nature medicine

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41591-018-0081-z

In the format provided by the authors and unedited.

Somatic mutations precede acute myeloid leukemia years before diagnosis

Pinkal Desai[1,7]*, Nuria Mencia-Trinchant[1,7], Oleksandr Savenkov[2], Michael S. Simon[3], Gloria Cheang[4], Sangmin Lee[1], Michael Samuel[1], Ellen K. Ritchie[1], Monica L. Guzman[1], Karla V. Ballman[2], Gail J. Roboz[1,8] and Duane C. Hassane[1,5,6,8]*

[1]Division of Hematology and Oncology, Weill Cornell Medical College, New York, NY, USA. [2]Heath Care Policy and Research, Weill Cornell Medical College, New York, NY, USA. [3]Barbara Ann Karmanos Cancer Institute, Detroit, MI, USA. [4]Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA. [5]Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA. [6]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medical College, New York, NY, USA. [7]These authors contributed equally: Pinkal Desai, Nuria Mencia-Trinchant. [8]These authors jointly supervised this work: Gail J. Roboz, Duane C. Hassane. *e-mail: pid9006@med.cornell.edu; dhassane@med.cornell.edu

SUPPLEMENTAL APPENDIX

Somatic Mutations Precede Acute Myeloid Leukemia Years Before Diagnosis

Pinkal Desai, MD, MPH[1†*], Nuria Mencia-Trinchant, PhD[1†], Oleksandr Savenkov, PhD[2], Michael S. Simon, MD MPH[3], Gloria Cheang[4], Sangmin Lee, MD[1], Michael Samuel[1], MD, Ellen K Ritchie, MD[1], Monica L. Guzman, PhD[1], Karla V Ballman, PhD[2], Gail J. Roboz, MD[1‡] and Duane C. Hassane, PhD[1,5,6‡*]

Table of Contents

SUPPLEMENTAL METHODS

Women's Health Initiative Cohort

Study population: The WHI enrolled more than 160,000 women in one or more of three clinical trials (CT group) or an observational study cohort (OS group) in 40 U.S. clinical centers from October 1, 1993 through December 31, 1998 with data collection updated through September 2012 and an average follow up of 10.8 years (SD 3.3 years). The participants in the CT group were followed at baseline and years 1, 3 and 6, with samples collected at WHI baseline, year 1 and 3 during these follow up visits. The participants in the OS group were followed at baseline and year 3 with samples collected similarly during these visits. Available clinical parameters for this study are white blood cell count (WBC), hemoglobin, platelet count, and hematocrit.

Case-control matching

Controls were matched to cases by age at baseline within 2 years, WHI component (clinical trial vs. observational study), history of non-myeloid cancers at baseline as well as the exact timing of blood draw and follow up. In addition to these criteria, controls were also matched by type and timing of any cancers that occurred in cases after WHI baseline, but before the diagnosis of AML. For example, if a clinical trial participant with AML had prior breast cancer and was age 60 at baseline, a corresponding 60-year-old control participant with prior breast cancer history was selected who never developed subsequent AML and had exactly the same timing of follow up visits and blood draws in follow up. Similarly, if a participant with AML (case) had breast cancer diagnosed after her year 1 PB draw, she was matched to a control who also had breast cancer diagnosed after her year 1 blood draw, but did not develop AML. Matching was done in a time forward manner to ensure that each control had as much control time as its matched case. For example, a participant who developed AML five years after randomization on the WHI protocol would be matched with a control with at least five years of follow-up.

Targeted Exome Sequencing

Genomic DNA was provided by WHI in a blinded manner, in which case-control status and clinical covariates were revealed only after variant calling was completed. Library generation and amplification were performed using a low error rate Hi-Fi DNA polymerase according to the Kapa HyperPrep protocol (Kapa Biosystems). Dual sample indexing, rather than single indexing, of libraries was performed to minimize signal spread errors arising from misidentification of multiplexed samples [35]. Targeted sequencing using a panel of 68 recurrently mutated genes in hematological malignancies was performed using a custom capture probes (Nimblegen) to a median coverage of 2000x for both AML cases and controls (Figure S1). Variant analysis was performed following rigorous quality control and filtration of low quality sequence information. To identify somatic variants, filtration based on population allele frequency data was applied so as to enrich somatic variants that are not likely inherited. To this end, variants were classified as probable somatic if exhibiting a dbSNP v142 or ExAC adjusted population allele frequency <= 0.25% or a median VAF in the cohort < 40%. Only mutations present at >1% VAF were evaluated for association with AML development and time-to-AML.

Statistical Analysis

Baseline characteristics of AML cases and matched controls were compared with the use of the two-sample t-test for continuous variables and the Fisher's exact test for categorical variables. Among the 188 cases, participants with baseline precursor mutations were compared to participants without precursor mutations with regard to demographic characteristics and baseline hematological characteristics (i.e. WBC count and differential counts, hemoglobin value and platelet count). The

relationship between specific precursor mutations and AML development were estimated by exact odds ratios (OR) and adjusted ORs were obtained from penalized-likelihood logistic regression [36]. Age was added as a continuous variable in all analysis and OR values were adjusted accordingly. OR and adjusted ORs are presented with their associated 95% confidence interval (95% CI). Multivariable penalized-likelihood logistic regression analysis was performed to assess the independent effect of demographic and prognostic factors of interest on precursor mutation status. Collinearity between predictors in the models was evaluated prior to the formulation of the final multivariable models. Time to development of AML was estimated with a Kaplan-Meier estimator. Differences between groups based on mutational status were evaluated with a log-rank test. Significant differences in variant allele fraction were determined in serial sampling using Fisher's Exact Test based on the count of supporting alternate and reference reads for each sample at a mutated site. For the absolute risk estimates we used a weighted partial likelihood approach and weighted baseline hazard [37,38]. We used sampling weights to adjust the contribution from controls, since the ratio of cases to controls is much higher in a case-control study than in the general population of interest. The weight for a control is proportional to a probability of being selected from a risk set (pool of potential controls for each case). This probability depends both on censoring time and on the matching variable. All p-values were two-sided with statistical significance set a priori at the 0.05 level. Ninety-five percent confidence intervals (95 % C.I.) were calculated to assess the precision of the obtained estimates.

Statistical tools.

All statistical analyses were performed with the use of R software v3.4.0[39]. Multivariable odds ratios and p-values were computed using Firth's Bias-Reduced Logistic Regression implemented in the *logistf v1.22* package[40]. Plots were produced using *ggplot2 v2.2.1* (https://github.com/tidyverse/ggplot2) . Data summarization and reshaping were performed using *plyr v1.8.4* (https://github.com/hadley/plyr), *dplyr v0.7.4* (https://github.com/hadley/dplyr) and *reshape2 v1.4.3* (https://github.com/hadley/reshape)

Next generation sequencing.

Following targeted enrichment according to Nimblegen protocols, libraries were sequenced on the Illumina HiSeq 4000 using dual-indexed sample adapters (Integrated DNA Technologies). To reduce errors arising from misalignment, reads were trimmed of contaminating adapter sequences and low-quality bases using Trimmomatic v0.32 (trimmed when median Illumina base quality score < 20 over 6 nt sliding window). To further improve sequence quality, overlapping paired end reads were merged into a single long consensus read using AdapterRemoval v2[41] when at least 12 bp overlap was present. The remaining high-quality reads were mapped against the 1000 genomes phase 2 human reference genome + decoy contigs (hs37d5) using BWA MEM[42]. Duplicate marking was performed using SamBlaster v0.1.21[43] and MarkDupsByStartEnd v0.2.1[44]. Single nucleotide variants (SNVs) and insertions/deletions (indels) were detected using VarDictJava v1.4.6[45] in single sample mode with indel realignment. Marked duplicates were excluded. Copy number variations (CNV) were detected using CNVkit v0.8.6[46]. Annotation of variants and their functional impact was performed using Variant Effect Predictor (VEP) v85[47] and snpEff v4.1g [48]. Indel representations in both the call set and annotation data were left-aligned and harmonized using vt analysis toolkit v0.5 to maximize concordance between variants and annotations[49].

The resulting call set was filtered with guidance regarding artifact removal as described[50]. Variants were classified as probable artifacts if any of the following conditions were met: occurs within a low complexity region subject to high alignment error defined in the hs37d5 reference genome according to the mdust-LC algorithm; exhibits strand bias; present within or immediately flanked by repetitive or low entropy regions (10 bp window); < 4 supporting reads per strand except for *NPM1* exon 12 (Ensembl

transcript ENST00000517671.5) where any supporting reads indicative of insertions longer than 4 nucleotides were considered; mean BWA MEM mapping quality of supporting reads < 45; mean Illumina base quality of variant-supporting bases < 30; exhibits read position bias toward the 5' or 3' end of reads; < 100x unique depth of high quality coverage at site of mutation; coverage depth x VAF < 8; exhibits high recurrence suggestive of artifact except for mutations known to be present >=10 times in COSMIC v74[51] or at least once in the TCGA AML study[52] when recurrence is defined by an identical mutation occurring in >20% of samples evaluated or the position is mutated in >40% of evaluated samples or identical mutation is present in >1 sequenced instance of NA12878 or the identical position is mutated in >2 sequenced instances of NA12878. Next, for each sample, we applied binomial probability filters based on the verifyBamID mix estimate that determined the lowest allelic fraction at which a somatic variant can be distinguished from sample cross-talk as described[53,54] and that ascertained the probability of being a heterozygous SNP based on the number of supporting alternate and reference reads controlling the false discovery rate at 1%. Subsequent filtering was performed for false variants that may arise from reads mapping to simple tandem repeats not flagged in previous steps. Variants above >=40% VAF that remained unfiltered following these steps were required to be present at <40% in same sample on serial evaluation or in the COSMIC or AML TCGA database. Finally, mutations classified by VEP were considered when categorized as missense, stop gain, splice acceptor, splice donor, frameshift insertion, frameshift deletion, in-frame insertion, and in-frame deletion mutations. Variants above 40% VAF that remained after these filtration steps included 4 instances JAK2 p.Val617Phe serially maintained in 2 study participants, 2 DNMT3A variants, and an isolated 18 nucleotide CREBBP in-frame insertion that is present in serial evaluation of the same subject and that may be a germline polymorphism exhibiting VAF under-representation as a result of capture bias.

Analyzed genes.

Coding exons and flanking DNA +/- 5 bp were evaluated for mutations in 68 genes including *ASXL1, ASXL2, ATRX, BCOR, BCORL1, BRAF, CALR, CARD11, CBL, CBLB, CBLC, CD70, CD79B, CDKN2A, CEBPA, CREBBP, CSF3R, CUX1, DIS3, DNMT3A, EPPK1, ETV6, EZH2, FAM46C, FBXW7, FLT1, FLT3, GATA1, GATA2, GNAS, HRAS, IDH1, IDH2, IKZF1, JAK1, JAK2, JAK3, KDM6A, KIT, KRAS, MPL, MYD88, NOTCH1, NPM1, NRAS, PAX5, PDGFRA, PHF6, PTEN, PTPN11, RAD21, RUNX1, SETBP1, SF3B1, SMC1A, SMC3, SRSF2, STAG1, STAG2, STAT6, TET1, TET2, TNF, TNFRSF14, TP53, U2AF1, WT1, ZRSR2*. Depth of coverage statistics were determined using Picard tools v2.6.0[55].

Pathogenicity of variants.

Detected variants were classified as "probable pathogenic" if any of the following conditions were met:

- Known hotspot in genes of known pathogenic significance in myeloid malignancies including *DNMT3A, IDH1, IDH2*, *SRSF2, SF3B1, TP53, JAK2, FLT3, NRAS, KRAS, KIT, MPL, CBL.*
- Disruptive mutations that produce frameshifts, duplications, stop codons, or affect splice acceptor or donor sites in genes where such alterations are known to be of pathogenic impact in myeloid malignancies including *DNMT3A, TET2, TP53, NPM1 exon 12, FLT3* exon 13/14*, RUNX1, ASXL1* exon 12/13, *CALR* exon 9, *CEBPA, ATRX, CBL, EZH2, BCOR, CREBBP, KDM6A, NOTCH1, RAD21, PHF6, CUX1*.
- Mutations demonstrating ≥ 10 instances in COSMIC v74 or previously identified in the AML TCGA study.
- SNVs classified as pathogenic using the Mendelian Clinically Applicable Pathogenicity (M-CAP) score classifier using the suggested cutoff of including variants scoring >0.025.[56,57]

Visualization and assessment of mutation spectrum, co-mutation spectrum, and clonal evolution.
Variants and clinical annotations were visualized as an OncoPrint using the *ComplexHeatmap v1.12.0* package[58] . Chord diagrams indicating co-mutational patterns of myeloid malignancy genes were producing using the *circlize v0.4.3* package[59]. Mutual exclusivity of and co-occurrence of mutated genes were determined using the *maftools v1.4.25* R package.

Determination of nucleotide composition and context of mutations.
Analysis of transitions, transversions, and relative proportion of mutation types was performed in R/BioConductor using the *maftools v1.4.25* package. Relative frequencies of transitions and transversions within trinucleotide and transcribed strand context was performed using the R/BioConductor *MutationalPatterns v1.0* package [60] using transcript strand notations derived from known transcript data present in the BioConductor *TxDb.Hsapiens.ucsc.hg19.knowngene* database with hg19 coordinates converted to hs37d5.

Functional domain analysis.
To determine the presence of mutations within known functional domains, individual mutations were mapped to Pfam functional domains based on the gene and protein-level HGVS description of each alteration (e.g. *DNMT3A* p.Arg882Cys) with the *maftools v1.4.25* package. Quantification and plotting of most frequently altered Pfam functional domains and the number of involved genes was performed using *maftools v1.4.25* package. 3-D special clusters were identified using mutation3D[61]. 3-D molecule visualizations were generated using PyMol 2.0 (Schrödinger, LLC; New York, NY).

Data availability statement
The datasets generated/or analyzed during the current study are available from the corresponding author on reasonable request.

SUPPLEMENTAL TABLES

Table S1. Summary of clinical parameters

Summary of clinical parameters for the entire cohort, AML cases or control participants (mutated, non-mutated and in total). Median and range values are provided for each variable, time to AML or last follow up (years), age, hematocrit, white blood cell (WBC), hemoglobin and platelet counts.

	All participants (n =369)			AML cases (n =188)			Controls (n =181)		
	Mutated n = 185	Non-mutated n = 184	Total n = 369	Mutated n = 129	Non-mutated n = 59	Total n = 188	Mutated n = 56	Non-mutated n =125	Total n = 181
Years to AML (cases) / follow up (controls)									
median	11.1	16.9	13.9	8.2	11.9	9.6	17.2	17.9	17.9
range	0.5-20.9	0.4-21	0.4-21	0.5-19.2	0.4-18.7	0.4-19.2	5-20.9	5.9-21	5-21
Age (years)									
median	68.1	64.4	66.1	67.6	62.9	66.2	69.2	65.3	66.1
range	51.5-79.7	51.4-79.7	51.4-79.7	51.5-79.7	52.6-78.4	51.5-79.7	52-79.4	51.4-79.7	51.4-79.7
Hematocrit (%)									
median	39.8	40	39.9	40	40	40	39.6	40	39.8
range	31.8-50.2	33.2-46.7	31.8-50.2	31.8-50.2	33.2-44.5	31.8-50.2	33.2-45.5	33.2-46.7	33.2-46.7
WBC (%)									
median	5.6	5.8	5.7	5.57	5.7	5.7	5.68	5.8	5.7
range	1.8-17.9	2.5-13	1.8-17.9	1.8-17.9	2.5-9.9	1.8-17.9	3.5-8.7	3.2-13	3.2-13
Hemoglobin (g/dL)									
median	13.5	13.5	13.5	13.5	13.3	13.5	13.3	13.6	13.5
range	10.4-16.4	10.8-16	10.4-16.4	10.4-16.4	10.8-15	10.4-16.4	10.8-15.3	11.7-16	10.8-16
Platelet ($1x10^9$/L)									
median	242	239.5	240	234	242	241.5	249	235	239
range	38-874	73-410	38-874	38-874	101-387	38-874	140-438	73-410	73-438

Table S2: Mutated genes in the case and control group

Number of participants with mutations in selected genes within the AML case or control groups overall and for patients < 65 or ≥ 65 years of age.

Gene		AML cases (N = 188) no. of participants (% of group)		Controls (N = 181) no. of participants (% of group)	
		Mutated	Non-mutated	Mutated	Non-mutated
DNMT3A	< 65	23 (28.75)	57 (71.25)	9 (11.69)	68 (88.31)
	≥ 65	46 (42.59)	62 (57.41)	25 (24.04)	79 (75.96)
	Total	69 (36.7)	119 (63.3)	34 (18.78)	147 (81.22)
TET2	< 65	8 (10)	72 (90)	2 (2.6)	75 (97.4)
	≥ 65	39 (36.11)	69 (63.89)	8 (7.69)	96 (92.31)
	Total	47 (25)	141 (75)	10 (5.52)	171 (94.48)
TP53	< 65	6 (7.5)	74 (92.5)	0 (0)	77 (100)
	≥ 65	15 (13.89)	93 (86.11)	0 (0)	104 (100)
	Total	21 (11.17)	167 (88.83)	0 (0)	181 (100)
SRSF2	< 65	2 (2.5)	78 (97.5)	0 (0)	77 (100)
	≥ 65	11 (10.19)	97 (89.81)	0 (0)	104 (100)
	Total	13 (6.91)	175 (93.09)	0 (0)	181 (100)
IDH2	< 65	2 (2.5)	78 (97.5)	0 (0)	77 (100)
	≥ 65	10 (9.26)	98 (90.74)	0 (0)	104 (100)
	Total	12 (6.38)	176 (93.62)	0 (0)	181 (100)
JAK2	< 65	5 (6.25)	75 (93.75)	0 (0)	77 (100)
	≥ 65	5 (4.63)	103 (95.37)	1 (0.96)	103 (99.04)
	Total	10 (5.32)	178 (94.68)	1 (0.55)	180 (99.45)
SF3B1	< 65	2 (2.5)	78 (97.5)	2 (2.6)	75 (97.4)
	≥ 65	9 (8.33)	99 (91.67)	0 (0)	104 (100)
	Total	11 (5.85)	177 (94.15)	2 (1.1)	179 (98.9)
ASXL1	< 65	1 (1.25)	79 (98.75)	2 (2.6)	75 (97.4)
	≥ 65	5 (4.63)	103 (95.37)	4 (3.85)	100 (96.15)
	Total	6 (3.19)	182 (96.81)	6 (3.31)	175 (96.69)
U2AF1	< 65	1 (1.25)	79 (98.75)	0 (0)	77 (100)
	≥ 65	5 (4.63)	103 (95.37)	0 (0)	104 (100)
	Total	6 (3.19)	182 (96.81)	0 (0)	181 (100)
RUNX1	< 65	1 (1.25)	79 (98.75)	0 (0)	77 (100)
	≥ 65	2 (1.85)	106 (98.15)	0 (0)	104 (100)
	Total	3 (1.6)	185 (98.4)	0 (0)	181 (100)
IDH1	< 65	1 (1.25)	79 (98.75)	0 (0)	77 (100)
	≥ 65	2 (1.85)	106 (98.15)	0 (0)	104 (100)
	Total	3 (1.6)	185 (98.4)	0 (0)	181 (100)

Table S3: Mutated cases in AML case group with time to AML less or greater than 5 years

Number of participants with mutations in selected genes within the AML case group with <5 or ≥ 5 years to the diagnosis of AML

Gene	Time to AML < 5 years (n = 48)	Time to AML ≥ 5 years (n = 140)
	Mutated cases (% of group)	Mutated cases (% of group)
TP53	11 (22.92)	10 (7.14)
RUNX1	3 (6.25)	0 (0)
DNMT3A	22 (45.83)	47 (33.57)
TET2	16 (33.33)	31 (22.14)
ASXL1	2 (4.17)	4 (2.86)
SRSF2	5 (10.42)	8 (5.71)
IDH2	4 (8.33)	8 (5.71)
JAK2	2 (4.17)	8 (5.71)
SF3B1	5 (10.42)	6 (4.29)
U2AF1	2 (4.17)	4 (2.86)
IDH1	1 (2.08)	2 (1.43)

Table S4. Effect of allelic fraction cutoff and variant classification on mutation rates in cases vs. controls.

Retained mutations are compared using different allelic fraction cutoffs and pathogenicity classification criteria. Approaches compared include (top to bottom on table below): **(i)** variants exhibiting >1% VAF are retained after filtering is performed as described in Supplemental Methods; **(ii)** SNVs present at >3.5% VAF and indels present >7% VAF are retained after filtering is performed as described in Supplemental Methods; **(iii)** same as (i) except only probable pathogenic mutations are retained; **(iv)** same as (ii) except only probable pathogenic mutations are retained; **(v)** variants present >1% VAF are retained when satisfying inclusion criteria described by Jaiswal et al 2014[62]; **(vi)** same as (v) using the same SNV and indel VAF cutoffs of >3.5% and >7% described by Jaiswal et al[62].

VAF cutoffs have a greater impact on clonal hematopoiesis (CH) rate in the control group than the variant classification criteria.

		AML cases (N=188)		Controls (N=181)		Odds Ratio	
		# Mutated (%)	# Non-mutated (%)	# Mutated (%)	# Non-mutated (%)	OR (95% CI)	P value
Desai classification, VAF cutoff 0.01	< 65	43 (53.75)	37 (46.25)	16 (20.78)	61 (79.22)	4.39 (2.08-9.61)	3.1×10^{-5}
	≥ 65	86 (79.63)	22 (20.37)	40 (38.46)	64 (61.54)	6.19 (3.25-12.14)	1.0×10^{-9}
	Total	129 (68.62)	59 (31.38)	56 (30.94)	125 (69.06)	4.86 (3.07-7.77)	3.8×10^{-13}
Desai classification, VAF cutoff 0.035 (SNV) or 0.07 (indels)	< 65	25 (31.25)	55 (68.75)	5 (6.49)	72 (93.51)	6.47 (2.24-23.05)	7.9×10^{-5}
	≥ 65	60 (55.56)	48 (44.44)	13 (12.5)	91 (87.5)	8.65 (4.2-18.98)	2.5×10^{-11}
	Total	85 (45.21)	103 (54.79)	18 (9.94)	163 (90.06)	7.43 (4.14-13.93)	1.1×10^{-14}
Desai classification, only Pathogenic variants, VAF cutoff 0.01	< 65	42 (52.5)	38 (47.5)	15 (19.48)	62 (80.52)	4.52 (2.12-10.05)	2.7×10^{-5}
	≥ 65	85 (78.7)	23 (21.3)	38 (36.54)	66 (63.46)	6.35 (3.35-12.4)	2.5×10^{-11}
	Total	127 (67.55)	61 (32.45)	53 (29.28)	128 (70.72)	5 (3.16-8.02)	1.1×10^{-14}
Desai classification, only Pathogenic variants, VAF cutoff 0.035 (SNV) or 0.07 (indels)	< 65	25 (31.25)	55 (68.75)	5 (6.49)	72 (93.51)	6.47 (2.24-23.05)	7.9×10^{-5}
	≥ 65	60 (55.56)	48 (44.44)	13 (12.5)	91 (87.5)	8.65 (4.2-18.98)	4.9×10^{-10}
	Total	85 (45.21)	103 (54.79)	18 (9.94)	163 (90.06)	7.43 (4.14-13.93)	1.6×10^{-13}
Jaiswal et al., 2014 classification, VAF cutoff 0.01	< 65	39 (48.75)	41 (51.25)	11 (14.29)	66 (85.71)	5.64 (2.5-13.65)	3.2×10^{-6}
	≥ 65	78 (72.22)	30 (27.78)	29 (27.88)	75 (72.12)	6.65 (3.54-12.84)	1.3×10^{-10}
	Total	117 (62.23)	71 (37.77)	40 (22.1)	141 (77.9)	5.78 (3.59-9.45)	3.3×10^{-15}
Jaiswal et al., 2014 classification, VAF cutoff 0.035 (SNV) or 0.07 (indels)	< 65	23 (28.75)	57 (71.25)	2 (2.6)	75 (97.4)	14.92 (3.45-136.04)	4.4×10^{-6}
	≥ 65	50 (46.3)	58 (53.7)	10 (9.62)	94 (90.38)	8.02 (3.67-19.16)	2.5×10^{-9}
	Total	73 (38.83)	115 (61.17)	12 (6.63)	169 (93.37)	8.89 (4.54-18.83)	3.9×10^{-14}

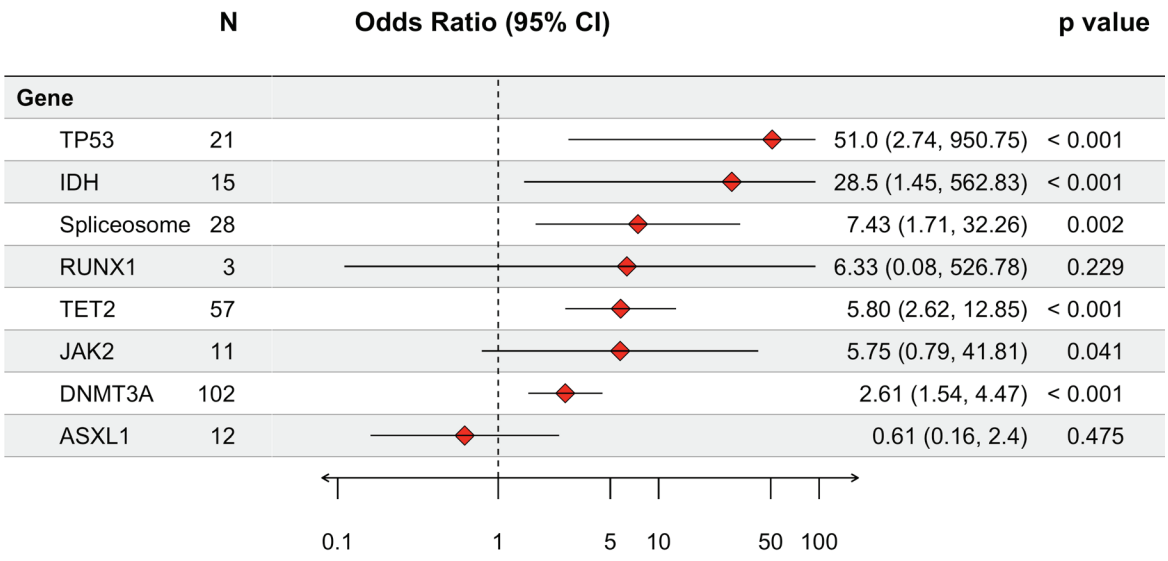
Table S5. Mutation frequency in AML cases and controls at varying SNP exclusion criteria

Comparison of the effect of varying population allele frequency cutoffs on the rate of mutations in cases and controls. (a) Population frequency cutoffs of 0.001%, 0.01%, 0.25% (applied in this study) and and 1% (recommendation of Association for Molecular Pathology (AMP) / American Society for Clinical Oncology (ASCO) / College of American Pathologists (CAP) for selecting somatic mutations) [63] (b) Forest plot indicating odds ratio of mutations in cases vs. controls using a population allele frequency cutoff 0.001%. Genes or gene categories significantly associated with AML include *TP53* ($P = 3.0 \times 10^{-6}$), IDH ($P = 3.0 \times 10^{-4}$), spliceosome, *TET2* ($P = 2.4 \times 10^{-6}$), and *DNMT3A* ($P = 3.4 \times 10^{-4}$). IDH category includes *IDH1* and *IDH2*. The spliceosome category includes SRSF2, SF3B1, and U2AF1. OR per gene are adjusted by age (years) as a continuous variable. Abbreviations: CI, confidence interval; N, number affected. P-values are shown for penalized likelihood multivariable logistic regression.

a

		AML cases (N=188)		Controls (N=181)		Odds Ratio	
		# Mutated (%)	# Non-mutated (%)	# Mutated (%)	# Non-mutated (%)	OR (95% CI)	P value
dbSNP & ExAC AF < 0.001%	< 65	43 (53.75)	37 (46.25)	16 (20.78)	61 (79.22)	4.39 (2.08-9.61)	3.1×10^{-5}
	≥ 65	86 (79.63)	22 (20.37)	40 (38.46)	64 (61.54)	6.19 (3.25-12.14)	1.0×10^{-9}
	Total	129 (68.62)	59 (31.38)	56 (30.94)	125 (69.06)	4.86 (3.07-7.77)	3.8×10^{-13}
dbSNP & ExAC AF < 0.01%	< 65	43 (53.75)	37 (46.25)	16 (20.78)	61 (79.22)	4.39 (2.08-9.61)	3.1×10^{-5}
	≥ 65	86 (79.63)	22 (20.37)	40 (38.46)	64 (61.54)	6.19 (3.25-12.14)	1.0×10^{-9}
	Total	129 (68.62)	59 (31.38)	56 (30.94)	125 (69.06)	4.86 (3.07-7.77)	3.8×10^{-13}
dbSNP & ExAC AF < 0.25%	< 65	43 (53.75)	37 (46.25)	16 (20.78)	61 (79.22)	4.39 (2.08-9.61)	3.1×10^{-5}
	≥ 65	86 (79.63)	22 (20.37)	40 (38.46)	64 (61.54)	6.19 (3.25-12.14)	1.0×10^{-9}
	Total	129 (68.62)	59 (31.38)	56 (30.94)	125 (69.06)	4.86 (3.07-7.77)	3.8×10^{-13}
dbSNP & ExAC AF < 1%	< 65	43 (53.75)	37 (46.25)	17 (22.08)	60 (77.92)	4.06 (1.95-8.78)	6.9×10^{-5}
	≥ 65	86 (79.63)	22 (20.37)	41 (39.42)	63 (60.58)	5.95 (3.13-11.65)	2.3×10^{-9}
	Total	129 (68.62)	59 (31.38)	58 (32.04)	123 (67.96)	4.62 (2.92-7.37)	1.9×10^{-12}

b



SUPPLEMENTAL FIGURES

Figure S1. Coverage across gene regions for top mutated genes – Part 1

Representative images for coverage of recurrently mutated genes in the cohort at the median depth of coverage of 2000x. For each gene, coverage across pertinent exons approaches or exceeds 2000x including important regions of driver genes such as *FLT3*. *CEBPA* also achieved >500x median coverage across its single coding exon. Median *NPM1* exon 12 coverage was relatively lower (~280x) potentially resulting in more false negatives. However, given the zero background rate of 4 nucleotide insertions in the *NPM1* insertion hotspot, this lower coverage maintains >80% power to detect *NPM1* insertions to a VAF > 1% (1-sample binomial power calculation; background mutation rate of 0.1%, alpha = 0.01).

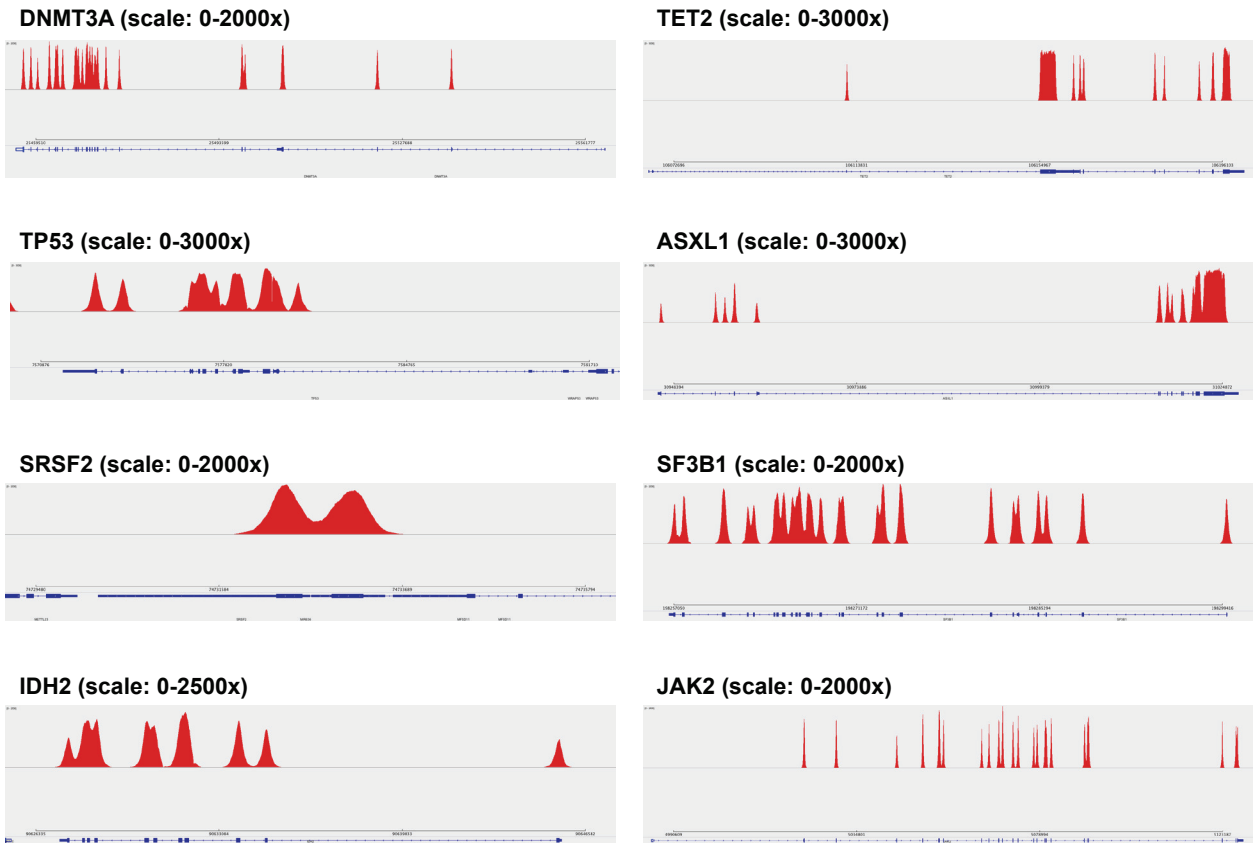


Figure S1. Coverage across gene regions for top mutated genes – Part 2

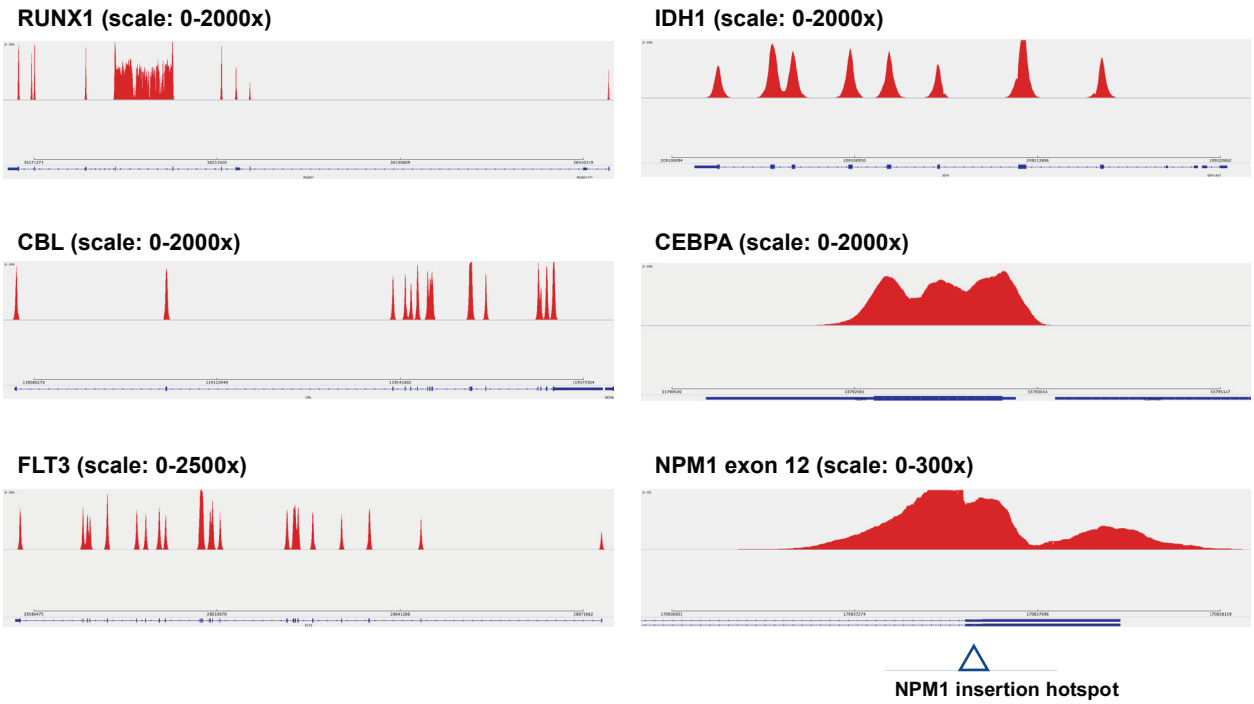


Figure S2. Spectrum and co-mutation pattern of pre-leukemic mutations – Part 1

(a) OncoPrint for all participants in the study at baseline. AML cases (N = 188) are represented in the left panel and controls (N = 181) in the right panel. Each row represents a gene and each column corresponds to a participant in the study. Bar plots indicate the number of mutations per patient (top bar plot), and the number of patients with mutations in each gene (side bar plot). For each patient, bottom panels show: time to AML diagnosis (Time to AML) for the cases or last follow up for the controls (Follow up time) and age at diagnosis (Age). Dark grey, patients older or equal than 65 years old; light grey, patients younger than 65 years old. Alterations classified as variant of unknown significance (VUS, green) or pathogenic (Pathogenic, magenta) according to the criteria specified in Supplemental Methods. All analyzed genes are included.

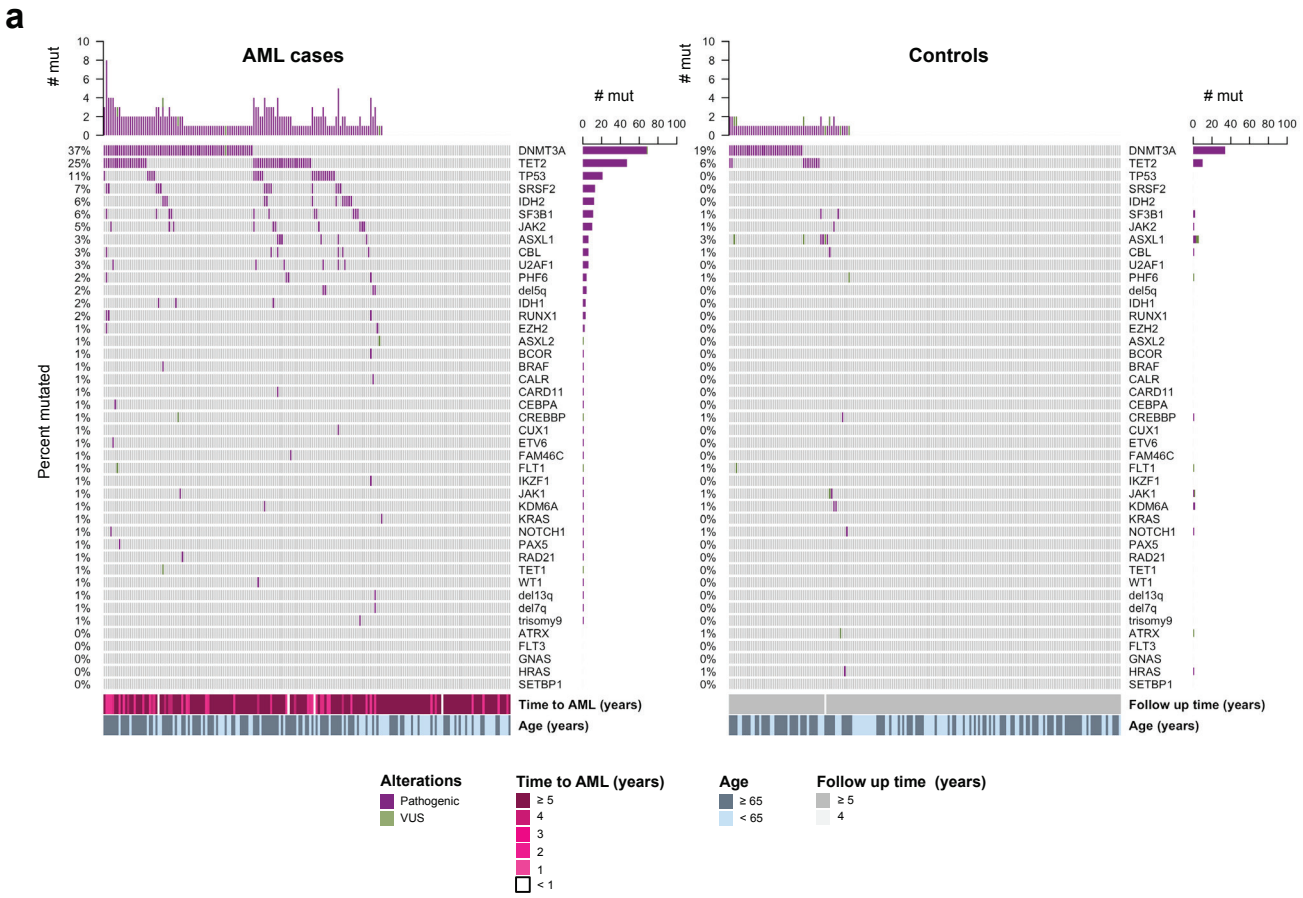


Figure S2. Spectrum and co-mutation pattern of pre-leukemic mutations – Part 2

(b) Oncoprint showing the number of variants per gene detected for each participant. Each row represents a gene and each column corresponds to a participant in the study. Bar plots indicate the number of mutations per patient (top bar plot), and the number of patients with mutations in each gene (side bar plot). For each patient, bottom panels indicate: time to AML diagnosis (Time to AML) for the cases or last follow up for the controls (Follow up time) and age at diagnosis (Age). Dark grey, patients older or equal than 65 years old; light grey, patients younger than 65 years old. Number of variants ranges from 1 to 4 (1, green; 2, light pink; 3, pink and 4, magenta).

Multiple variants per participant were mainly found in DNMT3A, TET2 and TP53 genes.

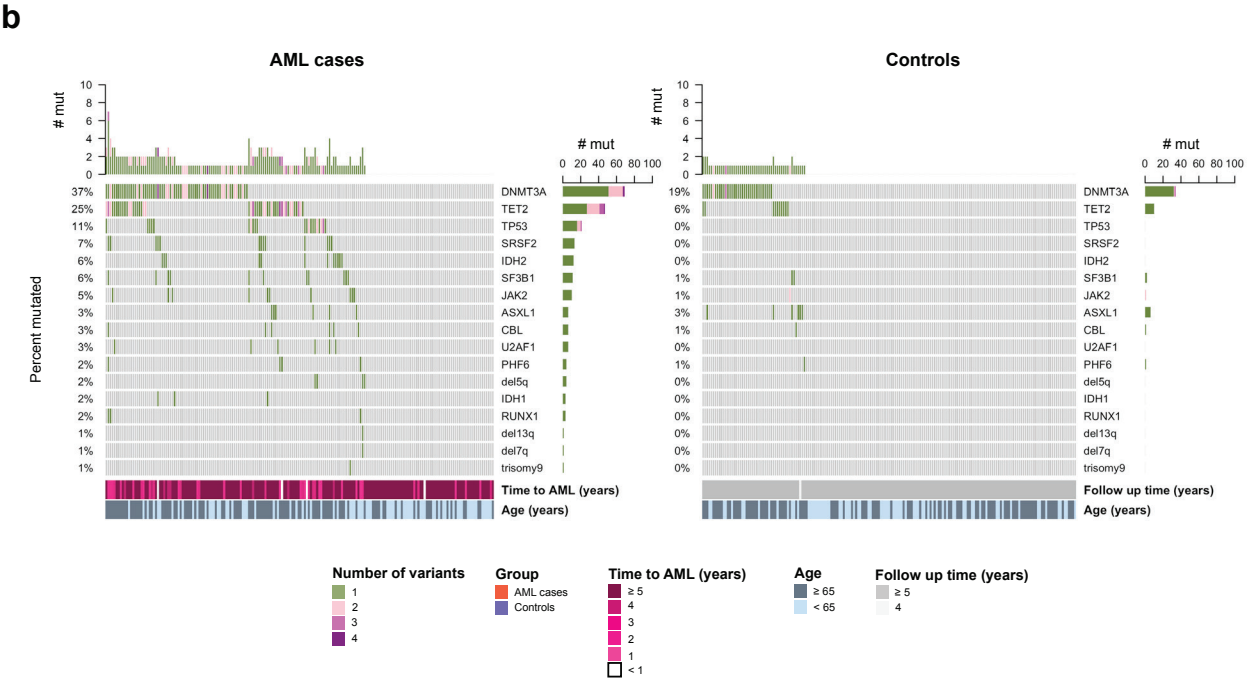
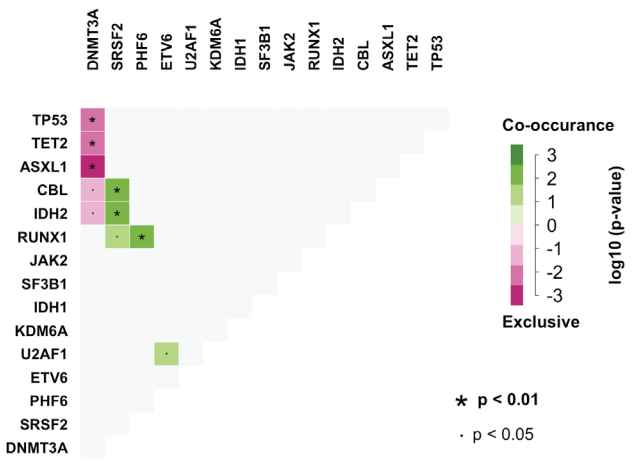


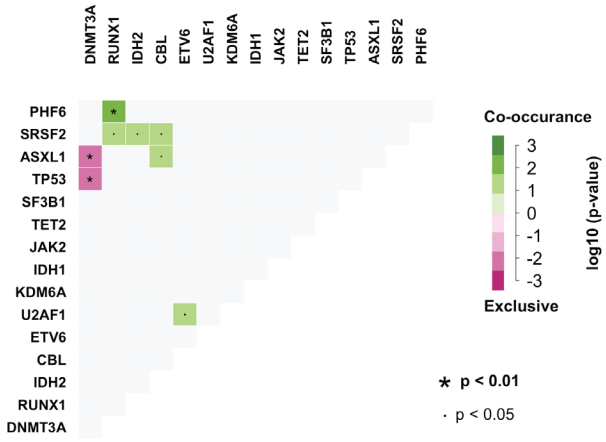
Figure S3. Co-occurrence and mutual exclusivity patterns

Tendency of co-occurrence and mutual exclusivity in mutated genes are shown for the (a) overall cohort, (b) AML case group, and (c) control group. (d) Table showing P values each significant gene pair association as well as tendency toward mutually exclusivity or co-occurrence. Pairwise significance of associations are determined using the *maftools* R package (* $p < 0.01$; • $p < 0.05$; two-sided Fisher's Exact Test). Mutations present >=3 times per group were considered. n = number mutated

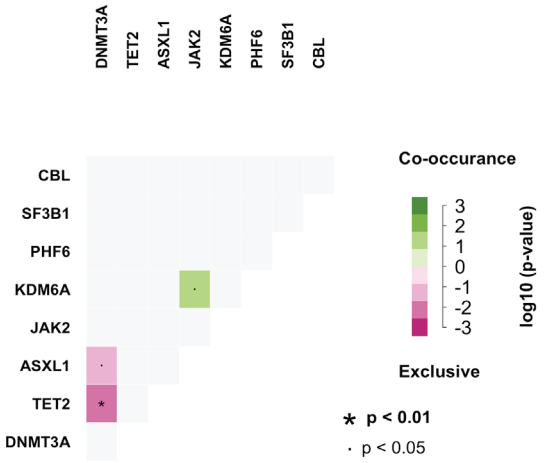
a



b



c



d

	Gene 1	Gene 2	Pvalue	Event
Overall	DNMT3A	ASXL1	0.0006	Mutually exclusive
	PHF6	RUNX1	0.0018	Co-occurrence
	TP53	DNMT3A	0.0021	Mutually exclusive
	TET2	DNMT3A	0.0022	Mutually exclusive
	IDH2	SRSF2	0.0055	Co-occurrence
	CBL	SRSF2	0.0083	Co-occurrence
	SRSF2	RUNX1	0.0133	Co-occurrence
	ETV6	U2AF1	0.0326	Co-occurrence
	DNMT3A	IDH2	0.0344	Mutually exclusive
	CBL	DNMT3A	0.0449	Mutually exclusive
Cases	RUNX1	PHF6	0.0022	Co-occurrence
	TP53	DNMT3A	0.0035	Mutually exclusive
	DNMT3A	ASXL1	0.0083	Mutually exclusive
	CBL	SRSF2	0.0139	Co-occurrence
	IDH2	SRSF2	0.0204	Co-occurrence
	CBL	ASXL1	0.0254	Co-occurrence
	RUNX1	SRSF2	0.0271	Co-occurrence
	ETV6	U2AF1	0.0469	Co-occurrence
Controls	TET2	DNMT3A	0.0092	Mutually exclusive
	ASXL1	DNMT3A	0.0299	Mutually exclusive
	JAK2	KDM6A	0.0357	Co-occurrence

Figure S4. Clonal hematopoiesis rate in the control group when applying different variant cutoffs and pathogenicity classification criteria.

(a) Percent of control participants exhibiting clonal hematopoiesis when defined as (left to right): any variant present at VAF > 1% irrespective of probable driver status (Desai, VAF > 0.01); any variant present irrespective of probable driver status using higher VAF cutoffs of 3.5% for SNVs and 7% for indels that were used by previous low depth whole exome sequencing studies [62] (Desai $VAF_{SNV} > 0.035$, $VAF_{indel} > 0.07$); pathogenic variants classified as in Supplemental Methods present at >1% VAF (Desai Pathogenic VAF > 0.01); pathogenic variants classified as in Supplemental Methods using higher VAF cutoffs of >3.5% for SNVs and >7% for indels (Desai Pathogenic $VAF_{SNV} > 0.035$, $VAF_{indel} > 0.07$); any variant present at VAF > 1% passing the pathogenicity criteria of Jaiswal and colleagues (Jaiswal VAF > 0.01)[62]; any variant passing both the pathogenicity criteria and VAF cutoffs used by Jaiswal and colleagues (Jaiswal $VAF_{SNV} > 0.035$, $VAF_{indel} > 0.07$)[62]. (b) Same classification and cutoff criteria as in (a) are applied stratified by age group (< 65 years; top) vs. (≥ 65 years; bottom). Each bar represents the fraction of mutated participants (black) vs. non-mutated participants (white). Percentages are shown.

VAF cutoffs have a greater impact on incidence of mutations than the variant classification criteria.

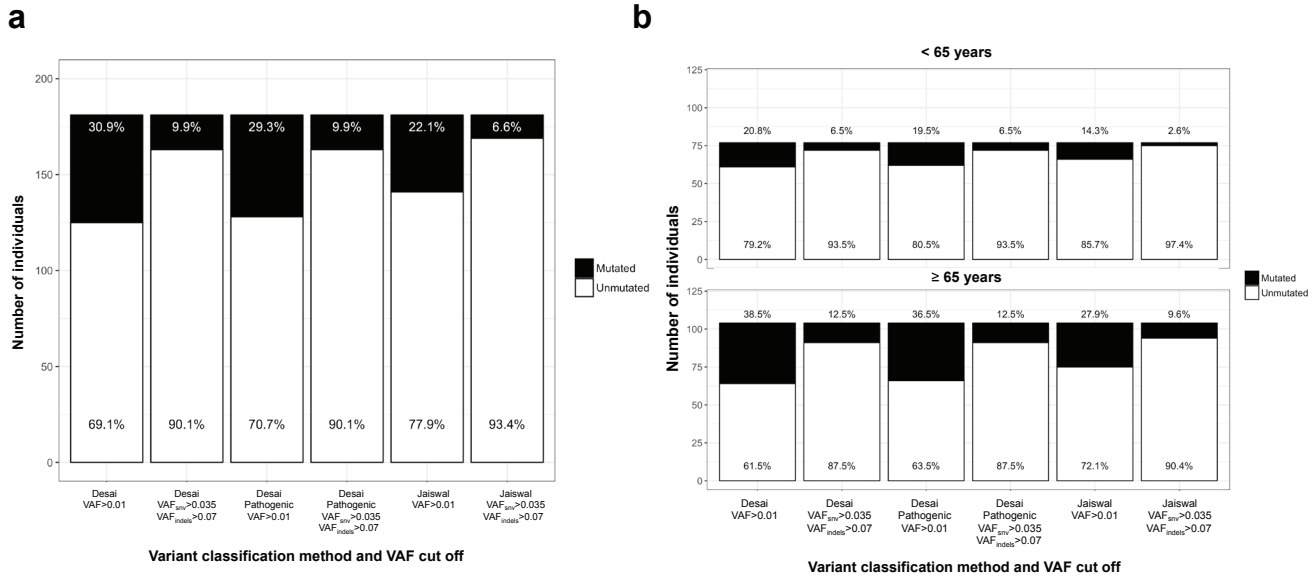


Figure S5. Mutation variant analysis summary in the AML case group.

A total of 289 variants were identified in AML case group (n = 129). Missense mutations accounted for 63% (181 mutations) followed by deletions (11.7%; 34 mutations), insertions (6.9%; 20 mutations), and 1.7% CNVs (5 total found exclusively in AML cases, denoted as "Complex" or "Other").

(a) Top row: bar plots enumerate variant classification (left); variant types (middle); SNV substitution (right). SNV, single nucleotide variant; INS, insertion; DEL, deletion. Bottom row: number of variants per sample with colors indicating variant classification (left); variant classification summary; and frequently mutated genes (right). N indicates number of mutations per sample. Colors indicate variant classifications. (b) Top row: boxplots indicating overall distribution of six different types of conversions (left) or transitions (Ti) and transversions (Tv) (right). Bottom row: Stacked bar plot shows the fraction of conversions per sample (bottom panel). Box and whiskers plots: box indicates the 1st quartile, median, and 3rd quartile whereas whiskers represent 1.5x the interquartile range. Plots are generated using *maftools*.

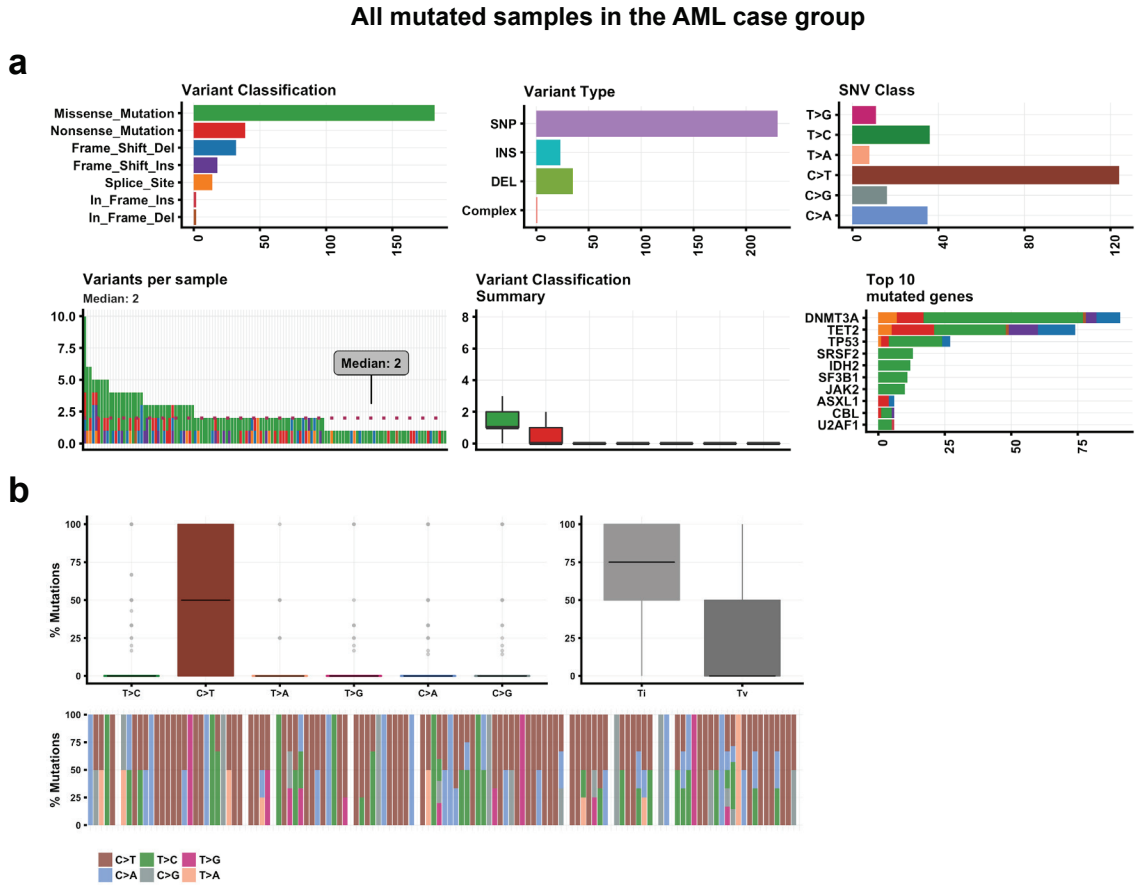


Figure S6. Mutation variant analysis summary in the control group.

A total of 68 variants were identified in the control group (n = 56). Missense mutations accounted for 54.4% of mutations (37 mutations total), followed by deletions (16.1%; 11 mutations), insertions (7.3%; 5 mutations).

As with the AML cases, the most common single-nucleotide change was a cytosine-to-thymine (C>T) transition occurring in CpG context for both groups (supplemental figures 3,4,5), a lesion and context associated with age-related mutagenesis[64] and consistent with other reports[62].

(a) Top row: bar plots enumerate variant classification (left); variant types (middle); SNV substitution (right). SNV, single nucleotide variant; INS, insertion; DEL, deletion. Bottom row: number of variants per sample with colors indicating variant classification (left); variant classification summary; and frequently mutated genes (right). N indicates number of mutations per sample. Colors indicate variant classifications. (b) Top row: boxplots indicating overall distribution of six different types of conversions (left) or transitions (Ti) and transversions (Tv) (right). Bottom row: Stacked bar plot shows the fraction of conversions per sample (bottom panel). Box and whiskers plots: box indicates the 1st quartile, median, and 3rd quartile whereas whiskers represent 1.5x the interquartile range. Plot is generated using *maftools*.

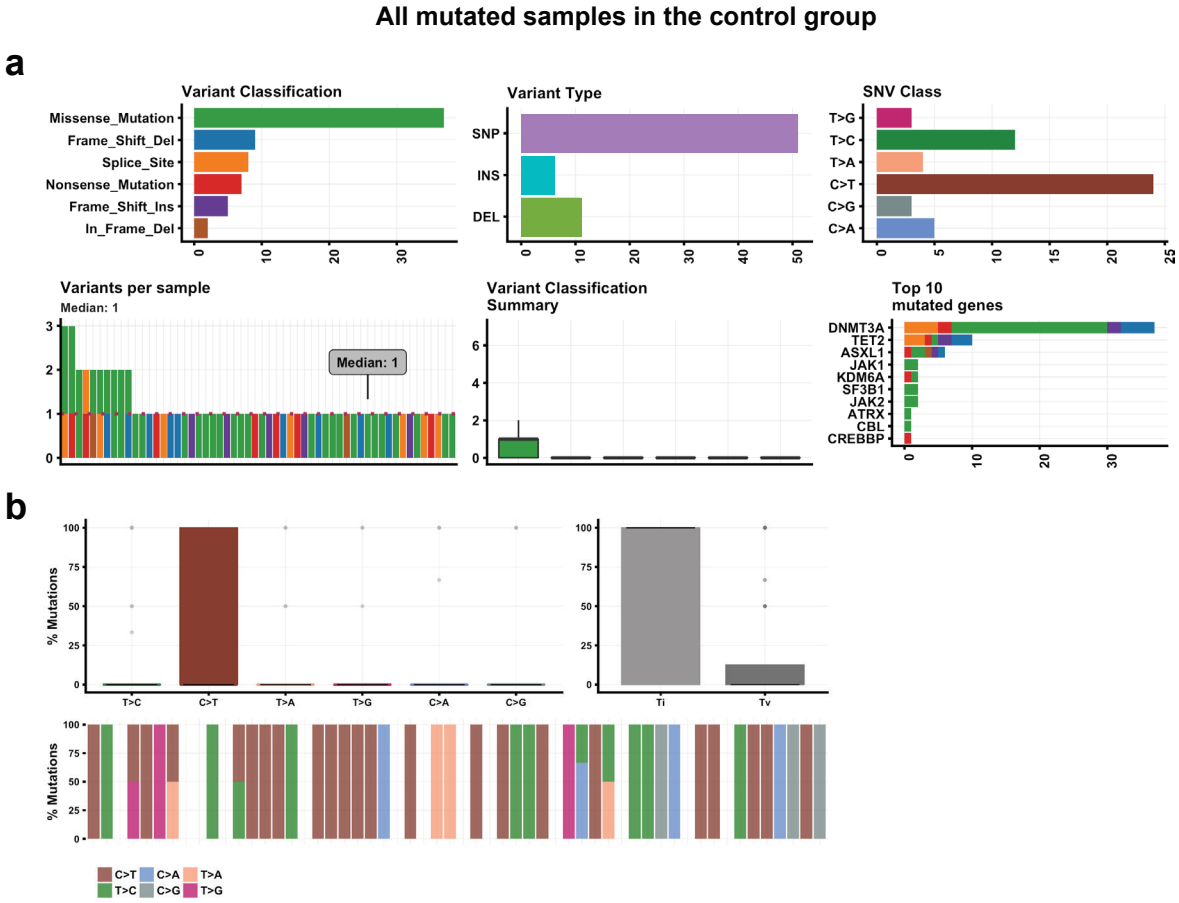


Figure S7. Distribution of point mutation types in the AML and control group.

Relative contribution of base substitutions when focusing on missense mutations enumerated in the AML case group (n = 181; Figure S1) and the control group (n = 37; Figure S2). The majority of C>T transitions occur in CpG context for both groups, suggesting acquisition from missrepair deamination of 5-methylcytosine. The most common single-nucleotide change was a cytosine-to-thymine (C>T) transition occurring in CpG context for both groups, a lesion and context associated with age-related mutagenesis[64] and consistent with other findings[62].

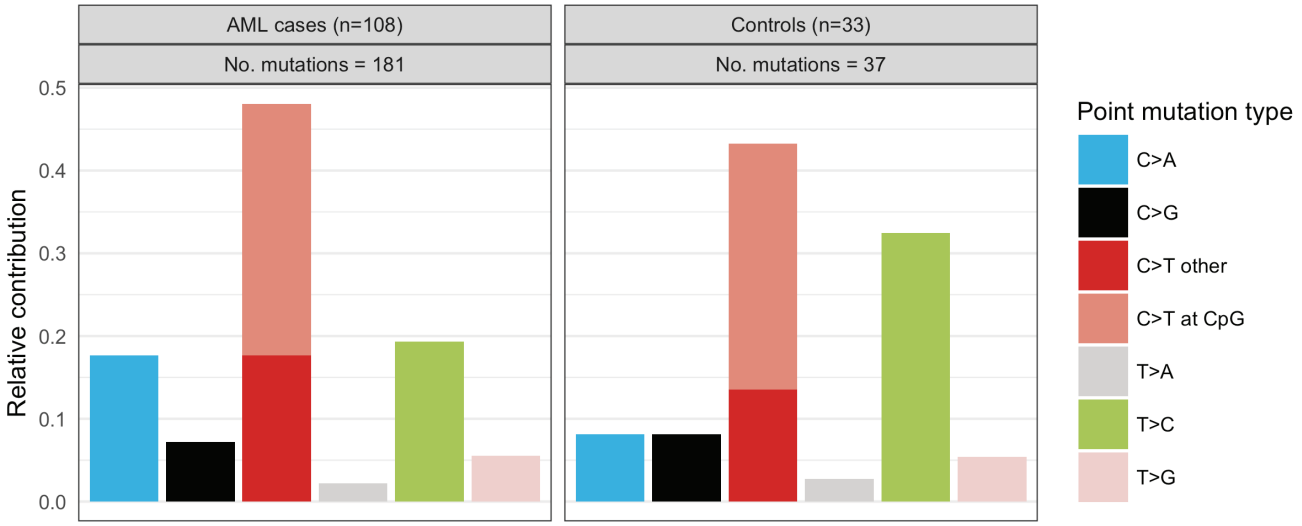


Figure S8. SNV signature analysis and transcribed strand bias in the case and control group.

A broader analysis of base substitution was performed taking into account the bases immediately upstream and downstream of the mutated base providing mutation context information (panel A below). An elevated rate of spontaneous deamination of 5-methyl-cytosine occurring predominantly NpCpG trinucleotide is consistent with reports of mutation patterns in AML[65]. Of note, a similar pattern is observed for AML cases and controls, which suggests common underlying mutation processes in the two groups most likely driven by aging[66]. Further analysis of the mutation pattern shows preference for certain substitutions in the transcribed strand over the untranscribed strand suggesting additional mutational processes driven by transcription-coupled repair (TCR) [67], a nucleotide excision repair (NER) process that has been shown to decrease in efficiency with normal aging[68]. While strand bias suggesting TCR approached significance ($P < 0.05$) for the AML group. The control group demonstrated the same trend toward strand bias also suggesting TCR but did approach significance because of lower number of cases.

(a) Trinucleotide context of C>A, C>G, C>T, T>A, T>C, and T>G point mutations is shown among non-synonymous point mutations in the AML case and control group. For each context, the stacked bar chart indicates mutations occurring on the untranscribed strand (filled with color) vs the transcribed strand (not filled). (b) Transcribed vs. untranscribed strand bias (log2 scale) is indicated for each type of point mutation type in the AML case (n = 181 variants, n = 108 participants) and control group (n = 37 variants, n = 33 participants). Significant differences are indicated by asterisks (*) using the Poisson test. * $P < 0.05$ as implemented in the *MutationalPatterns* R package.

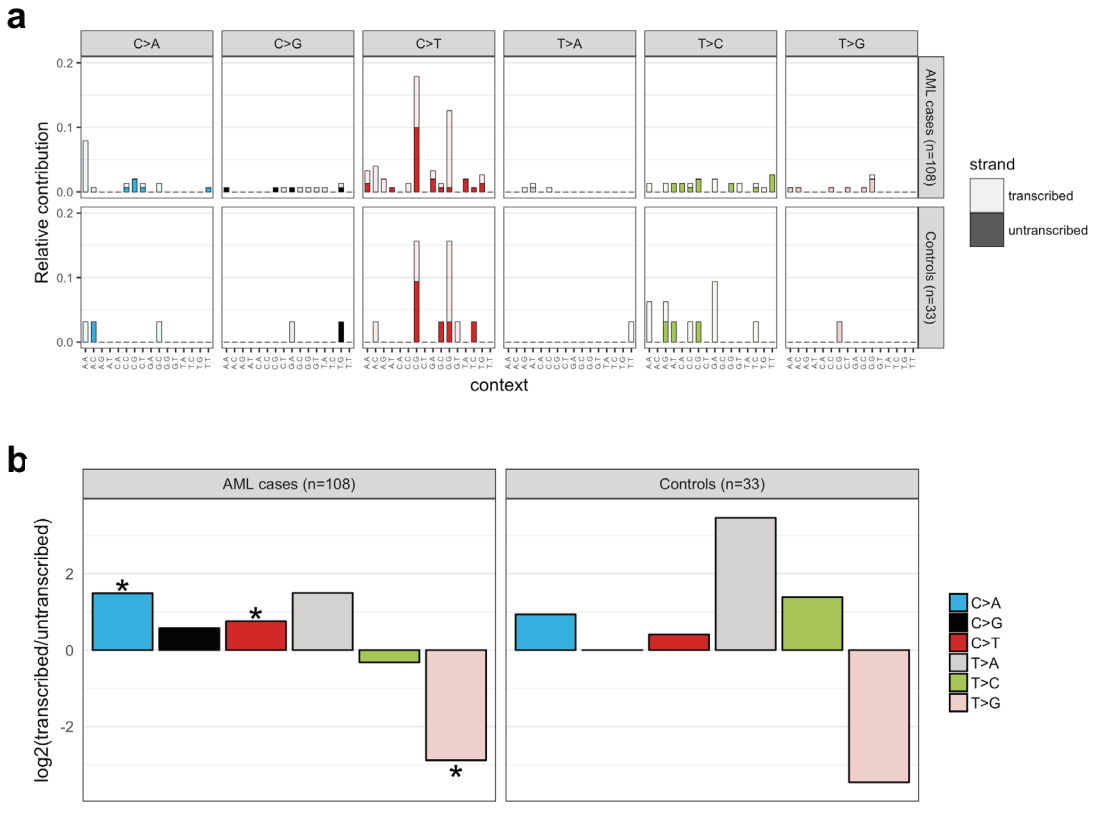


Figure S9. VAF cutoff tables for specificity and sensitivity

Analysis of sensitivity and specificity showed that the false positive rate for individuals bearing mutations in *TP53*, *SRSF2*, *IDH2*, *SF3B1* or *U2AF1* in the 1-2% VAF range is less than 1%.

Table of VAF cut offs for the significantly mutated genes producing no greater than a (a) 1% false positive rate, (b) 5% false positive rate, or (c) 10% false positive rate while maximizing sensitivity. False positives represent controls misclassified as AML cases.

a

< 1% false positive rate

Gene	VAF cutoff	True positive rate	False positive rate
DNMT3A	34.09	6.2	0
TET2	20.11	13.18	0
SRSF2	1.29	10.08	0
IDH2	1.18	9.3	0
TP53	1.02	16.28	0
SF3B1	1.95	7.75	0
U2AF1	1.79	4.65	0
Any of the above genes	31.21	11.63	0
Any of the above genes except DNMT3A	20.11	22.48	0

b

< 5% false positive rate

Gene	VAF cutoff	True positive rate	False positive rate
DNMT3A	21.65	10.08	3.57
TET2	5.84	19.38	3.57
SRSF2	1.29	10.08	0
IDH2	1.18	9.3	0
TP53	1.02	16.28	0
SF3B1	1.49	8.53	1.79
U2AF1	1.79	4.65	0
Any of the above genes	20.11	27.91	3.57
Any of the above genes except DNMT3A	5.84	37.98	3.57

c

< 10% false positive rate

Gene	VAF cutoff	True positive rate	False positive rate
DNMT3A	8.88	22.48	8.93
TET2	1.37	31.78	8.93
SRSF2	1.29	10.08	0
IDH2	1.18	9.3	0
TP53	1.02	16.28	0
SF3B1	1.49	8.53	1.79
U2AF1	1.79	4.65	0
Any of the above genes	9.66	44.96	8.93
Any of the above genes except DNMT3A	1.47	55.04	8.93

Figure S10. Variation in VAF over time in recurrently mutated genes.

Mutations in recurrently mutated genes in patients with serial samples were tracked over time. Generally, no changes were observed within the first year in AML cases or controls. At 3 years, however, the AML cases demonstrated elevations in VAF. In contrast, the VAF in the controls group remained mostly stable up to 3 years. Mutations in *IDH2, SRSF2, U2AF1, TP53 and TET2* showed significant increase in mean VAF between baseline and year 3 follow up.

Comparison of allelic fractions per mutated gene at baseline (horizontal axis) and (a) after 3 years of follow up (vertical axis) or (b) after 1 year of follow up (vertical axis) irrespective of whether mutation is present at baseline evaluation. The diagonal represents no change in VAF. (c) Forest plot indicating linear model estimates of mean allelic fraction changes when a mutation is present at baseline evaluation for specific genes: *IDH2* ($P = 1.9 \times 10^{-5}$, n=5); *SRSF2* ($P = 6.4 \times 10^{-4}$, n = 3); *U2AF1* ($P = 1.0 \times 10^{-3}$, n = 2); *TP53* ($P = 2.5 \times 10^{-2}$, n = 6); *TET2* ($P = 2.4 \times 10^{-2}$, n = 20); *IDH1* (P = 0.140, n = 2); *JAK2* (P = 0.120, n = 8); *ASXL1* (P = 0.294, n = 2); *SF3B1* (P = 0.227, n = 3); *DNMT3A* (n = 38, P = 0.169). * P <0.05, **P < 0.01, *** P<0.001; n, number of pairwise evaluations.

The diagonal represents no change in VAF. Individual AML cases mutated in the gene indicated are indicated with red dot (AML case) or blue dot (control). The maximum VAF was selected when participants harbored more than 1 mutation.

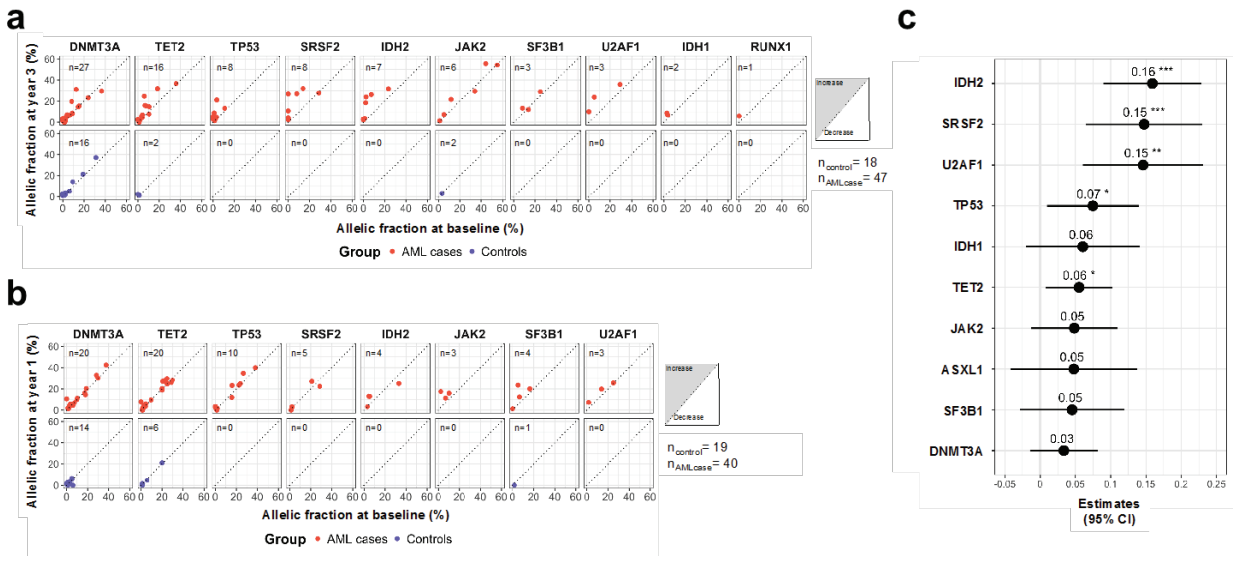


Figure S11. Persistence of mutations detected at baseline in longitudinally evaluated participants.

Heatmap indicating persistence of mutations in evaluable cases and controls with serial samples available at baseline and 1-year or 3-year follow-up. Blue to purple color gradient in heatmap indicates VAF ranging from 0% to 40%. >95% of variants present at baseline VAF > 1% are stably maintained at year 1 or year 3 (N=213/224 serially evaluable variants in 121 individuals). SNP signatures were verified in all longitudinally monitored participants and did not explain the non-persistent variants. Mutations were force-called down to 0.5% VAF to detect persistence of variants near the 1% VAF cutoff. Non-canonical variants in genes such as *FLT1* and *CARD11* are maintained similarly as canonical drivers such as *IDH2*, *TET2*, *TP53*, and *SRSF2*. Variants present at lower VAF demonstrated a tendency to drop out.

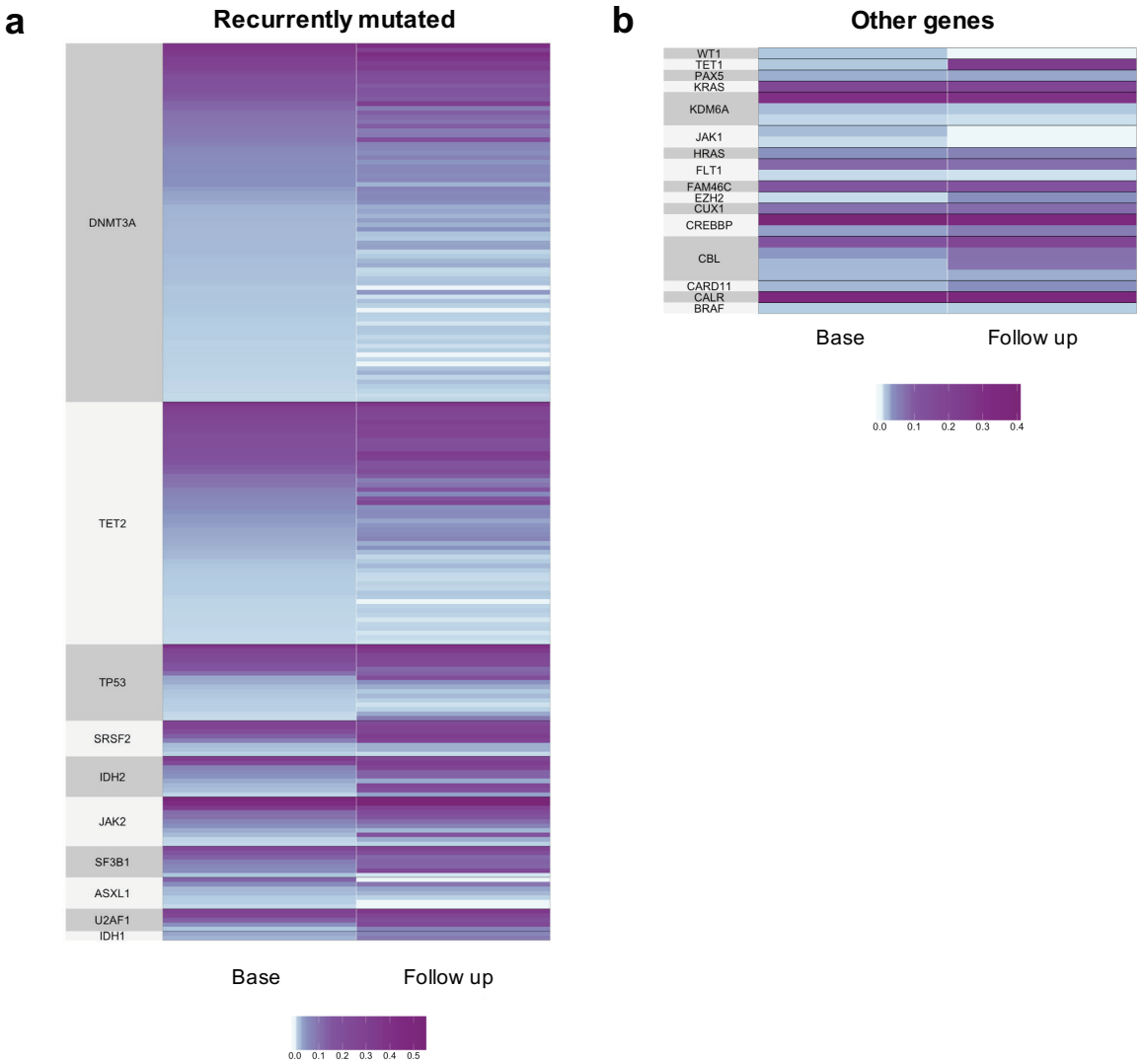
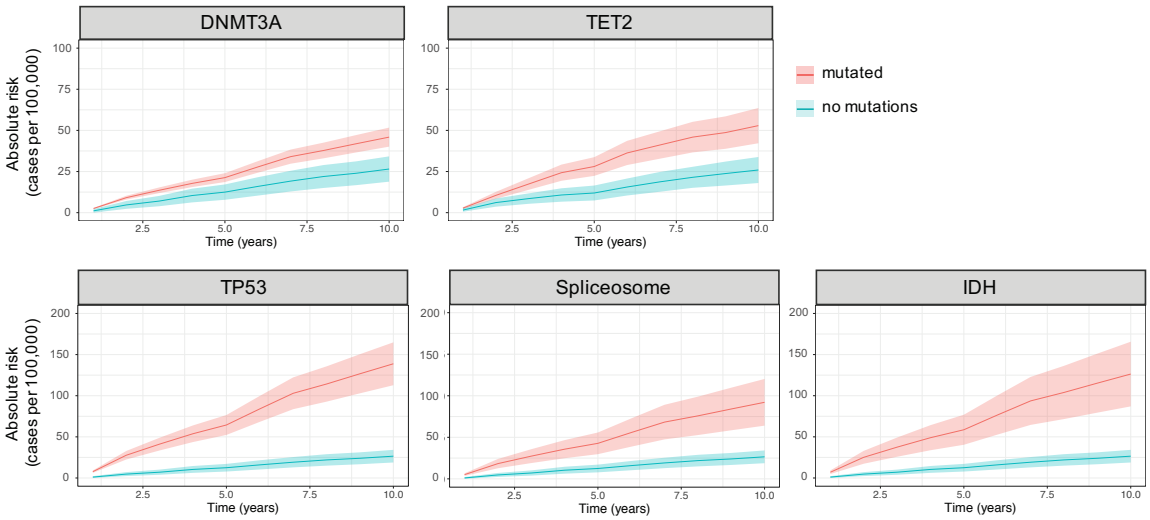


Figure S12. Absolute risk estimation for recurrently mutated genes.

Odds ratios from comparison of mutations in cases and controls are converted into estimated absolute risk estimates using a weighted partial likelihood method and weighted baseline hazard approach [38] described in Supplemental Methods. (a) Analysis shows estimated absolute risk over a 10-year period using demographic data from the WHI cohort for individual mutated in the genes indicated (mutated) alongside mutation-free participants (no mutations) (n = 59). Shaded region indicates the standard error (N=200 bootstrap estimates). Baseline incidence per year is 2.6 AML cases per 100,000 women in the absence of mutations. *DNMT3A* (n = 103), *TET2* (n = 57), *TP53* (n = 21), spliceosome (n = 28) and IDH genes (n= 15) (b) Incidence rate values per gene as well as for mutation-free participants expressed as cases per 100,000 persons per year.

a



b

Mutation	Incidence rate (per 100,000 persons per year)
No mutations	2.650
DNMT3A	4.591
TET2	5.462
TP53	13.884
Spliceosome	9.213
IDH	12.634

Figure S13. Mapping of coding alterations to protein domains – Part 1

Mapping of coding alterations to protein domains for recurrently mutated genes in the cohort: *DNMT3A, TET2, TP53, SRSF2, IDH2, JAK2, SF3B1, U2AF1, ASXL1, IDH1* and *RUNX1*. For each gene, mutations identified in the AML case and control groups are plotted except for genes that were not mutated in the control group: *TP53, U2AF1* and *RUNX1*. Mutations are classified as Missense (green), In-frame (brown) and Truncating (grey).

Mutations in *IDH2, SRSF2, JAK2, SF3B1* and *U2AF1* occurred in positions R140, P95, V617, K700 and Q157, respectively. Other point mutations were detected in the HEAT domain of *SF3B1* in close proximity to K700 or in the zinc finger domain of *U2AF1* in close proximity to Q157. All of these positions are known hotspots for the aforementioned genes and highly associated with hematological malignancies and especially AML[69].

Mutations in *DNMT3A* were localized in exons 8-23. 83% of the variants detected in the gene corresponded to SNV with missense mutations in the R882 position accounting for 26% of the SNVs. The second most common alteration were truncating variants affecting all the functional domains. SNVs demonstrated an overall tendency to occur in functional domains whereas truncating mutations occurred in the N-terminal half of the protein.

Missense mutations comprised 63% of the observed mutations in *TET2* with the majority of the missense SNVs being confined to oxygenase domain of *TET2* (Tet2_JDP). Truncating mutations were distributed across coding exons.

Mutations in *ASXL1* were predominantly found in exon 13, with the most common type of alteration being non-sense SNV. Truncations in the carboxy-terminus or premature stop variants have a disrupting effect whereas missense variants have an unknown significance[69].

Mutations in *TP53* were also found distributed along the gene. SNVs were concentrated in the DNA-binding domain with additional variations in tetramerization and transactivating domains. Truncating mutations occurred throughout. The vast majority of mutations were missense SNV (75%) including known hotspots with few frameshift deletions, consistent with the mutation pattern observed for mutations in *TP53* in human cancer [70].

A complete list of mutations is available as a supplemental spreadsheet.

Figure S13. Mapping of coding alterations to protein domains – Part 2

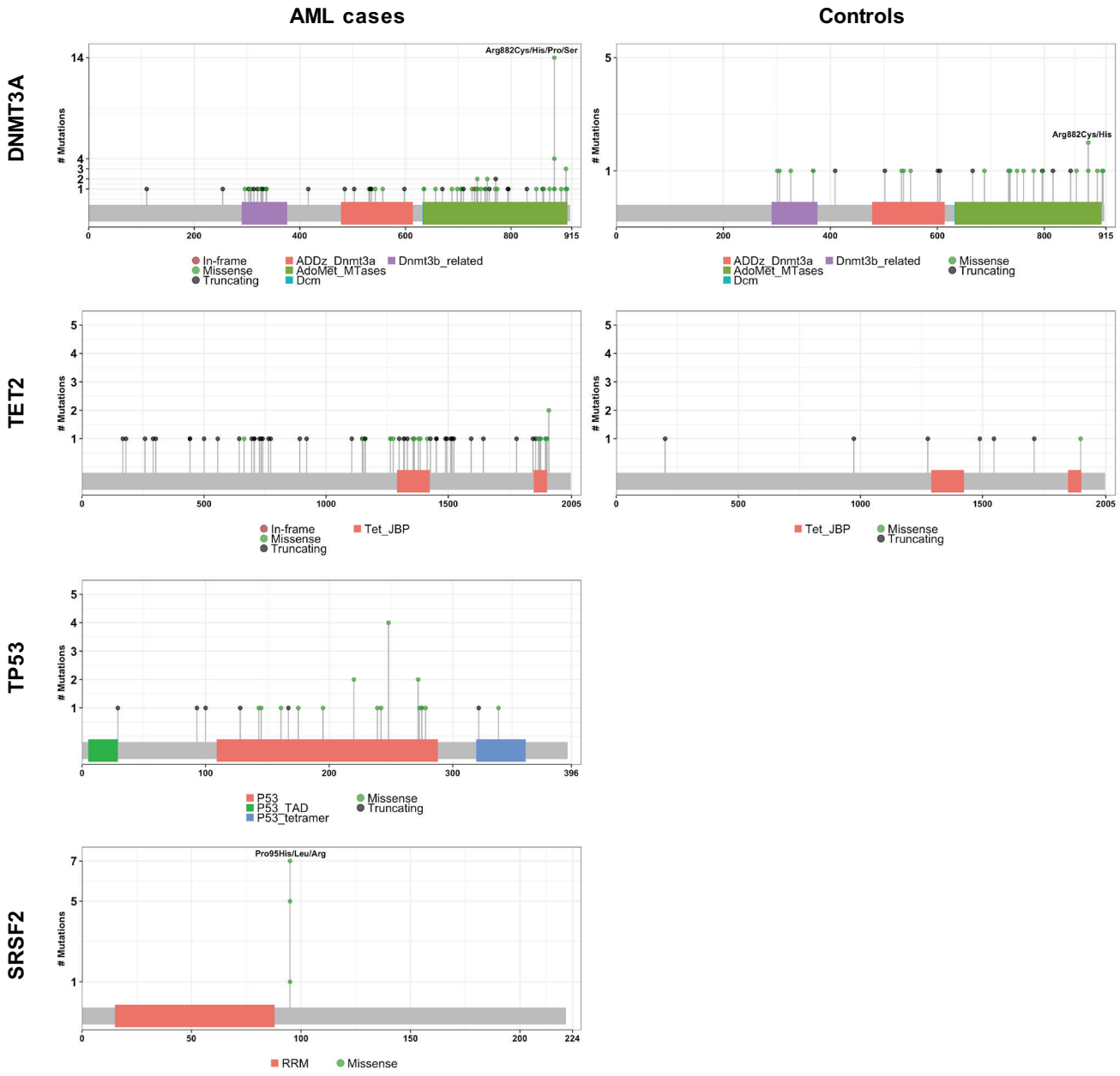


Figure S13. Mapping of coding alterations to protein domains – Part 3

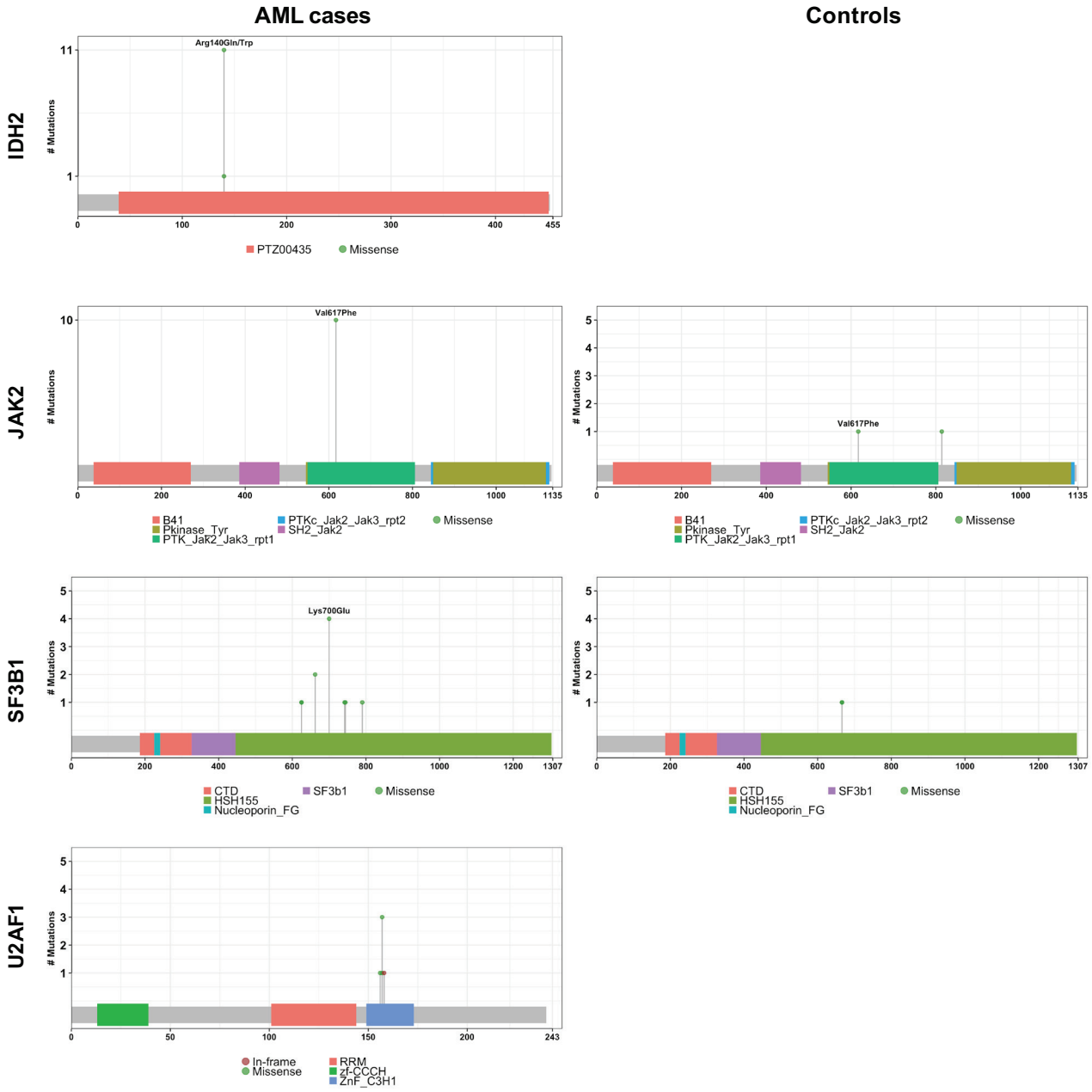


Figure S13. Mapping of coding alterations to protein domains – Part 4

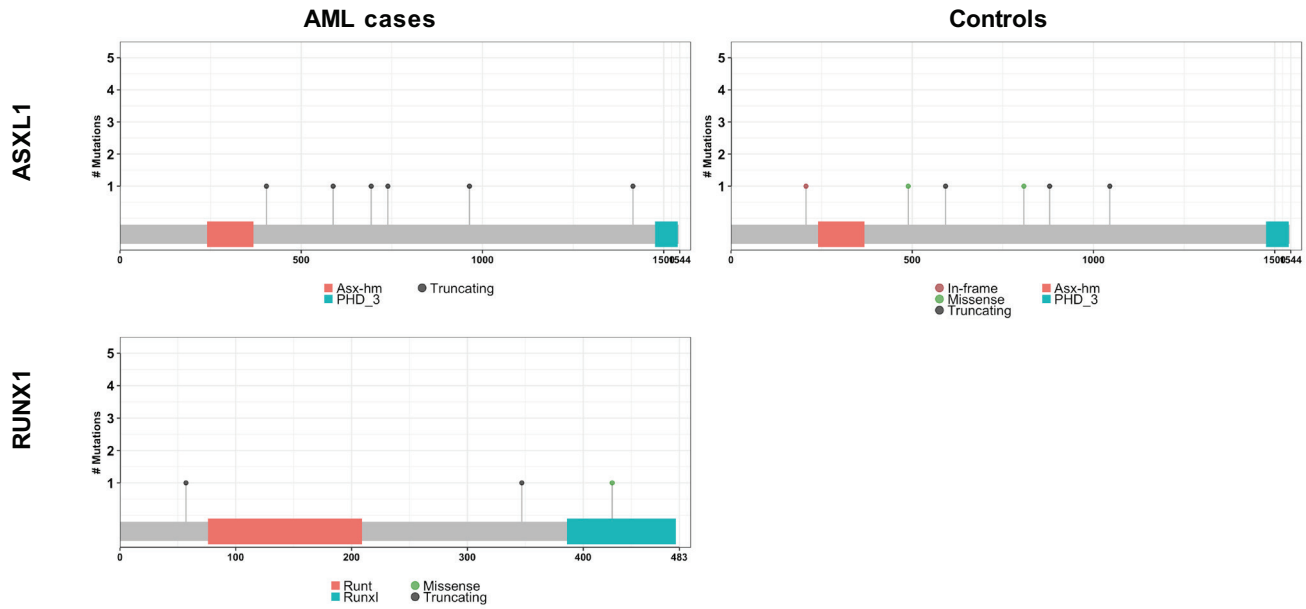


Figure S14. Spatial clustering of TP53 point mutations in 3-D structure.

Despite having distant amino acid positions, mutations within *TP53* across the AML cases were spatially localized when mapped to the tertiary structure of the protein. Positions of SNVs are indicated by spheres. Clustering was performed using mutation3D[61]. Two distinct clusters were identified each encompassing known regions involved in DNA contact (e.g. R248 and R273) and structural support (e.g. R175) [71].

Cluster 1 (red spheres) indicates alterations in amino acid positions 239, 242, 248 (4 participants), 272 (2 participants), 273 ,275 (2 participants), and 278 (P = 0.02, non-parametric bootstrap). Cluster 2 (blue spheres) indicates mutations in positions 161, 175, 195, and 234 (P = 0.06, non-parametric bootstrap). PDB accession number is 2XWR.

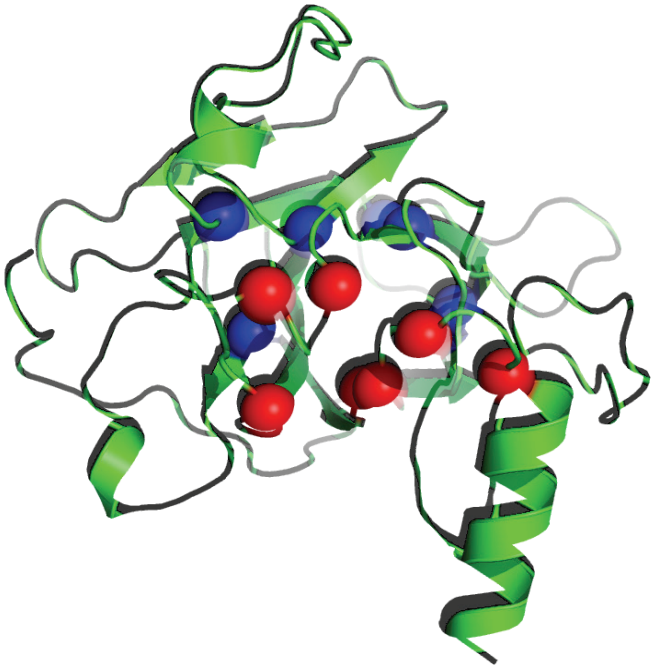


Figure S15. Distribution of mutations in protein domains (Pfam) in the AML case and control group.

Number of genes (nGenes; vertical axis) is plotted vs. number of mutations (nMuts; horizontal axis). AML case samples: Top mutated Pfam domains include PTZ00435 (isocitrate dehydrogenase *IDH1* Arg132 and *IDH2* Arg140), AdoMet_MTases (AdoMet methytransferase; *DNMT3A*), and P53 (DNA binding domain, *TP53*). Control samples: Top mutated Pfam domains include Dcm (DNA cytosine methyltransferase; *DNMT3A*), and Pkinase_Tyr (tyrosine kinase domain; *JAK1* Lys696 and *JAK2* Val617).

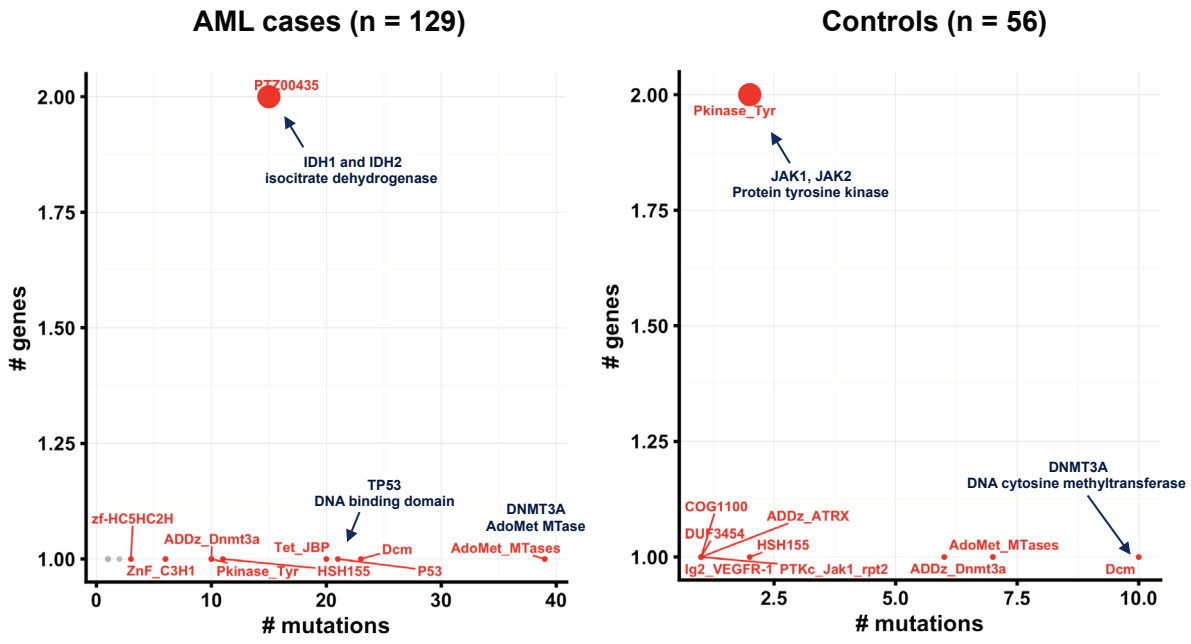


Figure S16. Mutation frequencies and association of probable pathogenic somatic variants with AML.

(a) Number and frequency of pathogenic mutations in AML cases vs. controls overall and for participants younger than 65 years vs. ≥ 65 years. (b) Forest plot indicating odds ratio of pathogenic mutations in each gene occurring in the AML cases vs. controls. Genes or gene categories significantly associated with AML include *TP53* ($P = 7.9 \times 10^{-6}$), *IDH* ($P = 2.6 \times 10^{-4}$), spliceosome, *TET2* ($P = 5.1 \times 10^{-6}$), and DNMT3A ($P = 4.6 \times 10^{-4}$). (c) Forest plot indicating odds of developing AML within 5 years from baseline, depicted as odds ratios for the specific pathogenic mutations. Mutations in TP53 and DNMT3A are significantly associated with rapid development of AML. (d) Forest plot indicating odds ratio of pathogenic mutations in each gene occurring in the AML cases vs. controls adjusted for presence of mutations in other genes (Others). $P < 0.001$: TET2 ($P = 4.8 \times 10^{-5}$) and DNMT3A ($P = 2.9 \times 10^{-4}$). OR per gene are adjusted by age (years) as a continuous variable. IDH category includes IDH1 and IDH2. The spliceosome category includes SRSF2, SF3B1, and U2AF1. Abbreviations: CI, confidence interval; N, number affected. P-values are shown for penalized likelihood multivariable logistic regression.

All participants bearing mutations in significant genes presented mutations classified as pathogenic, with the exception of a DNMT3A (see Figure 1A and Figure 1B).

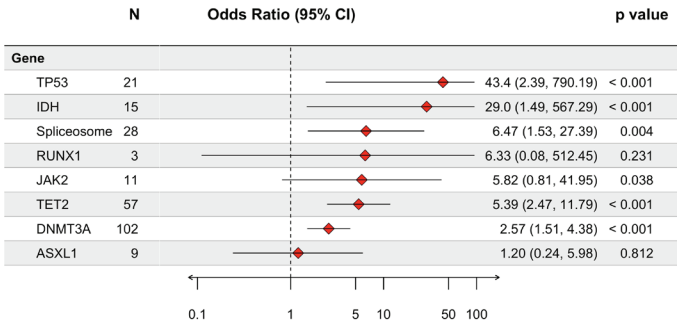
Figure S16. Mutation frequencies and association of probable pathogenic somatic variants with AML – continued.

a

	AML cases (N=188)		Controls (N=181)		Odds Ratio	
Age	# Mutated (%)	# Non-mutated (%)	# Mutated (%)	# Non-mutated (%)	OR (95% CI)	P value
< 65	42 (52.5)	38 (47.5)	15 (19.48)	62 (80.52)	4.52 (2.12-10.05)	2.7×10^{-5}
≥ 65	85 (78.7)	23 (21.3)	38 (36.54)	66 (63.46)	6.35 (3.35-12.4)	2.5×10^{-11}
Total	127 (67.55)	61 (32.45)	53 (29.28)	128 (70.72)	5.00 (3.16-8.02)	1.1×10^{-14}

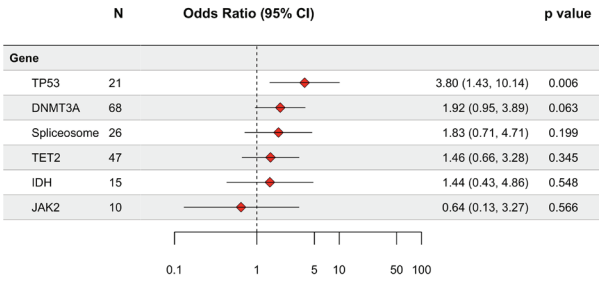
b

Odds ratio: AML cases vs. controls



c

Odds ratio: AML diagnosis < 5 years from baseline



d

Odds ratio: AML cases vs. controls, pathogenic variants with others

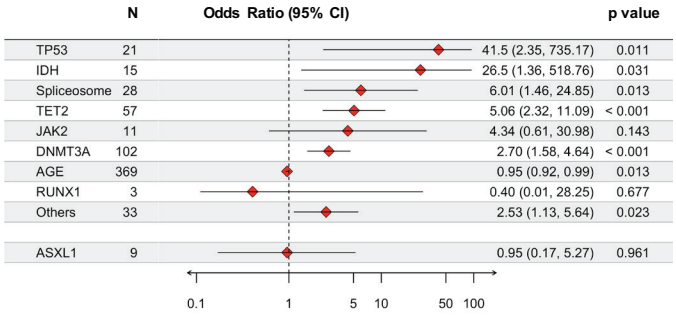


Figure S17. Odds of AML adjusted for the presence of additional mutations.

(a) Forest plot indicating odds of developing AML within 5 years from baseline, depicted as odds ratios for mutations in significant genes accounting for presence of 1 or multiple mutations in *DNMT3A* and *TET2*. Mutations in *TP53* are significantly associated with rapid development of AML. The number of mutations per participant in *DNMT3A* or *TET2* does not present significant differences in the odds to develop AML. (b) Forest plot indicating odds ratio of mutations in each gene occurring in the AML cases vs. controls. Odds ratios are corrected by the presence of mutations in other genes not represented in Table 1b. P < 0.001: *TET2* ($P = 2.6 \times 10^{-5}$) and *DNMT3A* ($P = 2.6 \times 10^{-4}$). OR per gene are adjusted by age (years) as a continuous variable. IDH category includes *IDH1* and *IDH2*. The spliceosome category includes *SRSF2*, *SF3B1*, and *U2AF1*. Abbreviations: CI, confidence interval; N, number affected. P-values are shown for penalized likelihood multivariable logistic regression.

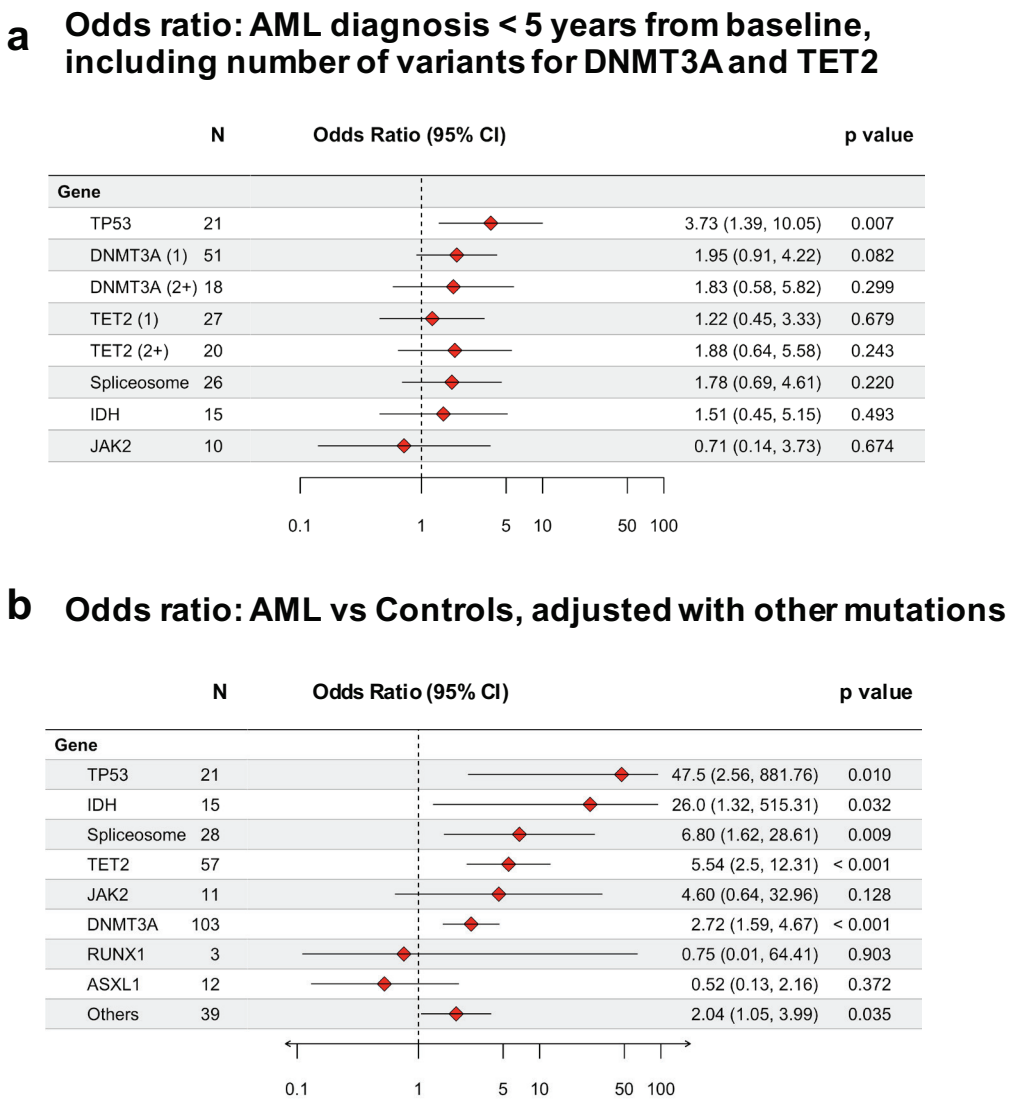
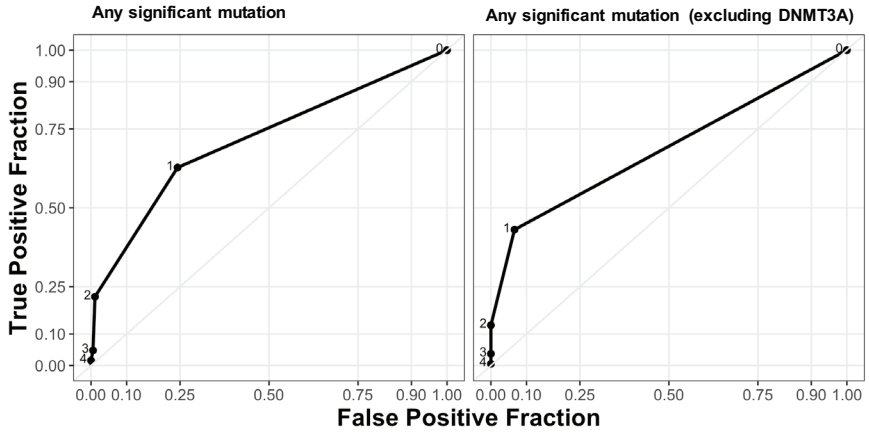


Figure S18. ROC analysis of number of mutations in significant high-risk genes.

Receiver operating characteristic (ROC) curves indicating the % true positive rate (vertical axis) vs. the % false positive rate (horizontal axis) of the number of mutations to detect AML cases. The curves indicate performance at decreasing number of variants per gene for (a) significant genes (left plot; *DNMT3A, TET2, IDH1, IDH2, SRSF2, SF3B1, U2AF1, TP53*; n = 164 [118 AML cases, 44 controls]) or the same set of genes excluding *DNMT3A*; n = 94 [81 AML cases, 12 controls] (right plot). (b) performance is shown for *DNMT3A* and *TET2* genes; left and right plot, respectively.

Presence of 2 or more mutations in selected genes was able to detect AML cases with less than 5% false positive fraction.

a



b

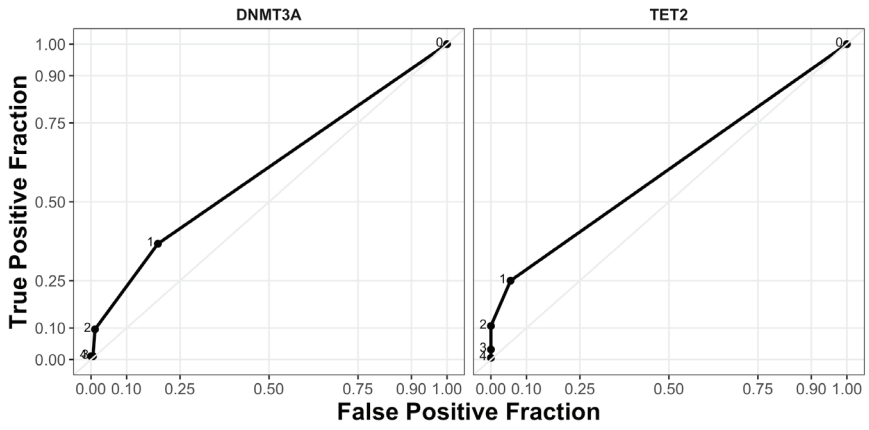
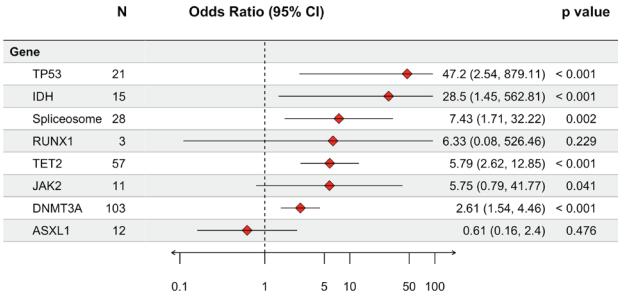


Figure S19. Associations between AML and mutations are robust to different variant classification methods and VAF cutoffs.

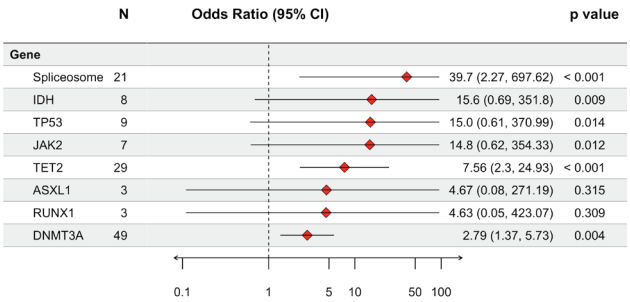
Forest plot indicating odds ratio of mutations per gene occurring in AML cases vs. controls applying different variant classification and VAF cutoff criteria. (a) Somatic variants the indicated genes at VAF > 1%. P < 0.001: *TP53* (P = 5.5 x 10^{-6}); *IDH* (P = 3.0 x 10^{-4}); *TET2* (P = 2.4 x 10^{-6}) and *DNMT3A* (P = 3.4 x 10^{-4}). (b) Somatic variants selected for the indicated genes at VAF > 3.5% for SNVs and VAF > 7% for indels. P < 0.001: spliceosome (P = 1.5 x 10^{-5}) and *TET2* (P = 9.8 x 10^{-5}) (c) Somatic variants selected according to the method of Jaiswal and colleagues [62] using a cutoff of VAF > 1%. P < 0.001: *TP53* (P = 4.8 x 10^{-6}); *IDH* (P = 5.0 x 10^{-4}); *TET2* (P = 5.0 x 10^{-4}) and *DNMT3A* (P = 7.5 x 10^{-5}) (d) Somatic variants selected according to the method of Jaiswal and colleagues [62] using a cutoff of VAF > 3.5% for SNVs and VAF > 7% for indels. P < 0.001: spliceosome (P = 9.8 x 10^{-6}) and DNMT3A (P = 9.5 x 10^{-4}). OR per gene are adjusted by age (years) as a continuous variable. *IDH* category includes *IDH1* and *IDH2*. The spliceosome category includes *SRSF2*, *SF3B1*, and *U2AF1*. Abbreviations: CI, confidence interval; N, number affected. P-values are shown for penalized likelihood multivariable logistic regression.

VAF cutoffs has a greater impact on the incidence of mutations than the variant classification criteria. See Table 1b, Table S4 and figure S4.

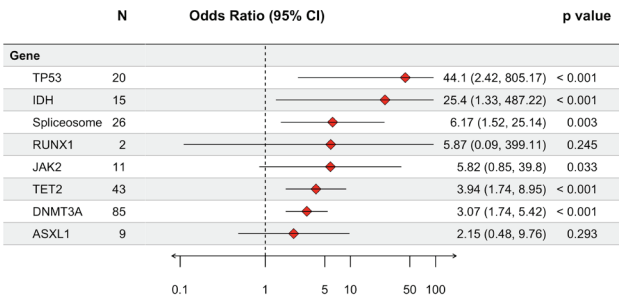
a Variants Desai VAF > 0.01



b VAF_{SNV} 0.035 & VAF_{indel} 0.07



c Variants in Jaiswal VAF > 0.01



d Variants in Jaiswal VAF_{SNV} 0.035 & VAF_{indel} 0.07

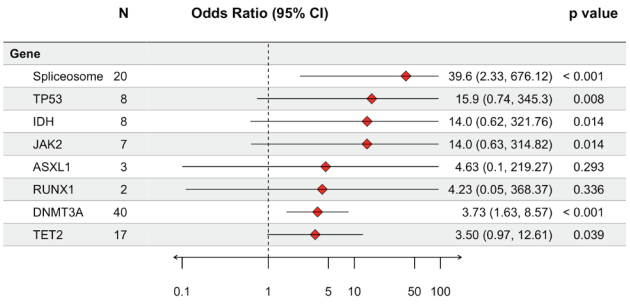


Figure S20. Association between number of somatic variants and clone size.

Significant positive correlation was found between the number of somatic variants and the clone size defined as the maximum VAF of all somatic mutations detected per participant ($\rho = 0.43$, $P = 1.5 \times 10^{-9}$, Spearman's correlation). Median, 1st quantile and 3rd quantile maximum clone VAF is shown for each number of variants (middle, lower and upper grey dots, respectively) for AML cases (n = 125) (red, left panel) and controls (n=56) (blue, right panel). AML cases: 1 (n = 43, VAF [0.01-0.55]), 2 (n = 46, VAF [0.01-0.41]), 3 (n = 18, VAF [0.02-0.37]), 4 (n = 12, VAF [0.06-0.38]) and 5 variants (n = 6, VAF [0.10-0.33]); Controls: 1 (n = 46, VAF [0.01-0.31]), 2 (n = 8, VAF [0.02-0.27]) and 3 variants (n = 2, VAF [0.03-0.07]). n, number of participants.

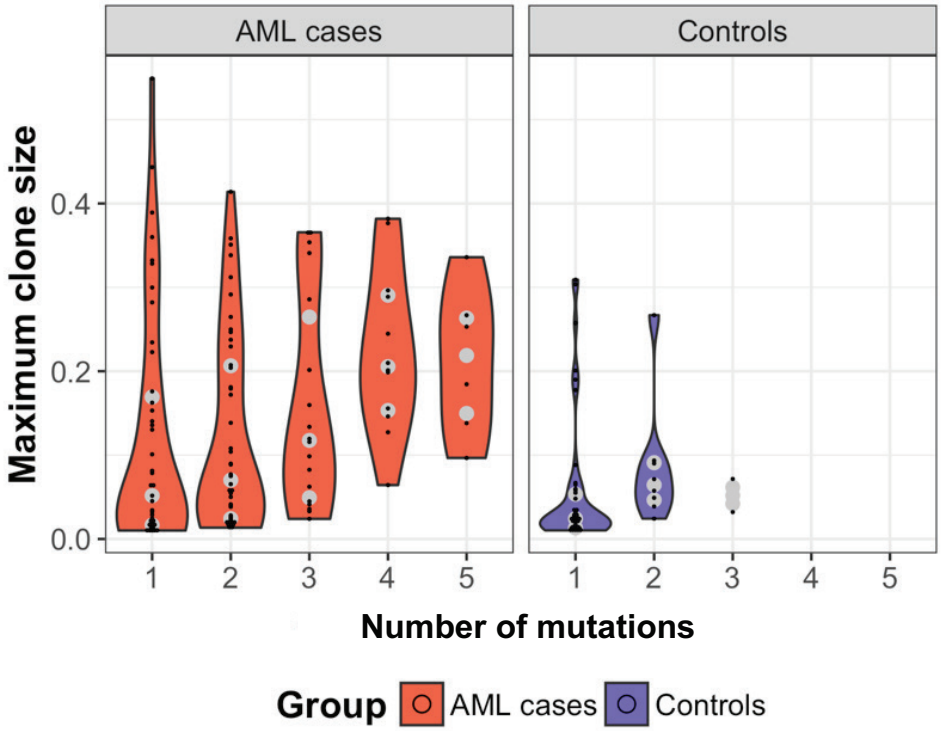


Figure S21. The pattern of subclonal somatic variations in cases and controls reveals similar driver genes as a de novo AML cohort.

Identification of driver genes was performed using the oncoDriveCLUST algorithm implemented in *maftools*[72]. The horizontal axis shows the fraction of mutations within clusters while the vertical axis indicates the –log10(false discovery rate). The FDR cutoff was set to 1%. Each red dot represents a probable driver gene informed by the variant call set. (a) Driver genes identified using all somatic variants at baseline evaluation for genes with >=5 mutations (N = 302 variants). (b) Driver genes identified using all somatic variants reported in the AML TCGA study for genes with >=5 mutations (N = 305 variants). The size of the dot is related to the number of clusters per gene with the number of clusters indicated by the number in brackets. Significant clusters were identified for *SRSF2, IDH2, JAK2, U2AF1, SF3B1, DNMT3A* and *TET2*. Similarly, the TCGA AML cohort reveals *IDH1, IDH2, NRAS, U2AF1, KIT, FLT3, DNMT3A, RUNX1, WT1, and KRAS*.

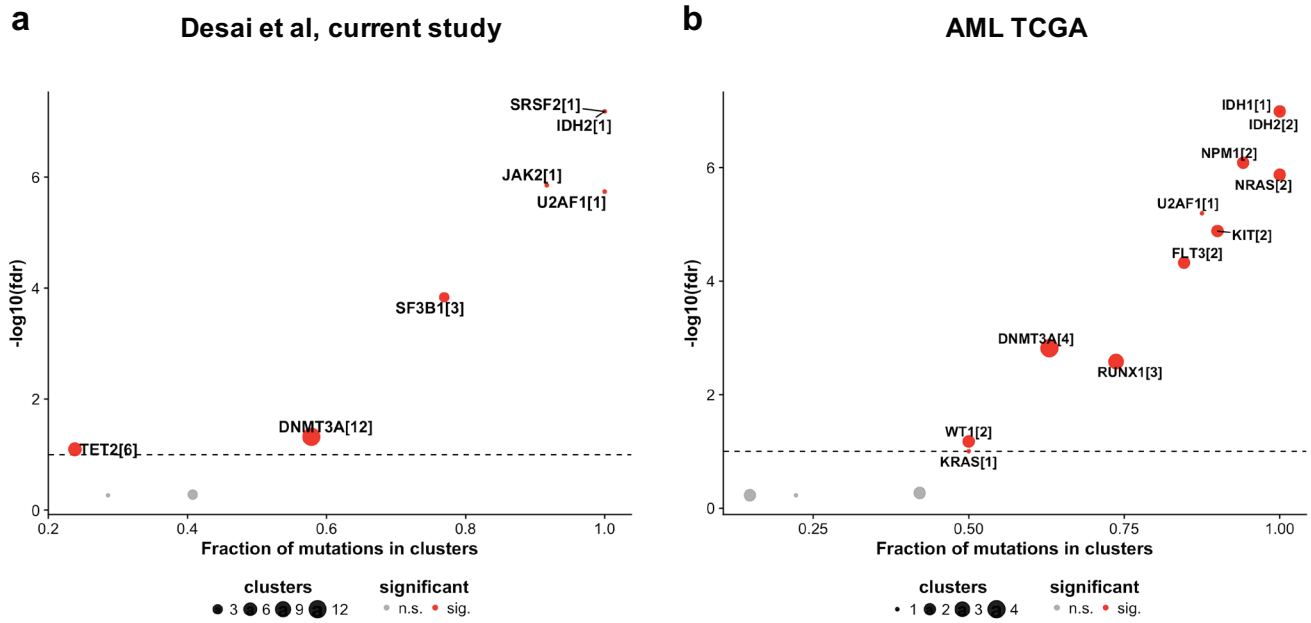


Figure S22. Odds of AML are elevated when mutations are present at higher VAF.

Forest plot indicating odds ratio of mutations in *DNMT3A*, *TET2* and spliceosome genes at high VAF (> 10%) vs low VAF (< 10%). The spliceosome category includes *SRSF2*, *SF3B1*, and *U2AF1*. Odds ratio (OR) per gene are adjusted by age (years) as a continuous variable. Abbreviations: CI, confidence interval; N, number affected. P-values are shown for penalized likelihood multivariable logistic regression. Exact p-values: TET2 ($P = 7.5 \times 10^{-6}$).

Participants with mutations in these genes and VAF > 10% have increased odds for AML development

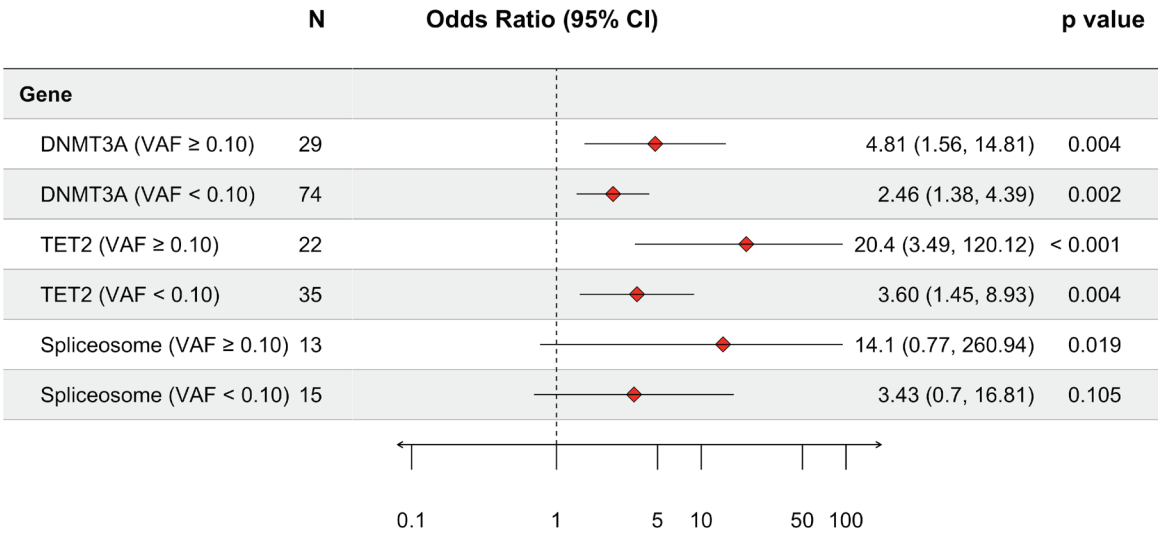
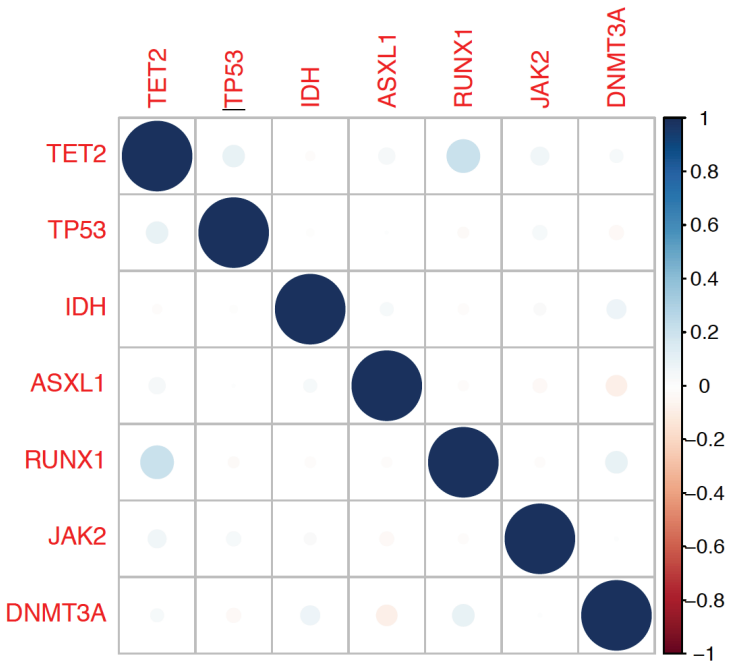


Figure S23. Absence of collinearity between predictors.

Correlation (Pearson's) matrix between recurrently mutated genes demonstrates the absence of collinear variables in multivariable models (N=185).



Short list of WHI investigators

Program Office: (National Heart, Lung, and Blood Institute, Bethesda, Maryland) Jacques Rossouw, Shari Ludlam, Joan McGowan, Leslie Ford, and Nancy Geller

Clinical Coordinating Center: (Fred Hutchinson Cancer Research Center, Seattle, WA) Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg I

Investigators and Academic Centers: (Brigham and Women's Hospital, Harvard Medical School, Boston, MA) JoAnn E. Manson; (MedStar Health Research Institute/Howard University, Washington, DC) Barbara V. Howard; (Stanford Prevention Research Center, Stanford, CA) Marcia L. Stefanick; (The Ohio State University, Columbus, OH) Rebecca Jackson; (University of Arizona, Tucson/Phoenix, AZ) Cynthia A. Thomson; (University at Buffalo, Buffalo, NY) Jean Wactawski-Wende; (University of Florida, Gainesville/Jacksonville, FL) Marian Limacher; (University of Iowa, Iowa City/Davenport, IA) Jennifer Robinson; (University of Pittsburgh, Pittsburgh, PA) Lewis Kuller; (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker; (University of Nevada, Reno, NV) Robert Brunner; (University of Minnesota, Minneapolis, MN) Karen L. Margolis

Women's Health Initiative Memory Study: (Wake Forest University School of Medicine, Winston-Salem, NC) Mark Espeland For a list of all the investigators who have contributed to WHI science, please visit:

https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf

References

35 Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**, e3, doi:10.1093/nar/gkr771 (2012).

36 Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80**, 27-38, doi:10.2307/2336755 (1993).

37 Cai, T. & Zheng, Y. Non-parametric Evaluation of Biomarker Accuracy under Nested Case-control Studies. *J Am Stat Assoc* **106**, 569-580, doi:10.1198/jasa.2011.tm09807 (2011).

38 Samuelsen, S. O. A psudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84**, 379-394, doi:10.1093/biomet/84.2.379 (1997).

39 Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299-314, doi:10.2307/1390807 (1996).

40 Heinze, G. *Logistic regression using Firth's bias reduction: a solution to the problem of separation in logistic regression*, <https://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/fllogistf/> (

41 Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes* **5**, 337, doi:10.1186/1756-0500-5-337 (2012).

42 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv* **00**, 3-3, doi:arXiv:1303.3997 [q-bio.GN] (2013).

43 Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503-2505, doi:10.1093/bioinformatics/btu314 (2014).

44 Beraldi, D. *MarkDupsByStartEnd*, <https://github.com/dariober/Java-cafe/tree/master/MarkDupsByStartEnd> (

45 Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108, doi:10.1093/nar/gkw227 (2016).

46 Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873, doi:10.1371/journal.pcbi.1004873 (2016).

47 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

48 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).

49 *Vt analysis toolkit*, <https://github.com/atks/vt> (2017).

50 Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843-2851, doi:10.1093/bioinformatics/btu356 (2014).

51 Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**, D777-D783, doi:10.1093/nar/gkw1121 (2017).

52 *The Cancer Genome Atlas. National Cancer Institure*, <https://cancergenome.nih.gov/> (

53 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

54 Landau, D. A. *et al.* The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nat Commun* **8**, 2185, doi:10.1038/s41467-017-02329-y (2017).

55 *Picard Tools*, <http://broadinstitute.github.io/picard/> (

56 Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586, doi:10.1038/ng.3703 (2016).

57 Jagadeesh, K. A., Wenger, A., Berger, M., Guturu, H., Stenson, P., Cooper, D., Bernstein, J., and Bejerano, G. *Mendelian Clinically Applicable Pathogenicity (M-CAP) Score*, <http://bejerano.stanford.edu/mcap/> (2016).
58 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849, doi:10.1093/bioinformatics/btw313 (2016).
59 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812, doi:10.1093/bioinformatics/btu393 (2014).
60 Janssen, R. *MutationalPatterns*, <https://github.com/UMCUGenetics/MutationalPatterns> (
61 Meyer, M. J. *et al.* mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum Mutat* **37**, 447-456, doi:10.1002/humu.22963 (2016).
62 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).
63 Jennings, L. J. *et al.* Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* **19**, 341-365, doi:10.1016/j.jmoldx.2017.01.011 (2017).
64 Ryan, S. L. *et al.* The role of the RAS pathway in iAMP21-ALL. *Leukemia* **30**, 1824-1831, doi:10.1038/leu.2016.80 (2016).
65 Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**, 970-976, doi:10.1038/ng.2702 (2013).
66 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
67 Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol* **9**, 958-970, doi:10.1038/nrm2549 (2008).
68 Gorbunova, V., Seluanov, A., Mao, Z. & Hine, C. Changes in DNA repair during aging. *Nucleic Acids Res* **35**, 7466-7474, doi:10.1093/nar/gkm756 (2007).
69 Khwaja, A. *et al.* Acute myeloid leukaemia. *Nat Rev Dis Primers* **2**, 16010, doi:10.1038/nrdp.2016.10 (2016).
70 Kadia, T. M. *et al.* TP53 mutations in newly diagnosed acute myeloid leukemia: Clinicomolecular characteristics, response to therapy, and outcomes. *Cancer*, doi:10.1002/cncr.30203 (2016).
71 Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* **2**, a001008, doi:10.1101/cshperspect.a001008 (2010).
72 Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-2244, doi:10.1093/bioinformatics/btt395 (2013).