was performed at the Ohio State University using a MassARRAY system as described below. Written informed consent for participation in these studies was obtained in accordance with the Declaration of Helsinki. Control samples were collected by random sampling of inhabitants of the city of Leipzig for the LIFE-Adult study[11]. Genotyping and imputation were performed as described[12]. One thousand six hundred sex-matched controls who were on average older than the cases (72 years compared to 57 years), and had never been diagnosed with leukemia or other cancer, were selected for this study.

Whole-genome genotyping and informatics methods.

Genotyping of all USA case and control samples was performed by deCODE Genetics (Reykjavík, Iceland) using Infinium Omni-1 Quad-bead arrays (Illumina, San Diego, CA). Samples with <94% genotyping yield were excluded. Variants were excluded if they had <94% yield, showed significant differences among genotyping batches, or if they significantly (P<10^{-6}) deviated from Hardy-Weinberg equilibrium. To achieve genome-wide coverage, phasing and imputation was performed using SHAPEIT[13] and IMPUTE2[14] with 1000 Genomes Phase 3 reference data[15]. To control for spurious associations due to cryptic ancestry differences between the cases and controls, population structure was determined using ADMIXTURE[16] and EIGENSOFT[17], and samples were excluded if they were not at least 90% European ancestry.

MassARRAY genotyping.

Genotyping of the suggestive association SNPs in the German AML cases was performed using iPLEX Gold kits and a MassARRAY® Analyzer Four System, 384/96 GT (Agena Bioscience, San Diego, CA). Oligonucleotides used for genotyping are listed in Supplementary Table S1.

Genome-wide association testing and combined analysis.

Logistic regression was performed using SNPTEST v2.5[18] to test for association between variants and AML, assuming an additive model, treating AML status as the response and expected genotype counts from imputation as covariates. Ten principal components were included as co-variates, and *P*-values were adjusted for age and sex. Results were combined from the two USA groups using METAL[19] assuming fixed effect *P*-values. We only included variants that were confidently genotyped by excluding variants with an imputation information score <0.8. The final association test combining USA1, USA2 and the German sample sets was performed using METAL[34] assuming fixed effect *P*-values. P<10^{-6} was considered a suggestive association and P<$5x10^{-8}$ was considered significant. The Manhattan plot and the quantile-quantile plot were made with the qqman R package and the regional association plots were made with LocusZoom[20].

Genomic characterization analyses using datasets

Correlation between rs75797233 genotype and *BICRA* expression was examined using the Genotype-Tissue Expression (GTEx) project database by querying the whole blood tissue type in GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)[21]. Methylation and transcription factor binding within the AML risk loci were assessed by examining the monomethylation of histone H3 lysine 4 (H3K4Me1) in seven cell lines from ENCODE[22] track, and the transcription factor chromatin immunoprecipitation sequencing (ChIP-seq) from ENCODE[22] with factorbook motifs (March 2012 Freeze) track in the University of California, Santa Cruz genome browser[23], respectively. Transcription factor binding motifs were assessed using the MotifbreakR R package[24] run on the Homo Sapiens Comprehensive Model Collection of Transcription Factor Binding Models dataset[25].

Lymphoblastoid culture from blood

Lymphoblastoid cell cultures were established by mixing RPMI-1640 (ThermoFisher, Waltham, MA [Gibco]) with peripheral blood, then separating lymphocytes using Ficoll-Paque (GE Healthcare, Chicago Il). Cells were re-suspended in warm Epstein-Barr virus containing media with cyclosporine-A and cultured for approximately seven days until they reached 75% confluence. Cells were then grown in RPMI-1640 and frozen into viable aliquots when they reached 75% confluence of 150 cm^2 growth area. For performing the indicated experiments cells were thawed and passaged for seven days.

Quantitative reverse transcription PCR

RNA from cell lines was isolated using TRizol reagent (ThermoFisher, Waltham, MA). Reverse transcription was performed on DNAse I treated RNA using the Superscript III First-Strand cDNA Synthesis Kit (ThermoFisher, Waltham, MA) with 2µg of input RNA. qPCR was performed using Fast SYBR Green Master Mix (ThermoFisher, Waltham, MA) and a 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster City, CA). Primer sequences for *BICRA* expression: F-ACCACAAAGTGGACGAGGAG; R-TTATTGAGCATGGCCTGGGT.

Chromatin immunoprecipitation

Chromatin immunoprecipitations were performed using the SimpleChiP Enzymatic Chromatin IP kit (Cell Signaling Technologies, Danvers, MA) and a GATA2 antibody (A0677, ABclomal Technology, Woburn, MA) Primer sequences for Sanger sequencing and qPCR of the rs75797233 locus: F-TCGTGATGATTCTAGTGGGTGT; R-GAAGGTGGTGAGGAAATGGC. Topoisomerase DNA I (TOPO) cloning of PCR products was performed with TOPO™ TA Cloning™ Kit for Subcloning (Invitrogen, Carlsbad CA). Twenty individual clones were sequenced for each PCR product.

SUPPLEMENTARY REFERENCES

1. Baer MR, George SL, Caligiuri MA, Sanford BL, Bothun SM, Mrózek K, *et al.* Low-dose interleukin-2 immunotherapy does not improve outcome of patients age 60 years and older with acute myeloid leukemia in first complete remission: Cancer and Leukemia Group B Study 9720. *J Clin Oncol* 2008; **26**: 4934-4939.
2. Blum W, Sanford BL, Klisovic R, DeAngelo DJ, Uy G, Powell BL, *et al.* Maintenance therapy with decitabine in younger adults with acute myeloid leukemia in first remission: a phase 2 Cancer and Leukemia Group B Study (CALGB 10503). *Leukemia* 2017; **31**: 34-39.
3. Kolitz, J.E. *et al.* Dose escalation studies of cytarabine, daunorubicin, and etoposide with and without multidrug resistance modulation with PSC-833 in untreated adults with acute myeloid leukemia younger than 60 years: final induction results of Cancer and Leukemia Group B Study 9621. *J Clin Oncol* 2004; **22**: 4290-4301.
4. Kolitz JE, George SL, Dodge RK, Hurd DD, Powell BL, Allen SL, *et al.* P-glycoprotein inhibition using valspodar (PSC-833) does not improve outcomes for patients younger than age 60 years with newly diagnosed acute myeloid leukemia: Cancer and Leukemia Group B study 19808. *Blood* 2010; **116**: 1413-1421.
5. Lee EJ, George SL, Caligiuri M, Szatrowski TP, Powell BL, Lemke S, *et al.* Parallel phase I studies of daunorubicin given with cytarabine and etoposide with or without the multidrug resistance modulator PSC-833 in previously untreated patients 60 years of age or older with acute myeloid leukemia: results of Cancer and Leukemia Group B study 9420. *J*. *Clin*. *Oncol*. 1999; **17**: 2831-2839.
6. Mayer RJ, Davis RB, Schiffer CA, Berg DT, Powell BL, Schulman P, *et al.* Intensive postremission chemotherapy in adults with acute myeloid leukemia. *N Engl J Med* 1994; **331**: 896-903.
7. Moore JO, Dodge RK, Amrein PC, Kolitz J, Lee EJ, Powell B, *et al.* Granulocyte-colony stimulating factor (filgrastim) accelerates granulocyte recovery after intensive postremission chemotherapy for acute myeloid leukemia with aziridinyl benzoquinone and mitoxantrone: Cancer and Leukemia Group B study 9022. *Blood* 1997; **89**: 780-788.

8. Moore JO, George SL, Dodge RK, Amrein PC, Powell BL, Kolitz JE, *et al.* Sequential multiagent chemotherapy is not superior to high-dose cytarabine alone as postremission intensification therapy for acute myeloid leukemia in adults under 60 years of age: Cancer and Leukemia Group B Study 9222. *Blood* 2005; **105**: 3420-3427.

9. Stone RM, Berg DT, George SL, Dodge RK, Paciucci PA, Schulman P, *et al.* Granulocyte-macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. *N Engl J Med* 1995 **332**: 1671-1677.

10. Gudmundsson J, Thorleifsson G, Sigurdsson JK, Stefansdottir L, Jonasson JG, Gudjonsson SA, *et al*. A genome-wide association study yields five novel thyroid cancer risk loci. *Nat Commun* 2017; **8**: 14417.

11. Loeffler M, Engel C, Ahnert P, Alfermann D, Arelin K, Baber R, *et al*. The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. *BMC Public Health* 2015; **15**: 691

12. Pott J, Burkhardt R, Beutner F, Horn K, Teren A, Kirsten H, *et al*. Genome-wide meta-analysis identifies novel loci of plaque burden in carotid artery. *Atherosclerosis* 2017; **259**: 32-40.

13. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013; **93**: 687-696.

14. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.

15. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015; **526**:68-74.

16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655-1664.

17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904-909.

18. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906-913.

19. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190-2191.

20. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, *et al*. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; **26**: 2336-2337.

21. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; **45**: 580-585.

22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57-74.

23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, *et al.* The human genome browser at UCSC. *Genome Res* 2002; **12**: 996-1006.

24. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* 2015; **31**: 3847-3849.

25. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, *et al*. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* 2018; **46**: D252-D259.

SUPPLEMENTARY TABLES

Supplementary Table S1. Oligonucleotides used for MassARRAY panel

SNP	Sequence	Primer type
rs6562807	ATAACTGCCACATCACCAGG	F
	TTCCCTTCCTCCTCCAAAAG	R
	AGTAGTAGTTTATCCCCTCCTGT	Ext
rs2039647	TGTTTGCTAACCTTCTTCCG	F
	CTTATGAGTCTACCATTGTC	R
	TCTACCATTGTCTATAATTCCT	Ext
rs4356363	CTTCATTCTCACAGGACAGC	F
	CAGGAATGAATGACCTTTCG	R
	AGAGACTGGCAGTTAGG	Ext
rs75797233	TGGGTGTGAGGTTGTATCTC	F
	GCCAAGAAACACAGGAAAGG	R
	CTTCCATTTCCCTAATACTG	E
rs139878336	ACTCCACCTTTATGAGAGTC	F
	TGGATCTGCCACTTGCTTTC	R
	ACTTGCTTTCTAGAAAAAAAAA	Ext
rs57706619	TTTATTTAGCAGAGACAGGG	F
	GTGGTCACTAAGGGATTAGG	R
	TTAGTTGCCCTGCCCTGA	Ext

Abbreviations: SNP, single nucleotide polymorphism; F, Forward primer; R, Reverse primer; Ext, Extension primer

Supplementary Table S2. Polymorphisms with *P*-values <10^{-6} in the acute myeloid leukemia association test with the USA1 and USA2 sample sets.

						USA1 (894 cases; 1714 controls)		USA2 (289 cases; 655 controls)		Combined analysis (USA1 and USA2)					
Locus	**Nearest gene**	**Polymorphism**	**RA**	**OA**	**Position**[a]	**OR**	***P*-value**	**OR**	***P*-value**	P_{het}	I^2	**RAF**$_{case}$	**RAF**$_{con}$	**OR**	***P*-value**[b]
1q41	*DUSP10*	rs731372	A	G	1:221550655	1.47	7.6 x 10^{-7}	1.19	0.21	0.19	41	0.17	0.14	1.40	7.2 x 10^{-7}
13q22	*KLF12*	rs6562807	G	A	13:74739819	1.42	1.0 x 10^{-6}	1.20	0.12	0.22	33	0.25	0.21	1.36	6.1 x 10^{-7}
		rs139878336	T	-	13:74744123	1.43	6.2 x 10^{-7}	1.19	0.15	0.17	46	0.25	0.21	1.36	5.2 x 10^{-7}
		rs2039647	G	A	13:74763378	1.41	9.9 x 10^{-7}	1.23	0.077	0.31	4	0.25	0.21	1.36	3.1 x 10^{-7}
		rs7337898	C	G	13:74769241	1.41	1.2 x 10^{-6}	1.23	0.069	0.33	0	0.26	0.21	1.36	3.2 x 10^{-7}
		rs4356363	A	G	13:74773253	1.40	1.3 x 10^{-6}	1.20	0.11	0.24	26	0.26	0.21	1.35	6.6 x 10^{-7}
18q22	*ZNF407*	rs200806632	TTAAG	-	18:72465223	2.29	1.4 x 10^{-6}	1.56	0.13	0.25	23	0.040	0.022	2.07	7.4 x 10^{-7}
19p13	*B3GNT3*	rs11670628	G	A	19:17904166	1.23	0.00102	1.65	6.62 x 10^{-6}	0.02	81	0.41	0.35	1.32	4.0 x 10^{-7}
		rs2240811	C	T	19:17907444	1.22	0.00198	1.66	5.09 x 10^{-6}	0.01	83	0.40	0.34	1.31	8.2 x 10^{-7}
19q13	*BICRA*	rs118082870	T	C	19:48094682	2.30	2.7 x 10^{-5}	2.21	0.0050	0.91	0	0.028	0.015	2.27	3.0 x 10^{-7}
		rs75797233	T	A	19:48099347	2.28	3.8 x 10^{-5}	2.58	0.0039	0.99	0	0.029	0.015	2.28	3.3 x 10^{-7}

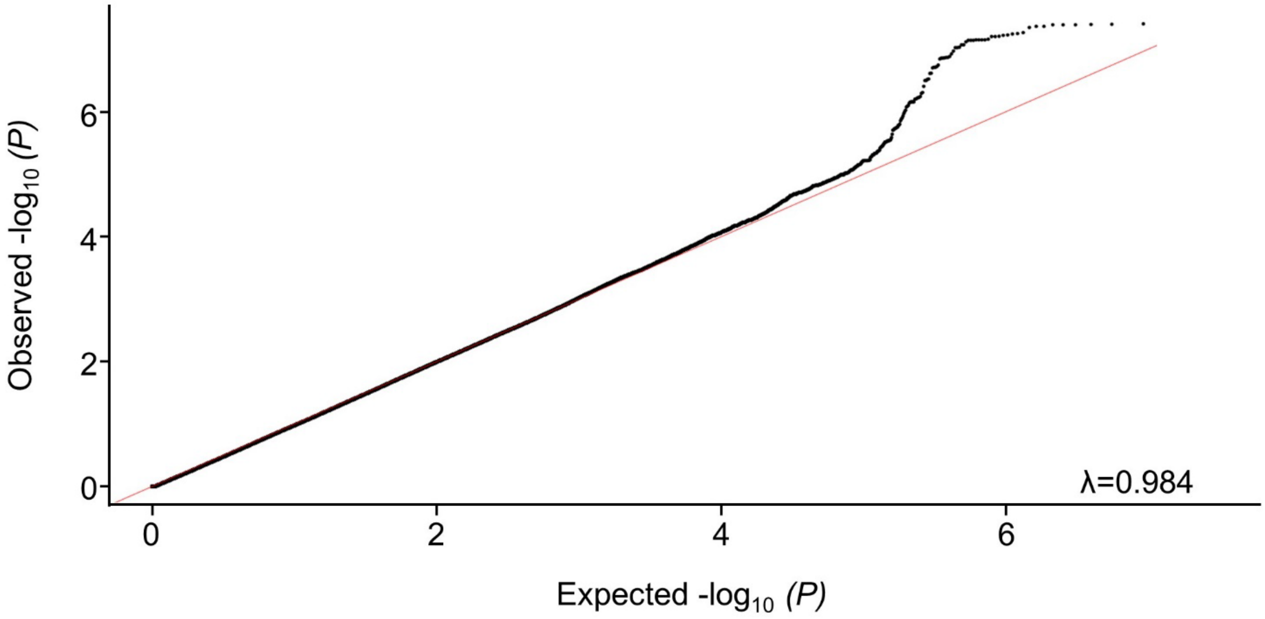
Abbreviations: RA, risk allele; OA, other allele; OR, odds ratio; I^2, heterogeneity statistic representing the fraction of variability due to heterogeneity between study groups; P_{het}, *P*-value from chi-squared test of heterogeneity; RAF$_{case}$, risk allele frequency in the cases; RAF$_{con}$, risk allele frequency in the controls; A, adenine; G, guanine; T, thymine; C, cytosine. [a]Position is given according to GRCh37 human genome build. [b]Association testing between variants and disease was performed using logistic regression.

Supplementary Table S3. Association test results for candidate single nucleotide polymorphisms (SNPs) in the German sample set.

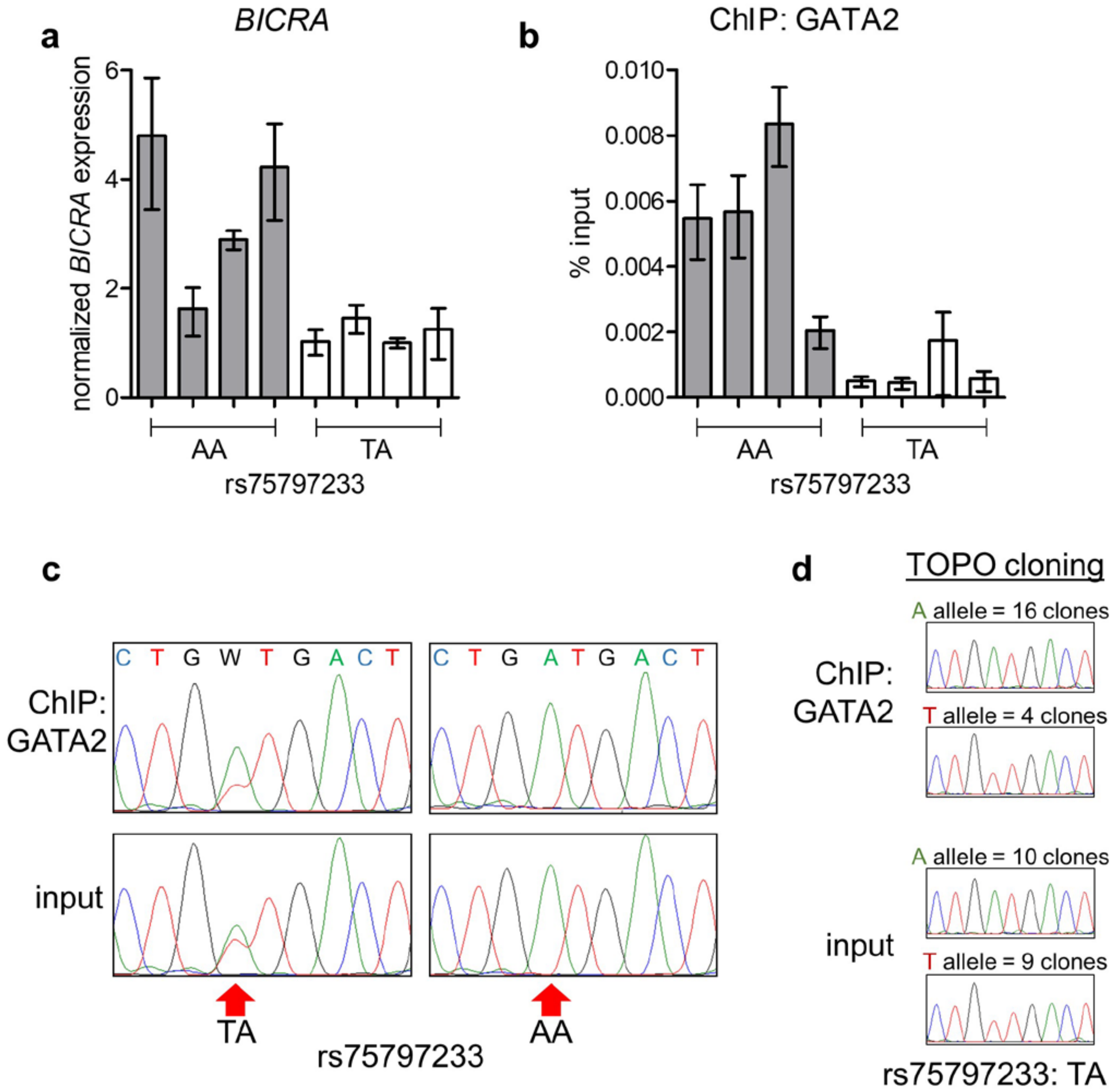
Locus	Nearest gene	SNP	RA	OA	Position[a]	RAF_{case}	RAF_{con}	OR	*P*-value[b]
13q22	*KLF12*	rs6562807	G	A	13:74739819	0.23	0.22	1.17	0.27
		rs139878336	T	-	13:74744123	0.22	0.22	1.07	0.66
		rs2039647	G	A	13:74763378	0.24	0.22	1.18	0.23
		rs4356363	A	G	13:74773253	0.23	0.23	1.19	0.23
19p13[c]	*B3GNT3*	rs57706619	T	C	19:17915881	0.43	0.35	1.44	0.038
19q13	*BICRA*	rs75797233	T	A	19:48099347	0.036	0.023	1.97	0.046

Abbreviations: RA, risk allele; OA, other allele; RAF_{case}, risk allele frequency in the cases; RAF_{con}, risk allele frequency in the controls; OR, odds ratio; G, guanine; A, adenine; T, thymine; C, cytosine. [a]Position is given according to GRCh37 human genome build. [b]Association testing between variants and disease was performed using logistic regression. [c]rs57706619 was chosen as a surrogate marker for genotyping the 19p13 locus because primers could not be designed for genotyping rs11670628 and rs2240811, and the genotype of rs57706619 is concordant with the genotypes of rs11670628 and rs2240811 in 99% of genomes from the study population.

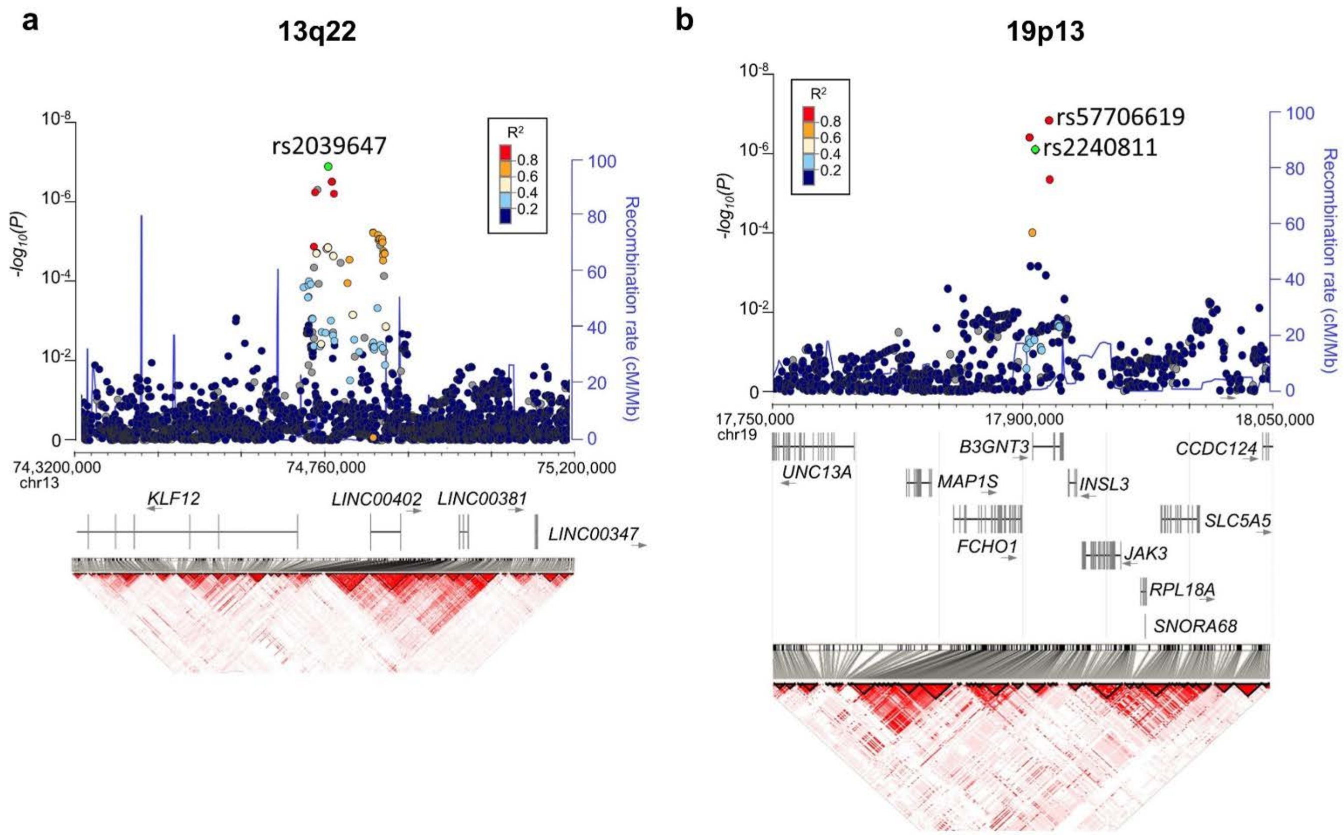
SUPPLEMENTARY FIGURES



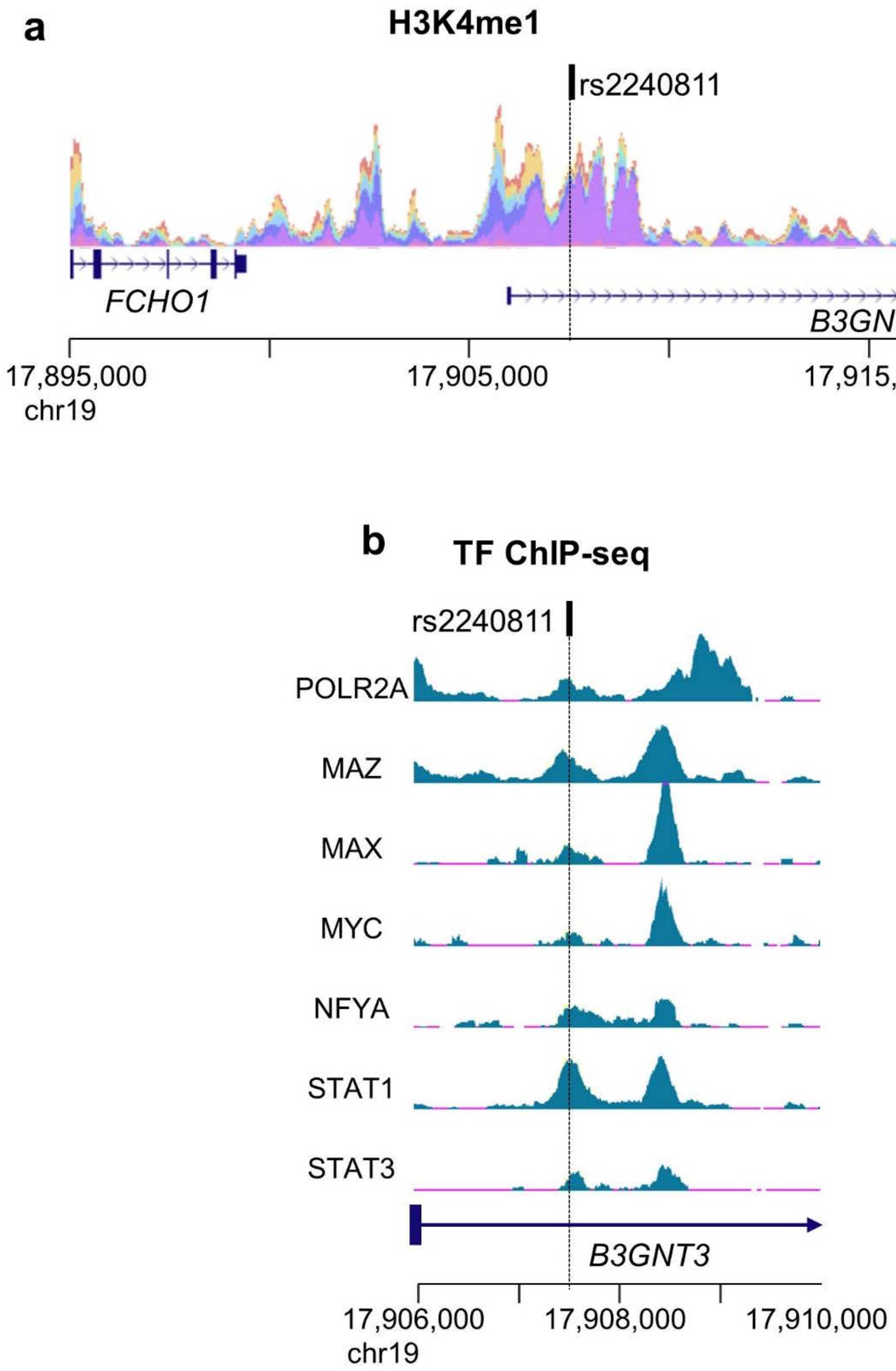
Supplementary Figure S1. Quantile-quantile plot of combined genome-wide association studies from USA1 and USA2 sample sets. The -log$_{10}$(*P)* of the observed (Y-axis) and expected (X-axis) *P*-values are shown for all tested polymorphisms. The red line represents the null hypothesis that observed *P*-values are exactly equal to expected *P*-values. Because deviation from the null hypothesis line is evident only in the tail area of the graph, it suggests that population stratification was adequately controlled, and discovered associations are not spurious.



Supplementary Figure S2. The rs75797233 AML risk allele (Thymine, T) reduces GATA2 binding and is associated with lower *BICRA* expression compared to rs75797233 (Alanine, A). **a** Quantitative reverse transcription PCR (qRT-PCR) performed on lymphoblastoid cell lines from four individuals who are homozygous for the rs75797233 non-risk allele (AA) and four heterozygotes (TA). Expression is significantly higher (*P*=0.02, two-sided T-test) in heterozygotes. **b** Chromatin immunoprecipitation (ChIP) for GATA2 was performed on the same cell lines as in (**a**), and binding to the rs75797233 locus was quantified by qPCR. Levels are significantly higher (*P*=0.01, two-sided T-test) in heterozygotes. Error bars in (**a**) and (**b**) are standard error of the mean from triplicates. **c** ChIP Sanger sequencing traces show preferential binding of GATA2 to rs75797233 (A) allele (green peak) compared to input in the heterozygous cell line (TA). **d** Topoisomerase DNA I (TOPO) cloning and Sanger sequencing was used to quantify the ratio of (A) to (T) alleles in the PCR products of the heterozygous cell line shown in (**c**). The alleles of 20 sequenced clones from the ChIP:GATA2 and input DNA PCR products are given.



Supplementary Figure S3. Regional association and linkage disequilibrium plots for suggestive association loci from combined analysis of USA1, USA2 and German sample sets. Plots show results for association with acute myeloid leukemia (AML) for all tested polymorphisms at 13q22 (**a**) and 19p13 (**b**). The Y-axes show the -log10 *P*-values of the polymorphisms and the X-axes show their chromosomal positions. 13q22 polymorphisms are colored according to their linkage disequilibrium (R^2) with rs2039647 (colored in green), and 19p13 polymorphisms are colored according to their linkage disequilibrium with rs2240811 (colored in green). The unbroken blue line behind the circles representing the polymorphisms is the recombination rate. The arrows under gene names show direction of transcription. The linkage disequilibrium plots presented in the lower tracks show the pairwise linkage disequilibrium between all polymorphisms in the region such that the red triangles indicate blocks of polymorphisms that are likely to be co-inherited. Linkage disequilibrium is plotted according to 1000 Genomes European population (November 2014 release), and genome is plotted according to human genome build GRCh37.



Supplementary Figure S4. Genomic context of the 19p13 single nucleotide polymorphism rs2240811. **a** Layered monomethylation of histone H3 lysine 4 (H3K4me1) shows a peak encompassing rs2240811 and the first intron of *B3GNT3,* which indicates that chromatin is open. **b** Transcription factor chromatin immunoprecipitation sequencing (TF ChIP-seq) for indicated transcription factors. Peaks indicate transcription factor binding to the encompassed region.