https://doi.org/10.1038/s41586-018-0317-6

Prediction of acute myeloid leukaemia risk in healthy individuals

Sagi Abelson[1,46], Grace Collord[2,3,46], Stanley W. K. Ng[4], Omer Weissbrod[5], Netta Mendelson Cohen[5], Elisabeth Niemeyer[6], Noam Barda[7], Philip C. Zuzarte[8], Lawrence Heisler[8], Yogi Sundaravadanam[8], Robert Luben[9], Shabina Hayat[9], Ting Ting Wang[1,10], Zhen Zhao[1], Iulia Cirlan[1], Trevor J. Pugh[1,8,10], David Soave[8], Karen Ng[8], Calli Latimer[2], Claire Hardy[2], Keiran Raine[2], David Jones[2], Diana Hoult[11], Abigail Britten[11], John D. McPherson[8], Mattias Johansson[12], Faridah Mbabaali[8], Jenna Eagles[8], Jessica K. Miller[8], Danielle Pasternack[8], Lee Timms[8], Paul Krzyzanowski[8], Philip Awadalla[8], Rui Costa[13], Eran Segal[5], Scott V. Bratman[1,8,14], Philip Beer[2], Sam Behjati[2,3], Inigo Martincorena[2], Jean C. Y. Wang[1,15,16], Kristian M. Bowles[17,18], J. Ramón Quirós[19], Anna Karakatsani[20,21], Carlo La Vecchia[20,22], Antonia Trichopoulou[20], Elena Salamanca-Fernández[23,24], José M. Huerta[24,25], Aurelio Barricarte[24,26,27], Ruth C. Travis[28], Rosario Tumino[29], Giovanna Masala[30], Heiner Boeing[31], Salvatore Panico[32], Rudolf Kaaks[33], Alwin Krämer[34], Sabina Sieri[35], Elio Riboli[36], Paolo Vineis[36], Matthieu Foll[12], James McKay[12], Silvia Polidoro[37], Núria Sala[38], Kay-Tee Khaw[39], Roel Vermeulen[40], Peter J. Campbell[2,41], Elli Papaemmanuil[2,42], Mark D. Minden[1,10,15,16], Amos Tanay[5], Ran D. Balicer[7], Nicholas J. Wareham[11], Moritz Gerstung[2,13,47]*, John E. Dick[1,43,47]*, Paul Brennan[12,47]*, George S. Vassiliou[2,41,44,47]* & Liran I. Shlush[1,6,45,47]*

The incidence of acute myeloid leukaemia (AML) increases with age and mortality exceeds 90% when diagnosed after age 65. Most cases arise without any detectable early symptoms and patients usually present with the acute complications of bone marrow failure[1]. The onset of such de novo AML cases is typically preceded by the accumulation of somatic mutations in preleukaemic haematopoietic stem and progenitor cells (HSPCs) that undergo clonal expansion[2,3]. However, recurrent AML mutations also accumulate in HSPCs during ageing of healthy individuals who do not develop AML, a phenomenon referred to as age-related clonal haematopoiesis (ARCH)[4–8]. Here we use deep sequencing to analyse genes that are recurrently mutated in AML to distinguish between individuals who have a high risk of developing AML and those with benign ARCH. We analysed peripheral blood cells from 95 individuals that were obtained on average 6.3 years before AML diagnosis (pre-AML group), together with 414 unselected age- and gender-matched individuals (control group). Pre-AML cases were distinct from controls and had more mutations per sample, higher variant allele frequencies, indicating greater clonal expansion, and showed enrichment of mutations in specific genes. Genetic parameters were used to derive a model that accurately predicted AML-free survival; this model was validated in an independent cohort of 29 pre-AML cases and 262 controls. Because AML is rare, we also developed an AML predictive model using a large electronic health record database that identified individuals at greater risk. Collectively our findings provide proof-of-concept that it is possible to discriminate ARCH from pre-AML many years before malignant transformation. This could in future enable earlier detection and monitoring, and may help to inform intervention.

To examine the occurrence of somatic mutations before the development of AML, we carried out deep error-corrected targeted sequencing of AML-associated genes in a discovery cohort of 95 pre-AML cases and 414 age- and gender-matched controls (Supplementary Table 1). A validation cohort comprising 29 pre-AML cases and 262 controls (Supplementary Table 1) was analysed using deep sequencing with an overlapping gene panel. Taking both cohorts together, ARCH, defined on the basis of putative driver mutations (ARCH-PD), was found in 73.4% of the pre-AML cases at a median of 7.6 years before diagnosis. By contrast, ARCH-PD was observed in 36.7% of controls ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Fig. 1a), consistent with data from a study of more than 2,000 unselected individuals assayed using a similarly sensitive method[9,10]. Additionally, 39% of pre-AML cases above the age of 50 had a driver mutation with a variant allele frequency (VAF) of more than 10%, compared to only 4% of controls,

[1]Princess Margaret Cancer Centre, University Health Network (UHN), Toronto, Ontario, Canada. [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [3]Department of Paediatrics, University of Cambridge, Cambridge, UK. [4]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada. [5]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [6]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [7]Clalit Research Institute, Tel Aviv, Israel. [8]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [9]Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge School of Clinical Medicine, Cambridge, UK. [10]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [11]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [12]International Agency for Research on Cancer, World Health Organization, Lyon, France. [13]European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Wellcome Genome Campus, Hinxton, UK. [14]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [15]Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [16]Division of Medical Oncology and Hematology, University Health Network, Toronto, Ontario, Canada. [17]Department of Molecular Haematology, Norwich Medical School, The University of East Anglia, Norwich, UK. [18]Department of Haematology, Norfolk and Norwich University Hospitals NHS Trust, Norwich, UK. [19]Public Health Directorate, Asturias, Spain. [20]Hellenic Health Foundation, Athens, Greece. [21]2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Athens, Greece. [22]Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy. [23]Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibs.GRANADA, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain. [24]CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain. [25]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [26]Navarra Public Health Institute, Pamplona, Spain. [27]Navarra Institute for Health Research, Pamplona, Spain. [28]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [29]Cancer Registry and Histopathology Department, Civic-M. P. Arezzo Hospital, Azienda Sanitaria Provinciale, Ragusa, Italy. [30]Cancer Risk Factors and Life-Style Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy. [31]Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany. [32]Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy. [33]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [34]Clinical Cooperation Unit Molecular Hematology/Oncology, German Cancer Research Center (DKFZ) and Department of Internal Medicine V, University of Heidelberg, Heidelberg, Germany. [35]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy. [36]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [37]Italian Institute for Genomic Medicine, Torino, Italy. [38]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program and Translational Research Laboratory, Catalan Institute of Oncology, ICO-IDIBELL, Barcelona, Spain. [39]University of Cambridge, Cambridge, UK. [40]Division of Environmental Epidemiology and Veterinary Public Health, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands. [41]Department of Haematology, University of Cambridge, Cambridge, UK. [42]Center for Molecular Oncology and Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [43]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [44]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [45]Division of Hematology, Rambam Healthcare Campus, Haifa, Israel. [46]These authors contributed equally: Sagi Abelson, Grace Collord. [47]These authors jointly supervised this work: Moritz Gerstung, John E. Dick, Paul Brennan, George S. Vassiliou, Liran I. Shlush. *e-mail: moritz.gerstung@ebi.ac.uk; John.Dick@uhnresearch.ca; BrennanP@iarc.fr; gsv20@sanger.ac.uk; liranshlush3@gmail.com

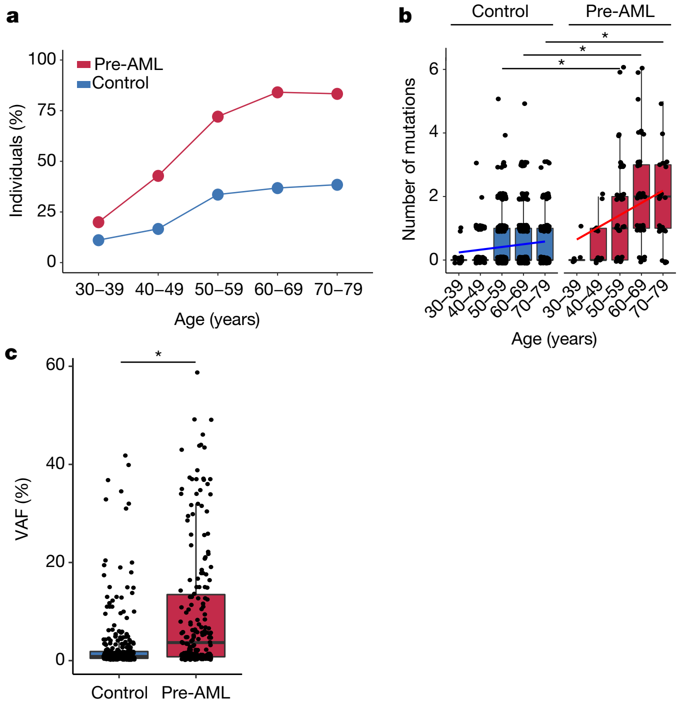


Fig. 1 | Prevalence of ARCH, number of mutations and clone size in individuals who developed AML. a, Prevalence of ARCH-PD among pre-AML cases (red) and controls (blue). **b**, The number of ARCH-PD mutations detected in cases and controls according to age. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. Individual values are indicated as dots. **c**, VAF of ARCH-PD mutations. *$P < 0.0005$, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. All panels show data for $n = 800$ biologically independent samples.

a prevalence that is in line with the largest studies of ARCH in the general population[4] ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Extended Data Fig. 1).

The median number of ARCH-PD mutations per individual increased with age and was significantly higher in the pre-AML group relative to controls (Fig. 1b and Supplementary Table 2). Furthermore, examination of ARCH-PD VAF distribution revealed significantly larger clones among the pre-AML cases ($P = 1.2 \times 10^{-13}$, two-sided Wilcoxon rank-sum test; Fig. 1c). To gain insight into clonal growth dynamics, we examined serially collected samples that were available for a subset of the validation cohort. We did not find significant differences in clonal expansion rates between pre-AML cases and controls (Extended Data Fig. 2a, b), although this may in part reflect the shorter follow-up of pre-AML cases, small sample size and large variance in growth rates (Extended Data Fig. 2c). The observed differences between pre-AML cases and controls may arise through cell-intrinsic or -extrinsic factors. Although these variables have not been adequately studied in ARCH, a number of observations in different contexts, such as aplasia, advanced age and after chemotherapy, have shown that increased clonal fitness is associated with distinct mutations depending on context[10–12]. Notably, mutations in splicing factor genes were significantly enriched among the pre-AML cases relative to the controls (odds ratio, 17.5; 95% confidence interval, 8.1–40.4; $P = 5.2 \times 10^{-16}$, two-sided Fisher's exact test) and were present in significantly younger individuals (median age 60.3 compared to 77.3 years, $P = 1.7 \times 10^{-4}$, two-sided Wilcoxon rank-sum test; Fig. 2a). Previous work suggests that spliceosome mutations appear to confer a competitive advantage in the context of ageing[10]. Therefore, it is possible that the significantly higher prevalence of such clones in younger pre-AML cases may reflect extrinsic selection pressures rather than earlier mutation acquisition.

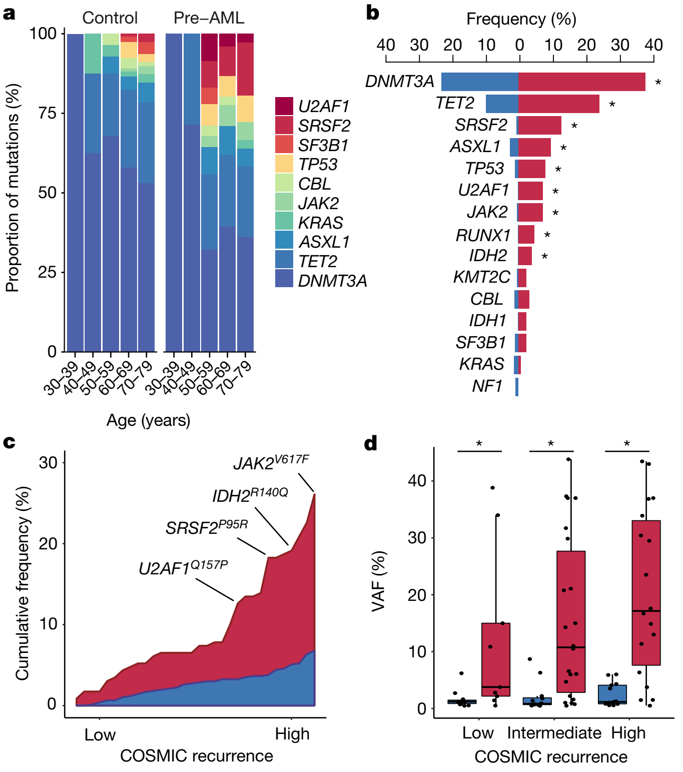


Fig. 2 | Accumulation of specific recurrent AML mutations in healthy individuals at a young age is associated with progression to AML. a, Relative frequency of mutations in the indicated genes according to age group for pre-AML cases and controls. **b**, Proportion of pre-AML cases (red) and controls (blue) who had ARCH-PD mutations in recurrently mutated genes. *$P < 0.05$, Fisher's exact test with Bonferroni multiple testing correction. **c**, The cumulative frequency of recurrent AML mutations (reported in >5 specimens in COSMIC) in pre-AML cases and controls. ARCH-PD mutations are ranked from left to right along the *x* axis from low to high recurrence. **d**, VAF of recurrent mutations in pre-AML cases and controls. Low, intermediate and highly recurrent COSMIC mutations are defined as those reported in 5–19 samples, 20–300 samples and >300 samples, respectively. Box plots indicate median, first and third quartiles and 1.5× interquartile range. *$P < 0.05$, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. All panels show data for $n = 800$ unique individuals.

In line with previous reports[5,6], we found that *DNMT3A* and *TET2* were the most commonly mutated genes in both groups (Fig. 2b). We could not identify any canonical *NPM1* mutations nor any *FLT3*-internal tandem duplication mutations, consistent with these arising late in leukaemogenesis[10,13]. Recurrent *CEBPA* mutations, which are implicated in around 10% of de novo AML[14], were also absent, suggesting that driver events in this gene may also be late events in AML evolution. In order to quantify the effect of different mutations on the likelihood of progression to AML, we ranked ARCH-PD mutations based on the number of times that they have been reported in Catalogue of Somatic Mutations in Cancer (COSMIC) database among individuals with haematological malignancies[15]. We found that mutations that are highly recurrent in cancer specimens were more common in pre-AML cases than in controls with ARCH-PD, whereas driver events in the controls tended to affect loci that are less frequently mutated in haematological malignancies and occurred at significantly lower VAF (Fig. 2c, d). Overall, these findings demonstrate notable differences in the mutational landscape of ARCH and pre-AML. Moreover, this work, in conjunction with recent insights into the origins of AML relapse[16], suggests that AML progression typically occurs over many years through clonal evolution of preleukaemic HSPCs before acquisition of late mutations leads to overt malignant transformation.

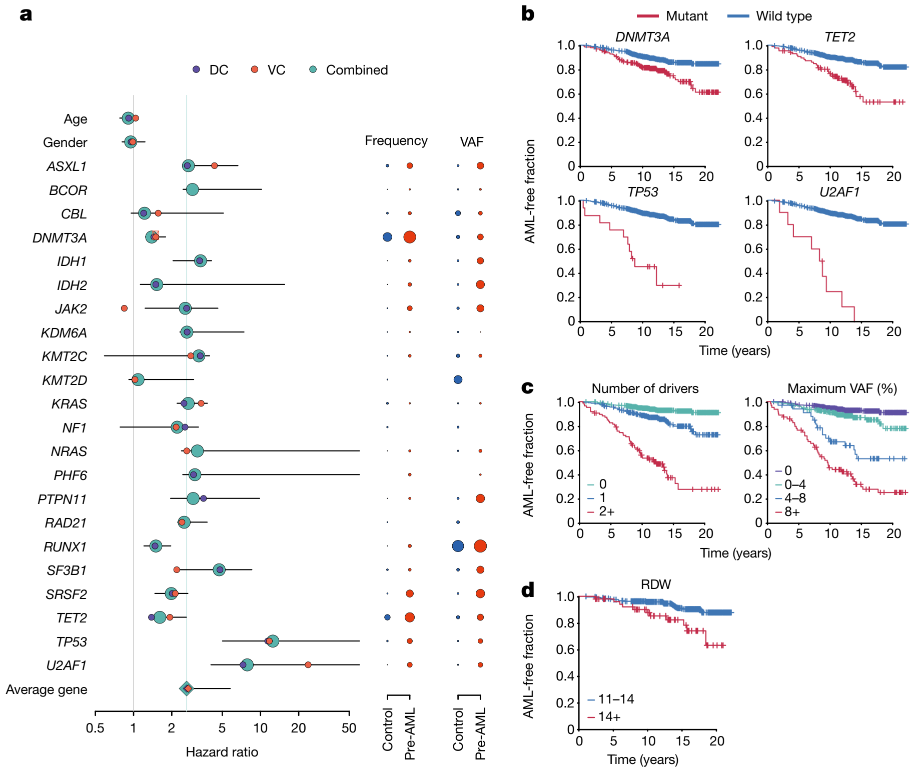


Fig. 3 | Model of future risk of AML. a, Forest plot of the risk of AML. Purple, orange and green circles indicate hazard ratios for the discovery (DC), validation (VC) and combined cohort, respectively. The horizontal lines denote 95% confidence intervals for the combined cohort. For each gene, the indicated hazard ratio applies to the 10-year risk of AML development conferred by each 5% increase in mutation VAF. The green vertical line indicates the mean hazard ratio across all genes. The hazard ratio for *RUNX1* must be interpreted with caution owing to the relatively high prevalence of deleterious germline variants in this gene, which may not be readily distinguishable from somatic mutations in unmatched sequencing assays (see Methods). The proportion of individuals with mutations in each gene and the average VAF are indicated to the right of the forest plot; red and blue circles represent pre-AML cases and controls, respectively, with circle sizes scaled to reflect mutation frequency and VAF. **b**–**d**, Kaplan–Meier curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status for selected genes (**b**), number of driver mutations per individual and largest clone detected (**c**) and RDW (**d**). Data for $n = 796$ unique individuals (**a**–**c**); $n = 299$ individuals for whom RDW measurements were available (**d**).

On the basis of these findings, we next developed an approach to quantify the relative contributions of driver mutations and clone sizes to the risk of progressing to AML. We tested different regularised logistic and Cox proportional hazards regression approaches, which achieved similar performance in both the discovery cohort (concordance $(C) = 0.77 \pm 0.03$) and the validation cohort ($C = 0.84 \pm 0.05$; Extended Data Figs. 3, 4 and Supplementary Table 3). Models that were only trained on data from the discovery or validation cohort had similar coefficients (Fig. 3a). We therefore combined the datasets for a more accurate analysis of the contributions of mutations in individual genes to risk ($C = 0.77 \pm 0.05$; area under curve, 0.79; Supplementary Table 3). Quantitatively, we found that driver mutations in most genes conferred an approximately twofold increased risk of developing AML per 5% increase in clone size (Fig. 3a and Supplementary Table 3). Notable exceptions to this trend are the most frequently mutated ARCH genes, *DNMT3A* and *TET2*, which confer a lower risk of progression to AML (Fig. 3a, b and Supplementary Table 3). By contrast, a larger effect size was apparent for *TP53* (hazard ratio, 12.5; 95% confidence interval, 5.0–160.5) and *U2AF1* (hazard ratio, 7.9; 95% confidence interval, 4.1–192.2) mutations (Fig. 3a, b). However, we note that other ARCH-PD genes, such as *SRSF2*, can contribute a similar relative risk owing to their presence at a higher VAF in pre-AML cases (Fig. 3a, Extended Data Fig. 5a and Supplementary Note). Of note, mutations in *TP53* and spliceosome genes (including *U2AF1*) are also associated with a poorer prognosis in AML[14]. Because the effect of each ARCH-PD mutation is deleterious and the effect of multiple mutations that are present in the same individual is multiplicative, a higher number of mutations is predicted to increase the risk of progression to AML (Fig. 3c). Similarly, the size of the largest driver clone was also strongly associated with the risk of progression to AML, in agreement with the risk of individual mutations generally being proportional to VAF (Fig. 3c). Collectively, although the VAF and the number of mutations confer much of the predictive value, this model does demonstrate distinct gene-level risk factors, and is able to quantify the cumulative impact of multiple mutations and clonal size on the likelihood of progression to AML.

Although our predictive model performs well in identifying those at risk of developing AML in our experimental cohorts, AML incidence rates in the general population are low (4:100,000)[1], and thus millions of individuals would need to be screened to identify the few pre-AML cases, with many false positives. We therefore sought to determine whether routinely available clinical information could improve prediction accuracy or identify a high-risk population for targeted genetic screening. We first analysed complete blood count and biochemistry data that were available for 37 of the pre-AML cases and 262 controls. As reported previously[5,10,17], ARCH-PD was overwhelmingly associated with normal blood counts and this was also the case for pre-AML cases, indicating that these did not represent undiagnosed myelodysplastic syndrome[18]. We identified a significant association between higher red blood cell distribution width (RDW) and risk of progression to AML ($P = 0.0016$, Wald test with Bonferroni multiple-testing correction, Fig. 3d). Although traditionally used in the evaluation of anaemia, raised RDW has been correlated with inflammation, ineffective erythropoiesis, cardiovascular disease and adverse outcomes in several inflammatory and malignant conditions[19]. The correlation between RDW and risk of AML development remained highly significant when controls without ARCH-PD were excluded

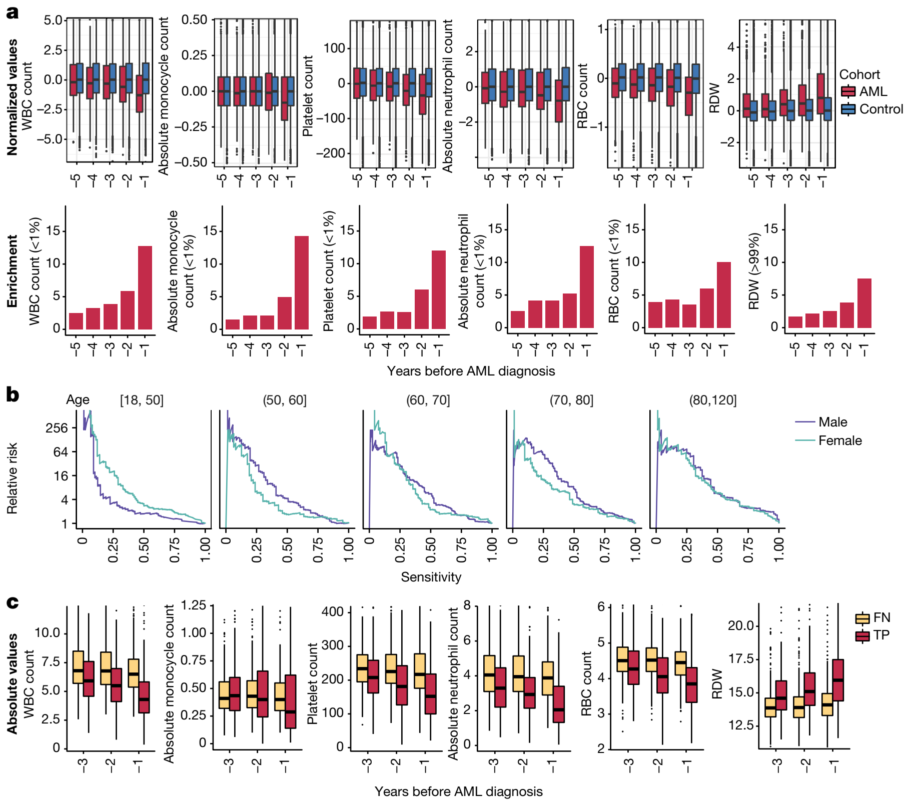


Fig. 4 | Increased risk of AML development inferred from electronic health records. a, Box plot of normalized laboratory measurements. Increased RDW, reduction in monocyte, platelet, red blood cell (RBC) and white blood cell (WBC) counts (top) show a high association (bottom) with a higher risk of AML development and differed at least a year before AML diagnosis. **b**, Model performance stratification by age and gender. Age ranges are indicated above each graph. **c**, Absolute laboratory values for true positive (TP) and false negative (FN) predictions. Box plots indicate median, first and third quartiles and 1.5× interquartile range.

from the analysis ($P = 3.5 \times 10^{-6}$, Wald test with Bonferroni multiple testing correction; Extended Data Fig. 5b). Higher RDW has previously been associated with ARCH and overall mortality[5], but has never been shown to distinguish ARCH from pre-leukaemia. In order to verify RDW as a predictive factor and determine whether additional clinical parameters are associated with risk of AML development, we studied the Clalit database[20], which contains electronic health records that include an average of 3.45 million individuals per year and data that were collected over a 15-year period[21]. We identified 875 cases with AML using stringent criteria based on diagnostic codes and treatment records (Extended Data Fig. 6 and Supplementary Table 4). Analysis of RDW trends revealed significantly raised measurements several years before AML diagnosis relative to age and sex-matched controls (Fig. 4a). Additional parameters that correlated with risk of AML development included reductions in monocyte, platelet, red blood cell and white blood cell counts, albeit usually remaining above the thresholds for clinically relevant cytopenias[18] (Fig. 4a and Extended Data Fig. 7). These findings suggest that evolving de novo AML may sometimes have a considerable prodrome with subtle but discernible clinical manifestations. We next applied a machine-learning approach to construct an AML prediction model based entirely on variables that are routinely documented in electronic health records (Extended Data Fig. 8 and Supplementary Table 4). This model was able to predict AML 6–12 months before diagnosis with a sensitivity of 25.7% and overall specificity of 98.2%. The model performed consistently across different age groups with an increased relative risk of 28 and 24 for males and females, respectively, between the age of 60 and 70 years (Fig. 4b). To better understand which patients are most likely to be accurately classified by this model, we compared absolute laboratory values for true positives and false negatives. We found that 35.5% of false-negative predictions were for patients for whom infrequent blood count data were available (Extended Data Fig. 9). Some of the true-positive cases had mildly abnormal blood counts that would not initiate a diagnostic work-up (Fig. 4c), and cytopenias that would be compatible with undiagnosed myelodysplastic syndrome[18] were uncommon.

Collectively, our findings provide new insights into the pre-clinical evolution of AML and support the hypothesis that individuals at high risk of AML development can be identified years before they develop overt disease. To this end, we present two distinct models for the prediction of de novo AML: one based on somatic point mutations and the other on routinely documented clinical information. We find that basic clinical and laboratory data can identify a high-risk subgroup 6–12 months before AML presentation, while genetic information can identify a substantial fraction of cases several years to more than a decade before diagnosis. By characterizing features that distinguish benign ARCH from pre-leukaemia, our models give valuable insights into leukaemogenesis. It is evident from the current study, together with our recent analysis of mutation acquisition from pre-leukaemic development through to relapse[16], that long-term pre-leukaemic HSPCs frequently carry mutations and undergo considerable clonal expansion while retaining differentiation capacity for years before AML diagnosis. Furthermore, it is clear that some mutations, particularly those affecting *TP53* and *U2AF1*, impart a relatively high risk of subsequent AML, whereas mutations in other genes, for example *DNMT3A* and *TET2*, confer a lesser risk of malignant transformation. Previous studies suggest that oncogenic mutations in *TP53* and spliceosome genes confer little or no competitive advantage in the absence of particular selective pressures[11,22], indicating that cell-extrinsic factors may be important determinants of clonal trajectory.

Cancer predictive models have enabled successful early detection and intervention programmes for several solid tumours[23–25]. However, screening tests are unavailable for the sub-clinical stages of most haematological malignancies. Our study provides proof-of-concept for the feasibility of early detection of healthy individuals at high risk of developing AML, and is a first step in the design of future clinical studies to investigate the potential benefits of early interventions in this deadly disease. However, the infrequency of AML necessitates that future screening tests provide high sensitivity and specificity. Our findings suggest that basic clinical data may identify a higher risk population that might benefit from targeted genetic screening. Equally, combining clinical and genetic information in a single model and including structural driver events is likely to improve model accuracy further. Nevertheless, establishing the utility of such a tandem approach will require extensive clinical and genetic analysis on the same population cohort, in a prospective setting. Furthermore, ARCH is associated with several non-malignant conditions[4,5], and may have a causal role in cardiovascular disease[26,27]. Therefore, genetic testing for ARCH may also prove useful in the management of common age-related diseases. Moreover, this study has broader implications for cancer screening and early intervention beyond AML. Advances in sequencing technologies have revealed a remarkable degree of somatic genetic diversity in normal ageing tissues, often characterized by the presence of clones that have canonical oncogenic mutations[28]. The degree to which clones at high risk of malignant transformation can be reliably distinguished from their indolent counterparts is an important biological question with compelling clinical ramifications. Understanding the selective pressures and cell-intrinsic mechanisms governing clonal fate is the next important step in developing strategies to predict and prevent progression to overt malignancy.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0317-6

Received: 9 July 2017; Accepted: 3 May 2018;
Published online 9 July 2018.

1. Deschler, B. & Lübbert, M. Acute myeloid leukemia: epidemiology and etiology. *Cancer* **107**, 2099–2107 (2006).
2. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl Acad. Sci. USA* **111**, 2548–2553 (2014).
3. Shlush, L. I. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
4. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
5. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
6. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
7. Busque, L. et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).
8. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
9. Acuna-Hidalgo, R. et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am. J. Hum. Genet.* **101**, 50–64 (2017).
10. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
11. Wong, T. N., et al. Role of *TP53* mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518**, 552–555 (2015).
12. Yoshizato, T. et al. Somatic mutations and clonal hematopoiesis in aplastic anemia. *N. Engl. J. Med.* **373**, 35–47 (2015).
13. Krönke, J. et al. Clonal evolution in relapsed *NPM1*-mutated acute myeloid leukemia. *Blood* **122**, 100–108 (2013).
14. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
15. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
16. Shlush, L. I. et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
17. Buscarlet, M. et al. *DNMT3A* and *TET2* dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* **130**, 753–762 (2017).
18. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
19. Hu, L. et al. Prognostic value of RDW in cancers: a systematic review and meta-analysis. *Oncotarget* **8**, 16027–16035 (2017).
20. Balicer, R. D. & Afek, A. Digital health nation: Israel's global big data innovation hub. *Lancet* **389**, 2451–2453 (2017).
21. Dagan, N., Cohen-Stavi, C., Leventer-Roberts, M. & Balicer, R. D. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. *Br. Med. J.* **356**, i6755 (2017).
22. McKerrell, T. & Vassiliou, G. S. Aging as a driver of leukemogenesis. *Sci. Transl. Med.* **7**, 306fs38 (2015).
23. Vickers, A. J. Prediction models in cancer care. *CA Cancer J. Clin.* **61**, 315–326 (2011).
24. Cassidy, A. et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br. J. Cancer* **98**, 270–276 (2008).
25. Wang, X., Oldani, M. J., Zhao, X., Huang, X. & Qian, D. A review of cancer risk prediction models with genetic variants. *Cancer Inform.* **13**, 19–28 (2014).
26. Fuster, J. J. et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
27. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
28. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).

Acknowledgements This work was supported by a Quest for Cure grant to L.I.S., J.C.Y.W. and M.D.M. from the Leukemia and Lymphoma Society, and the following grants to L.I.S from: ERC Horizon 2020 MAMLE, Abisch-Frenkel foundation and an American Society of Hematology Scholar Award. Further funding to J.E.D. was provided by the Canada Research Chair Program, Ontario Institute for Cancer Research, the province of Ontario, Canadian Cancer Society, the Canadian Institutes for Health Research and the Ontario Ministry of Health and Long Term Care to UHN, whose views are not expressed here. Work conducted at the Sanger Institute was supported by the Wellcome Trust and UK Medical Research Council. S.A. was personally funded by the Benjamin Pearl fellowship from the McEwen Centre for Regenerative Medicine, G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); G.S.V. by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA) and a Cancer Research UK Senior Cancer Research Fellowship (C22324/A23015). G.S.V.'s laboratory is also funded by the Kay Kendall Leukaemia Fund and Bloodwise. We thank A. Mitchell and all members of the Dick and Shlush laboratories for comments and T. Hudson for early study planning; G. Barabash for organising the Clalit dataset collaboration. The EPIC study centres were supported by the Hellenic Health Foundation, Regional Government of Asturias, the Regional Government of Murcia (no. 6236), the Spanish Ministry of Health network RTICCC (ISCIII RD12/0036/0018), FEDER funds/European Regional Development Fund (ERDF), "a way to build Europe", Generalitat de Catalunya, AGAUR 2014SGR726; EPIC Ragusa in Italy-Aire-Onlus Ragusa; Epic Italy-Associazione Italiana per la Ricerca sul Cancro (AIRC) Milan, Italy. S.V.B. and T.J.P. are supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at the Princess Margaret Cancer Centre.

Reviewer information *Nature* thanks R. Levine, P. Van Loo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.W.K.N., O.W., N.M.C. and E.N. contributed equally to the work. S.A. performed error-corrected sequencing, analysed sequencing data, performed statistical analyses, contributed to genetic predictive model derivation and wrote the manuscript. G.C. performed variant calling, statistical analyses, derived genetic predictive models and wrote the manuscript. M.G., S.W.K.N., O.W. and R.C. derived genetic predictive models. N.M.C., E.N. and N.B. derived the clinical prediction model. P.C.Z., Z.Z., I.C., K.N., C.L., C.H., D.H., F.M., J.E., J.K.M., D.P., L.T., P.K., S.V.B. and A.Br. and A.Ba. provided sequencing and technical support and enabled sample acquisition. L.H., Y.S., T.T.W., T.J.P., K.R. and D.J. provided bioinformatics support. R.L., S.H., M.J., K.M.B., A.Kr. and N.J.W. enabled sample acquisition, clinical data curation and/or provided clinical expertise. D.S., J.D.M., P.A., E.S., S.B., P.Be., M.D.M and I.M. contributed to data analysis and interpretation. P.J.C. and E.P. contributed to data interpretation and designed the targeted sequencing assay for the validation cohort. J.C.Y.W. revised the manuscript. J.R.Q., A.Ka., C.L.V., A.T., E.S.-F., J.M.H., R.C.T., R.T., G.M., H.B., S.Pa., R.K., S.S., S.Po., N.J.W., N.S., K.-T.K., M.F., J.M.K., E.R., P.V. and R.V. enabled sample acquisition (EPIC). A.T. and R.D.B. analysed Clalit data and derived the clinical prediction model. M.G. derived predictive genetic models, contributed to sequencing data analysis and manuscript writing. J.E.D. contributed to funding applications, study supervision and manuscript writing. P.Br. supervised sample acquisition from all EPIC centres. G.S.V. and L.I.S. designed and supervised all aspects of the study and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0317-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0317-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.G., J.E.D., P.B., G.S.V. or L.I.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study participants. Samples for both the discovery and validation cohort were obtained from participants in the EPIC study[29]. All relevant ethical regulations were followed. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols were approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, the Weizmann Institute of Science Ethics board approval #60-1 and East of England–Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). Patients with AML were identified based on the following ICD9 codes: 9861/3, 9860/3, 9801/3, 9866/3, 9891/3, 9867/3, 9874/3, 9840/3, 9872/3, 9895/3, 9873/3, which included only cases of de novo AML, and no secondary AML. All patients provided peripheral blood samples for which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen before DNA extraction.

Discovery cohort. In total, 509 DNA samples were collected from individuals upon enrolment into the EPIC study between 1993 and 1998 across 17 different centres[29] (Supplementary Table 1). Altogether, 95 individuals who developed AML an average of 6.3 years (interquartile range (IQR) = 4.8 years) after the sample was collected were included in the pre-AML group. For the control group, 414 age- and gender-matched individuals were selected, as they did not develop any haematological disorders during the average follow-up period of 11.6 years (IQR = 2.1 years). The median age at recruitment was 56.7 years (range, 36.08–74.42). In order to minimize any possible demographic biases, an approximate 1:4.5 pre-AML to control ratio was maintained across the different centres.

Validation cohort. Samples were obtained from individuals enrolled in the EPIC-Norfolk longitudinal cohort study between 1994 and 2010. Samples and clinical metadata were available from 37 patients with AML (of which 8 were already included in the discovery cohort) and 262 age- and gender-matched controls without a history of cancer or any haematological conditions. The average time between the first blood sampling and AML diagnosis was 10.5 years (IQR = 8.3 years). The average follow-up period for the control cohort was 17.5 years (IQR = 3.8). For 12 individuals in the pre-AML cohort, 2–3 blood specimens were available, taken a median of 3.4 years apart. Of the 262 controls, 141 had multiple blood samples available, spanning a median of 10.5 years. Blood counts and other clinical parameters were available for all study participants (Supplementary Table 1).

Targeted sequencing. *Discovery cohort sequencing.* Targeted deep sequencing was performed using error-corrected sequencing as follows.

Shearing of genomic DNA, preparation of pre-capture sequencing libraries, hybridization-based enrichment, assessment of the libraries quality and enrichment following hybridization were performed as previously described[30]. In brief, 100 ng of genomic DNA was sheared before library construction (KAPA Hyper Prep Kit KK8504, Kapa Biosystems) with a Covaris E220 instrument using the recommended settings for 250-bp fragments. Following end repair and A-tailing, adaptor ligation was performed using 100-fold molar excess of Molecular Index Adaptor. Library clean-up was performed with Agencourt AMPure XP beads (Beckman-Coulter) and the ligated fragments were then amplified for eight cycles using 0.5 μM Illumina universal and indexing primers.

Targeted capture was carried out on pools containing three indexed libraries. Each pool of adaptor-ligated DNA was combined with 5 μl of 1 mg ml^{-1} Cot-I DNA (Invitrogen), and 1 nmol each of xGEN Universal Blocking Oligo, TS-p5, and xGen Universal Blocking Oligo, TS-p7 (8 nucleotides). The mixture was dried using a SpeedVac and then re-suspended in 1.1 μl water, 8.5 μl NimbleGen 2× hybridization buffer and 3.4 μl NimbleGen hybridization component A. The mixture was heat denatured at 95 °C for 10 min before addition of 4 μl of xGen Lockdown Probes (xGen AML Cancer Panel v.1.0, 3 pmol). Each pool was then hybridized at 47 °C for 72 h. Washing and recovery of the captured DNA was performed according to the manufacturer's specifications. In brief, 100 μl of clean streptavidin beads was added to each capture. Following separation and removal of the supernatant using a magnet, 200 μl 1× Stringent Wash Buffer was added and the reaction was incubated at 65 °C for 5 min. The supernatant containing unbound DNA was removed before repeating the high stringency wash one additional time. Then, the bound DNA was washed as follows: (1) 200 μl 1× Wash Buffer I and separation of the supernatants by magnetic separation; (2) 200 μl 1× Wash Buffer II after magnetic separation; (3) 200 μl 1× Wash Buffer III and removal of the supernatants using magnetic separation. The captured DNA on beads was resuspended in 40 μl of Nuclease-Free water before dividing the total volume into two PCR tubes and subjecting the libraries to 10 cycles of post-capture amplification (manufacturer-recommended conditions; Kapa Biosystems). Before sequencing, libraries were spiked with 2% PhiX.

Validation cohort sequencing. Targeted sequencing was performed using a custom complementary RNA bait set (SureSelect, Agilent, ELID 0537771) designed complementary to all coding exons of 111 genes that have been implicated in myeloid leukaemogenesis (Extended Data Table 1). Genomic DNA was extracted from peripheral whole blood and sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75-bp paired-end sequencing as per Illumina and Agilent SureSelect protocols.

Variant calling. *Discovery cohort variant calling and error correction.* The 126-bp paired-end reads sequencing data from the Illumina platform were converted to FASTQ format, the 2-bp molecular barcode information at each read of the pair was trimmed and was written in the reads' name. The thymine nucleotide required for ligation was removed from the sequences. Burrows–Wheeler aligner (BWA-mem)[31] was used for alignment of the processed FASTQ files to the reference hg19 genome, after realignment of insertions and deletions (indels) using GATK[32]. An in-house algorithm was written to collapse read families that share the same molecular barcode sequence, the left-most genomic position of where each read of the pair maps to the reference and the CIGAR string. Families that consisted of at least two reads were used to generate consensus reads and a consensus base was called when there was at least 70% agreement. When a consensus base was called, it was assigned with the maximum base quality score observed in its corresponding pre-collapsed reads. Furthermore, when possible, duplex reads[33] were generated from two consensus reads, from a singleton read and a consensus read, or from two singleton reads. For each sequenced sample, we generated two BAM files, called BAM1 and BAM2. BAM1 consisted of duplex reads, consensus reads and singleton reads, thereby including some error-corrected and non-error corrected reads, while still containing all the genomic information encoded in the data in the form of unique DNA molecules. BAM2 consisted of duplex reads and consensus reads but not singleton reads. Both files were then analysed to detect single nucleotide variants (SNVs) and small indels using Varscan2[34]. To further remove sequencing artefacts and improve sensitivity, we applied a two-step polishing statistical approach that models the error rate for each sequenced genomic position. For both steps, BAM1 was used and all samples except the sample that was investigated were included for error rate modelling. At step one, as previously described[30], the error rates were modelled by fitting Weibull distribution curves to the non-reference allele fractions. SNVs with allele fractions that were statistically distinguishable from the background error rates ($P = 0$) were further analysed. At step 2, the coverage of the non-reference allele fractions was considered using linear line fitting that describes the negative correlation that exist between the log(non-reference allele fraction) and the corresponding log(coverage) values. This allowed us to estimate different error rates at different coverage depths. Because indel errors are rare and cannot be appropriately modelled by the same statistical framework, they were called using barcode-mediated error correction alone. At least 10 consensus reads, 5 supporting reads on the forward strand, 5 supporting reads on the reverse strand and 2 duplex reads were required to call an indel. Additional post-processing steps applied to data from both the discovery cohort and validation cohort are detailed in 'Additional post-processing filters applied to discovery and validation cohort data'. Variants were annotated using Annovar[35].

Validation cohort variant calling. Sequencing reads were aligned to the reference genome (GRCh37d5) using the Burrows–Wheeler aligner (BWA-aln)[31]. Unmapped reads, PCR duplicates and reads mapping to regions outside the target regions (merged exonic regions and 10 bp either side of each exon) were excluded from analysis. Sequencing depth at each base was assessed using Bedtools coverage v.2.24.0[36].

Somatic SNVs were called using shearwater, an algorithm developed for detecting subclonal mutations in deep-sequencing experiments (https://github.com/gerstung-lab/deepSNV v.1.21.5)[37–39] considering only reads with minimum nucleotide and mapping quality of 25 and 40, respectively. This algorithm models the error rate at individual loci using information from multiple unrelated samples. Additionally, allele counts at the recurrent AML mutation hotspots listed in 'Curation of oncogenic variants' were generated using an in-house script (https://github.com/cancerit/alleleCount) and manually inspected in the Jbrowse genome browser[40]. To further complement our SNV calling approach, we applied an extensively validated in-house version of CaVEMan v.1.11.2 (Cancer variants through expectation maximization)[41]. CaVEMan compares sequencing reads between study and nominated normal samples and uses a naive Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan).

Post-processing filters required that the following criteria were met for CaVEMan to call a somatic substitution. (1) If coverage of the mutant allele was less than 8, at least one mutant allele was detected in the first two-thirds of the read. (2) Less than 3% of the mutant alleles with base quality ≥ 15 were found in the nominated normal sample. (3) Mean mapping quality of the mutant allele reads was ≥ 21. (4) The mutation does not fall in a simple repeat or centromeric region. (5) Fewer than 10% of the reads covering the position contained an indel according to mapping. (6) Less than 80% of the reads report the mutant allele at the same read position. (7) At least a third of the reads calling the variant had a base quality of 25 or higher. (8) Not all mutant alleles reported in the second half of the read. (9) Position does not fall within a germline insertion or deletion.

The following additional post-processing criteria were applied to all SNV calls. (1) Minimum VAF = 0.5% with a minimum of five bidirectional calls reporting the mutant allele (with at least two reads in forward and reverse directions). (2) No indel called within a read length (75 bp) of the putative substitution.

Small indels were sought using two complementary bioinformatics approaches. First, an in-house version of Pindel v.2.2[42] (https://github.com/cancerit/cgpPindel) was applied. We additionally used the aforementioned deepSNV algorithm in order to increase sensitivity for indels present at low VAF. VAF correction was performed using an in-house script (https://github.com/cancerit/vafCorrect).

Post-processing filters required that the following criteria were met for a variant to be called. (1) A minimum of five reads supporting the variant with a minimum of two reads in each direction. For Pindel, the total read count was based on the union of the BWA and Pindel reads reporting the mutant allele. (2) VAF $\geq$ 0.5%. (3) Variant not present within an unmatched normal panel of approximately 400 samples. (4) No reads supporting the variant identified in the nominated normal sample.

Mutations were annotated according to ENSEMBL v.58 using VAGrENT[43] for transcript and protein effects (https://github.com/cancerit/VAGrENT) and Annovar[35] for additional functional annotation.

Additional post-processing filters applied to discovery and validation cohort data. The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event. (1) Any mutant allele reported within 75 bp of another variant. (2) Any mutant allele with a population allele frequency $>$1 in 1,000 according to any of five large polymorphism databases (ExAC, 1000 Genomes Project, ESP6500, CG46 and Kaviar) that is not a canonical hotspot driver mutation with COSMIC recurrence $>$100. (3) Mutations that were present in $>$10% of the control cohort but not recurrent in COSMIC were flagged as potential germline variants or sequencing artefacts. (4) As artefactual indels tend to be recurrent, any indels occurring in $>$2 samples were flagged as for additional inspection.

Curation of oncogenic variants. Putative oncogenic variants were identified according to evidence for functional relevance in AML as previously described and used to define ARCH-PD[14].

Variants were annotated as likely driver events if they fulfilled any of the following criteria. (1) Truncating mutations (nonsense, essential splice site or frameshift indel) in the following genes implicated in AML pathogenesis by loss-of-function: *NF1*, *DNMT3A*, *TET2*, *IKZF1*, *RAD21*, *WT1*, *KMT2D*, *SH2B3*, *TP53*, *CEBPA*, *ASXL1*, *RUNX1*, *BCOR*, *KDM6A*, *STAG2*, *PHF6* and *KMT2C*. (2) Truncating variants in *CALR* exon 9. (3) $JAK2^{V617F}$. (4) *FLT3* internal tandem duplication. (5) Non-synonymous variants at the following hotspot residues: *CBL* E366, L380, C384, C404, R420 and C396; *DNMT3A* R882; *FLT3* D835; *IDH1* R132; *IDH2* R172 and R140; *KIT* W557, V559 and D816; *KRAS* A146, Q61, G13 and G12; *MPL* W515; *NRAS* Q61, G12 and G13; *SF3B1* K700 and K666; *SRSF2* P95; *U2AF1* Q157, R156 and S34. (6) Non-synonymous variants reported at least 10 times in COSMIC with VAF $<$42% and population allele frequency $<$0.003. (7) Non-synonymous variants clustering within a functionally validated locus or within four amino acids of a hotspot variant with population allele frequency $<$0.003 and VAF $<$42%. (8) Non-synonymous variants reported in COSMIC $>$100 times with population allele frequency $<$0.003 regardless of VAF.

Our driver curation strategy inevitably runs a small risk of including germline variants in familial AML genes. We feel that in the real world, where a matched constitutional DNA sample would be unavailable, this is the best approach.

Statistical analysis. All statistical analyses were performed in the R statistical programming environment. A two-sided Wilcoxon rank-sum test was used to assign significance level for differences in the median number of somatic mutations among the pre-AML and control groups, the median VAF of mutations among groups. and the age of individuals with spliceosome mutations. Fisher's exact test was used to assess the significance of differences in the prevalence of ARCH among the groups and spliceosome mutations in the pre-AML group.

Predictive modelling. *Cox proportional hazards model with random effects.* We used a Cox proportional hazards regression to model AML progression-free survival as previously described[14,38]. We used random effects for the Cox proportional hazards model in the CoxHD R package (http://github.com/gerstung-lab/CoxHD). A key strength of this approach is the ability to include many variables in one model while shrinking estimated effects for parameters with weak support in the data, thus controlling for overfitting. We used weighting to minimize the biases introduced by the artificial case–control ratio[44,45] and calculated hazard ratios relative to the (approximate) true cumulative incidence of about 1–3/1,000 in the given age range over a follow up of 10–20 years. The observed driver mutation frequency and VAF in pre-AML cases closely resembled values expected based on the estimated risks, indicating that risk model and driver prevalence are well aligned (Extended Data Fig. 4). Full details of model derivation and comparisons with alternative methods are included in the accompanying code (Supplementary Note, also available at https://github.com/gerstung-lab/preAML). In brief, variables comprised age, gender and the VAF of putative driver mutations (see 'Curation of oncogenic variants' for details of variant curation). We performed agnostic imputation of missing variables by mean and linear rescaling of gene variables by a power of 10 to a magnitude of 1. The model was first trained separately on the discovery cohort and validation cohort. For each of these two models, we evaluated the following measures of predictive accuracy before and after leave-one-out cross-validation (LOOCV): concordance (*C*)[46] and time-dependent area under the receiver-operating characteristic curve (AUC)[47]. The models trained on the validation and discovery cohorts were then cross-validated using the data from the other cohort. In view of the cross-validation results and close correlation between coefficients (Supplementary Table 3), we derived a model on the combined cohorts using both cohorts in order to achieve greater accuracy on the individual effects. Confidence intervals were calculated using 100 bootstrap samples. The coefficients and performance metrics for each iteration of the model are included in Supplementary Table 3.

Concordance measures were obtained using the survConcordance() function implemented in the survival R package[45]. Dynamic AUC was calculated with AUC.uno() implemented in the survAUC package. Time-independent AUCs were calculated using the performance function implemented in the ROCR package. The expected incidence of AML was calculated from the UK office of national statistics, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence. All-cause mortality data was obtained from the office of national statistics (https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables).

Ridge-regularized logistic regression. Using the same covariates as in 'Cox proportional hazards model with random effects', we fitted a ridge-regularized logistic regression model to dichotomised outcome data. While logistic regression is a common choice for case–control analyses, a downside of this approach is the inability to explicitly use time-dependent covariates. The penalty parameter was chosen using LOOCV on the full cohort; this value was then used on the discovery cohort and validation cohort to yield the same scaling of coefficients. Confidence intervals were calculated using 100 bootstrap samples. Fitting was performed using the glmnet R package. AUC as the primary performance metric was calculated using the ROCR R package.

Additional regression models. Two alternative predictive models were developed. Model 1 performs logistic-regression-based predictions using four types of features: gender, age at blood sampling, the sum of the VAFs ARCH-PD reported in COSMIC v.80 to be recurrent (at least two case reports in haematopoietic and lymphoid tissues) and somatic mutation burden of selected genes, where each gene was represented by the sum of the VAFs corresponding to ARCH-PD mutations in that gene. We measured the predictive performance of each gene via the AUC obtained in a fivefold cross-validation when using only the gene as a predictive feature, and only retained genes with AUC $>$ 55% in the final model.

For model 2 we applied LASSO regression as implemented in the glmnet R package, while enabling LOOCV to fit a Cox regression model. A minimal subset of ARCH-PD variants was selected for which the respective weighted combined VAFs were highly predictive of AML development in the training set. Scores were calculated for each patient as a linear combination of VAF of mutations weighted by regression coefficients that were estimated from the training data. As most scores were zero in the training subset, non-zero scores were discretized to take on a value of 1 that corresponds to AML prediction.

Models 1 and 2 were trained on the discovery cohort and tested for their association with AML development using the validation cohort data. Survival analysis was performed using the Kaplan–Meier and Cox proportional hazards models. Wald's test was used to evaluate the significance of hazard ratios. Logistic regression models were used with the positive predictive value metric to determine the ability of various mutations and other patient parameters to predict AML development. The rms R package was used for logistic regression analysis, and the pROC 1.8 R package was used for receiver-operating characteristic curve analysis.

AML-predictive model based on electronic health records. *Clalit database.* The Clalit database includes information from patients covered by the Clalit health services in Israel[20] during the years 2002–2017. The Clalit training-set data, contains the electronic health records (EHR) of 3.45 million individuals per year on average. All data was anonymized through hashing of personal identifiers and addresses and randomization of dates by sampling a random number of weeks for each patient and adding it to all dates in the patient diagnoses, laboratory and medication records. This approach maintained differential data analysis per patient. Diagnoses codes were acquired from both primary care and hospitalization records, and were mapped to the ICD-9 coding system for historical reasons, with few exceptions that used a partial ICD-10 coding system. Laboratory records were normalized for age and gender by subtracting raw test values from the median levels observed among all test values with matching gender and age (using a bin size of five years). We observed some chronological biases in laboratory ranges, but avoid normalizing these and instead insured case and controls are matched for chronological distributions.

Defining AML cases. We screened for all active patients ($18 < \text{age} < 100$) who were diagnosed with AML (ICD-9 code 205.0*) between the years 2003 and 2016. We then excluded cases based on the following criteria. (1) We excluded patients with prior myeloid malignancies to omit secondary AML, consistent with the case selection for the genetic model. The following diagnosis were excluded if documented within five years before the diagnosis of AML: essential thrombocythemia (ICD-9 238.71), low-grade myelodysplastic syndrome (MDS) (ICD-9 238.72); high-grade MDS lesions (ICD-9 238.73); MDS with 5q deletion (ICD-9 238.74); MDS, unspecified (ICD-9 238.75); polycythemia vera (ICD-9 238.4); myelofibrosis (ICD-9 289.83); chronic myelomonocytic leukaemia (ICD-9 206.10-206.22).

(2) Patients that had any procedures performed on bone marrow or spleen (ICD-10 code Z41) in the five-year period before first mention of AML diagnosis code in their record. These patients were presumed to have an inaccurate AML diagnosis date or misdiagnosis recorded.

(3) Patients that received medications suggestive of an alternative diagnosis of chronic myeloid leukaemia, lymphoid malignancy or acute promyelocytic leukaemia (APL). At any time before diagnosis: imatinib, dasatinib, anagrelide, hydroxycarbamide, asparaginase, pegaspargase or arsenic trioxide. At any time after diagnosis: imatinib, dasatinib, methotrexate, tretinoin or arsenic trioxide. At any time after diagnosis, along with any acute lymphoblastic leukaemia diagnosis (ICD-9 204) or more than single dose: mercaptopurine. APL cases were excluded as early diagnosis of APL will most probably not change its outcome, as treatment is successful already.

(4) Patients without a hospitalization record within three months before or after the onset diagnosis. This parameter was used as it is unlikely that a patient with AML will not be hospitalized close to diagnosis. This filter reduced false-positive cases and better defined the onset date.

We refined the estimated time of onset using the earliest time at which any of the following diagnosis appeared in the patient's history: amyloidosis (ICD-9 277.3), lymphoid leukaemia (ICD-9 204), myeloid leukaemia (ICD-9 205), leukaemia of unspecified cell type (ICD-9 208).

This strategy retained 875 AML cases in the training set for further analysis. These were further validated by manual expert inspection of the complete records of 8% of the cases.

To define the control set, we included all Clalit individuals that were not cases. Since our analysis was aggregating data from a historical time window of 15 years, we associated each control with a randomized time point for evaluation. Using this approach, both cases and controls represented a specific time point in the historical record of a patient, with matching calendric, age and gender distributions. Through this strategy 5,238,528 controls were used.

Defining features for construction of a predictive a score. We extracted the following features for discriminative analysis of cases and controls (this procedure was applied repeatedly in cross-validation as discussed below). (1) Age (in years) at time point. (2) Gender. (3) Laboratory features. Out of 2,770 different types of laboratory tests, we selected the top 50 most frequent laboratory tests (Supplementary Table 4). For each laboratory measurement, we used median age- and gender-normalized test values per patient in three time windows for 6–12 months before onset, 1–2 years before onset and 2–3 years before onset. In addition, we compute the slope of the normalized laboratory measurements for the 6–12 month time window using a linear regression model. (4) Diagnosis features. Of the 1780 different major ICD-9 diagnosis codes, we selected only diagnoses that were previously observed in at least 10 different cases and have an increased relative risk for AML >twofold (as observed in the training set, Supplementary Table 4). For each diagnosis code, we mark whether it appeared in each of the patients in time intervals of 6 months to 3 years, and 3–5 years before onset. (5) BMI features. For each patient in the cohort, we extracted median BMI, weight and height as measured in time intervals of 6 months to 2 years, and 2–3 years before onset.

Gradient boosting. We used the R package xgboost to infer parameters for a classifier given cases and controls. Objective was set to binary:logistic, the evaluation metric to AUC. We set nrounds = 5000, eta = 0.001, gamma = 0.1, lambda = 0.01, alpha = 0.01, max_depth = 6, min_child_weight = 2, subsample = 0.7 and colsample_bytree = 0.7. The boosting algorithm reports a function f that computes a predictive score given the features. Given a threshold T the expression $f(\text{patient features}) > T$ defines a classifier. To standardise thresholds we estimate quantiles for the scores on the training set $T(p) = \text{quantile}(f(\text{train}),p)$ and define the classifier for specificity level p as $f(\text{patient features}) > T(p)$ (Supplementary Table 4).

Cross-validation and relative risk evaluation. To evaluate the predictive value of the classification scheme while considering the strong age and gender biases in the incidence of AML, we performed fivefold cross-validation after splitting the cases and controls into five age- and gender-matched groups. For each fold, we sampled 100,000 controls and combined with the cases, constructed the feature set and trained the model. The model was then tested on the fold cases along with 200,000 sampled controls. We used standardized classifier parameters and standardized thresholds that were inferred based on each training set to generate a series of classifications on each test set and merged these based on the control quantiles in the test as described above. Given a threshold p to define high and low prediction score, we counted for each bin b that defines a patient in a specific age (<40, 40–50, 50–60, 60–70, 70–80, >80) and gender group: the number of cases in bin b ($N^b{}_{\text{case}}$) and the number of controls in bin b ($N^b{}_{\text{control}}$) where N^b is the number of patients in bin b (entire database minus recall controls that are only a sample of the cohort). $N^b(\text{case, high score}) = N^b{}_{\text{TP}}$ indicates the number of true positives (TP); $N^b(\text{case, low score}) = N^b{}_{\text{FN}}$ indicates the number of false negatives (FN); $N^b(\text{control, high score}) = N^b{}_{\text{FP}}$ indicates the number of false positives (FP); $N^b(\text{control, low score}) = N^b{}_{\text{TN}}$ indicates number of true negatives (TN).

For each age and gender group, the absolute risk for AML in the bin is computed by $r^b{}_{\text{abs}} = N^b{}_{\text{case}}/N^b$. The absolute risk given a high score is estimated as $r^b{}_{\text{abs,high}} = N^b{}_{\text{TP}}/(N^b{}_{\text{FP}}+ N^b{}_{\text{TP}})$. The relative risk in the bin is defined by $\text{rr}^b = r^b{}_{\text{abs,high}}/r^b{}_{\text{abs}}$ where the sensitivity level for the classifier threshold level is defined as $\text{sense}^b = N^b{}_{\text{TP}}/N^b{}_{\text{case}}$.

$$\text{rr} = \frac{\dfrac{\dfrac{\text{TP} \times \text{cases}}{(\text{TP} + \text{FN})}}{\dfrac{\text{TP} \times \text{cases}}{(\text{TP} + \text{FN})} + \dfrac{\text{FP} \times \text{controls}}{(\text{FP} + \text{TN})}}}{\dfrac{\text{cases}}{\text{cases}+\text{controls}}}$$

Clonal growth rate calculation. Individual clones were defined by different mutations in different study participants. Per clone we calculated α according to the following equation:

$$a = \log(V/V_0)\ /(T - T_0)$$

where T and T_0 indicate the age of the individual at the two measurement time points. V and V_0 correspond to the VAF at T and T_0, respectively.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code for derivation of the prediction model is publically available on Github (https://github.com/gerstung-lab/preAML). Code for the analysis of error-corrected sequencing is available from the Shlush lab upon request.

Data availability. Targeted sequencing data for the discovery cohort are deposited as BAM files at the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAD00001003583. All other data are available from the corresponding authors upon reasonable request. Sequencing data for the validation cohort are deposited at the European Genome-phenome Archive with accession number EGAD00001003703.

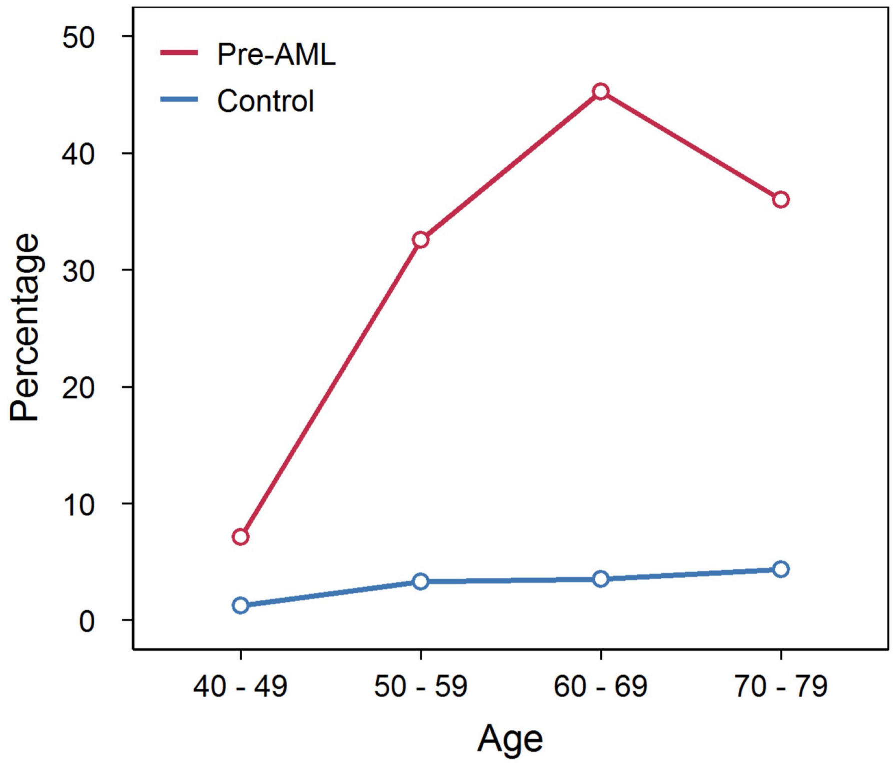
29. Riboli, E. et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
30. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
31. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
32. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
33. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
34. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
35. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10**, 1556–1566 (2015).
36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
37. Gerstung, M. et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
38. Gerstung, M. et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49**, 332–340 (2017).
39. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
40. Buels, R. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
41. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
42. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.17.1–15.7.12 (2015).
43. Menzies, A. et al. VAGrENT: Variation Annotation Generator. *Curr. Protoc. Bioinformatics* **52**, 15.18.1–15.18.11 (2015).

44. Antoniou, A. C. et al. A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. *Genet. Epidemiol*. **29**, 1–11 (2005).

45. Therneau, T. & Grambsch P. M. *Modeling Survival Data: Extending the Cox Model* 1st edn (Springer-Verlag, New York, 2000).

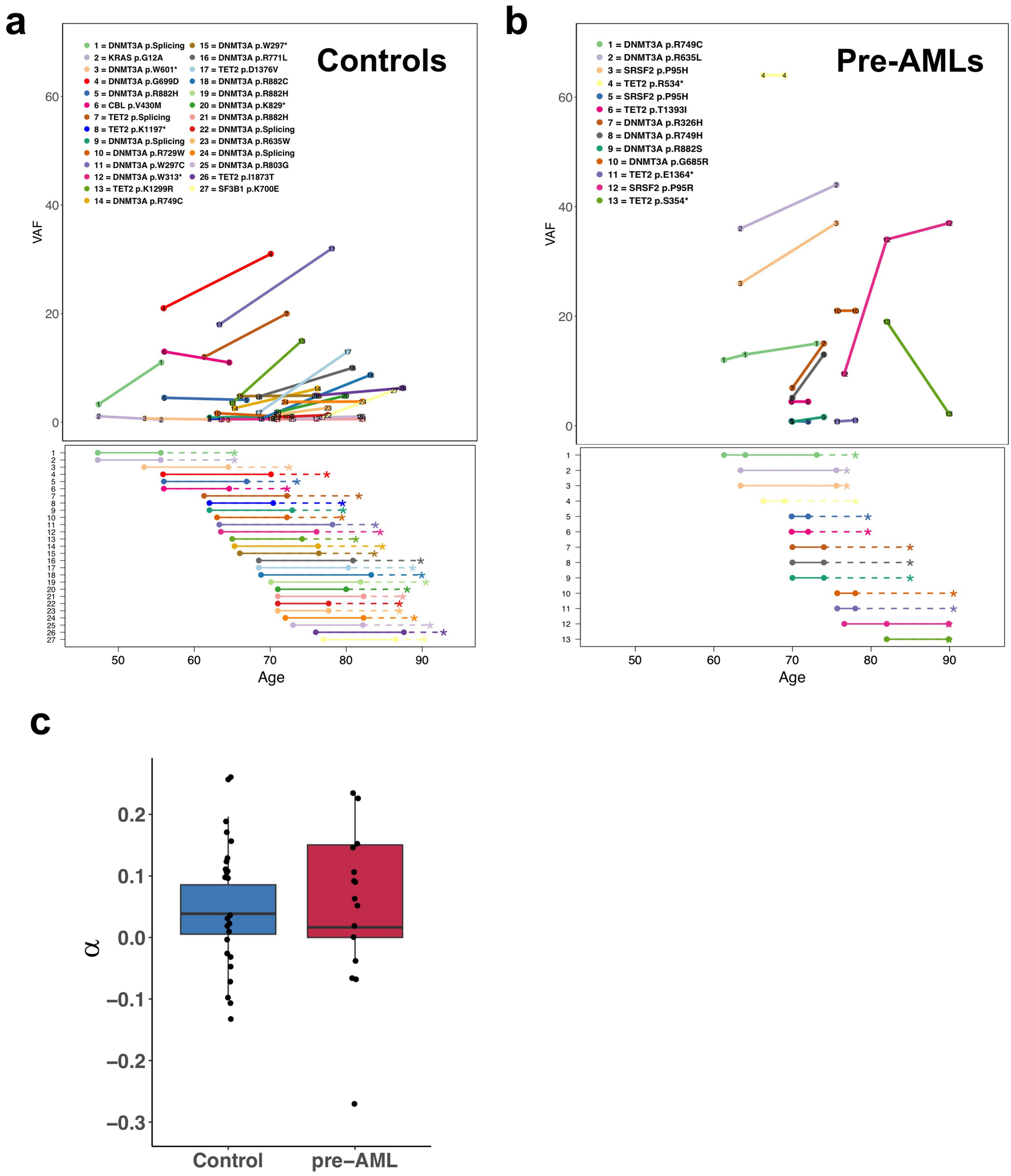
46. Harrell, F. E. Jr, Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med*. **15**, 361–387 (1996).

47. O'Quigley, J., Xu, R. & Stare, J. Explained randomness in proportional hazards models. *Stat. Med*. **24**, 479–489 (2005).

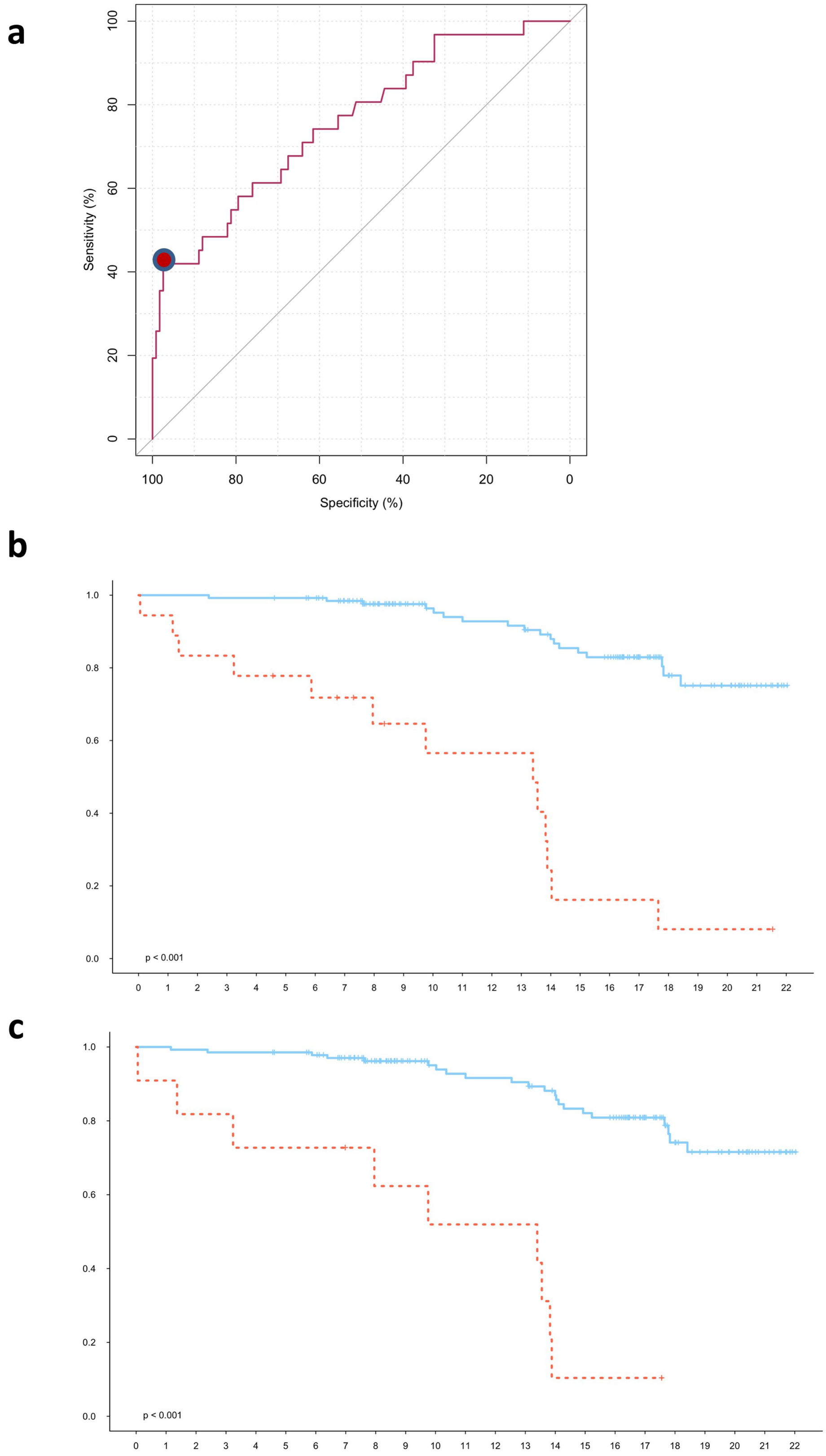


No. with mutation				
Pre-AML	1	14	19	7
Control	1	8	9	9
Total				
Pre-AML	14	43	42	25
Control	82	242	254	161

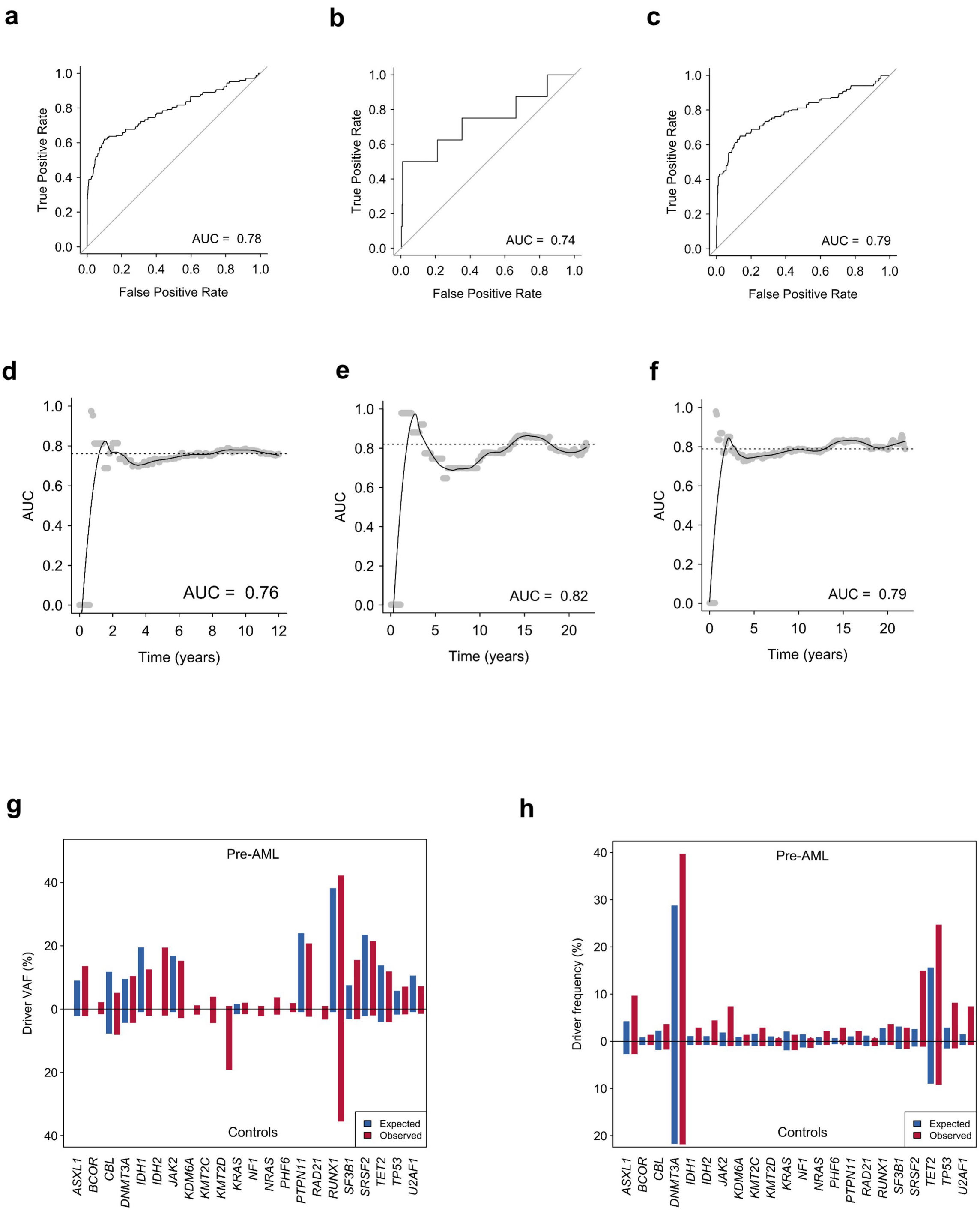
Extended Data Fig. 1 | Prevalence of ARCH-PD mutations with VAF ≥ 10% according to age. Red and blue lines represent the proportion of pre-AML cases and controls, respectively, that had ARCH-PD mutations with VAF ≥ 10%.



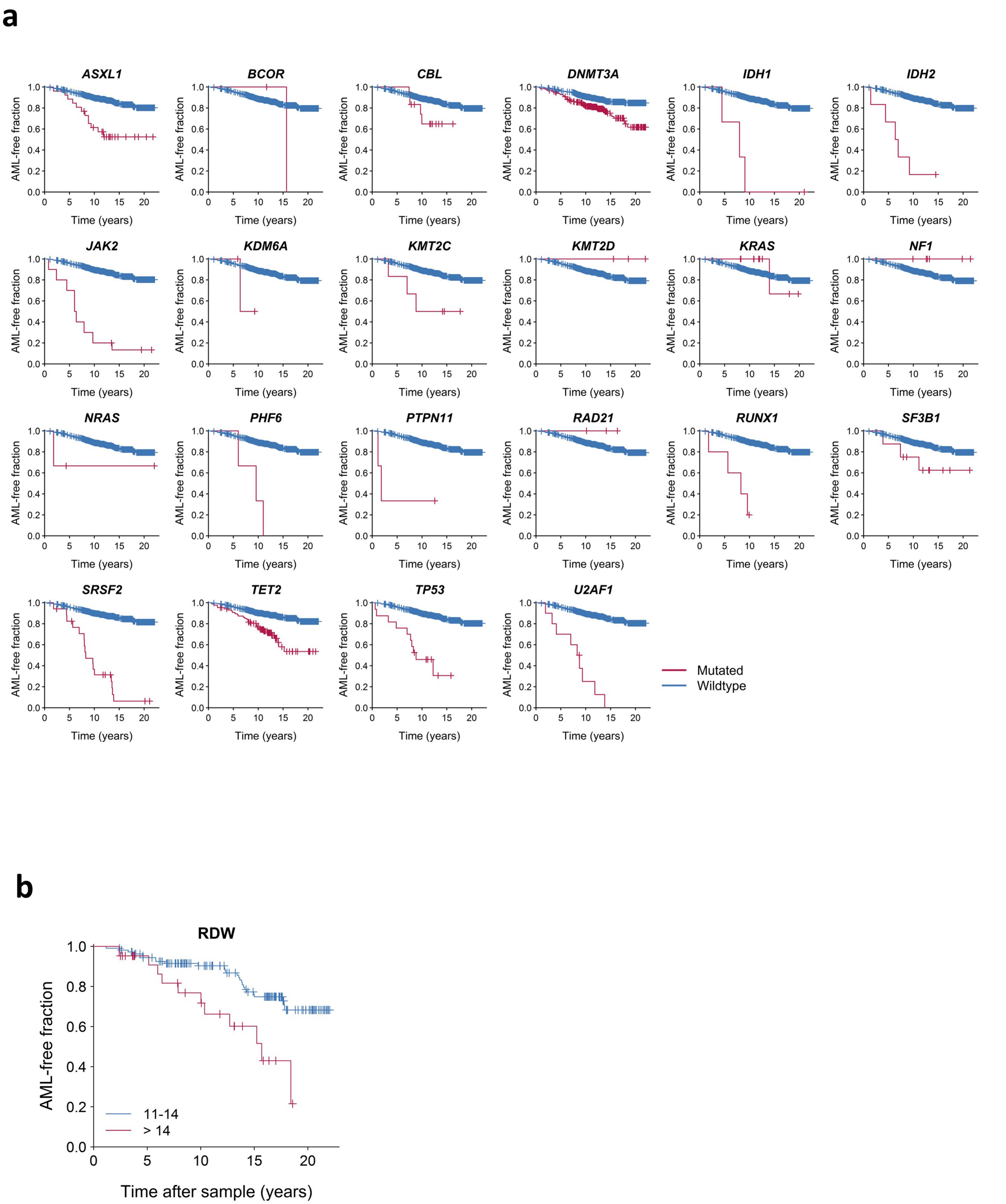
Extended Data Fig. 2 | Serially collected sampling supports a long-lived HSPCs as the cell of origin for most ARCH-PD clones. a, **b**, VAF trajectory of persistent clones carrying putative driver mutations in controls (**a**) and pre-AML cases (**b**). Age is indicated on the *x* axis. Top, VAF is shown on the *y* axis and each persistent mutation is shown in a different colour, with circles denoting individual serial samples and solid lines representing the growth trajectory between serial samples. Bottom, dashed lines indicate the time interval between the last sampling and the end of follow-up (controls) or AML diagnosis (cases). **c**, Clonal growth rates (α) are shown for 27 control clones corresponding to 54 time points and 13 pre-AML clones corresponding to 15 time points. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5 × interquartile range, respectively.



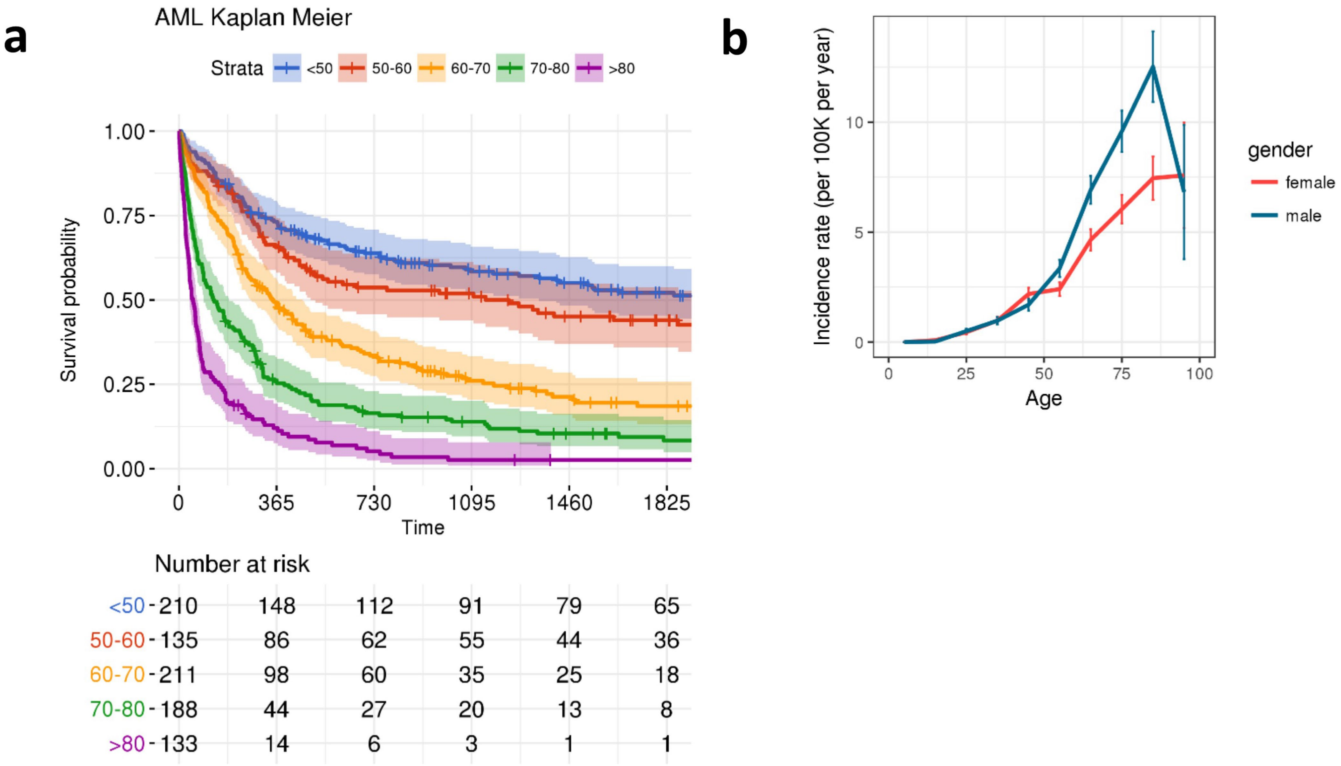
Extended Data Fig. 3 | Performance of the combined model in predicting progression to AML. a, Receiver operating characteristic curve for prediction of AML development using model 1 (see Methods). The red dot indicates the point on the curve with the highest positive predictive value with sensitivity of 41.9% and specificity of 95.7%. **b**, **c**, Kaplan–Meier estimates of time to AML diagnosis for individuals predicted to develop AML (red) and not develop AML (blue) using model 1 (**b**; hazard ratio, 10.38; $P = 4.2 \times 10^{-10}$, Wald test) and model 2 (**c**; hazard ratio, 10.75; $P = 1.75 \times 10^{-8}$, Wald test), from the point of enrolment until the end of follow-up for patients enrolled in the EPIC study.



Extended Data Fig. 4 | AML predictive models. a–**c**, Time-dependent receiver operating characteristic curve for Cox proportional hazards model trained on the discovery cohort ($n = 505$ unique individuals, 91 pre-AML and 414 controls) (**a**), validation cohort ($n = 291$ unique individuals, 29 pre-AML and 262 controls) (**b**) and combined cohorts (**c**). **d**–**f**, Dynamic AUC for Cox proportional hazards models trained on the discovery cohort (**d**), validation cohort (**e**) or combined cohort (**f**). **g**, **h**, Red and blue bars indicate the observed and expected VAF (**g**) and driver frequency (**h**) of pre-AML cases and controls for each gene indicated on the *x* axis.

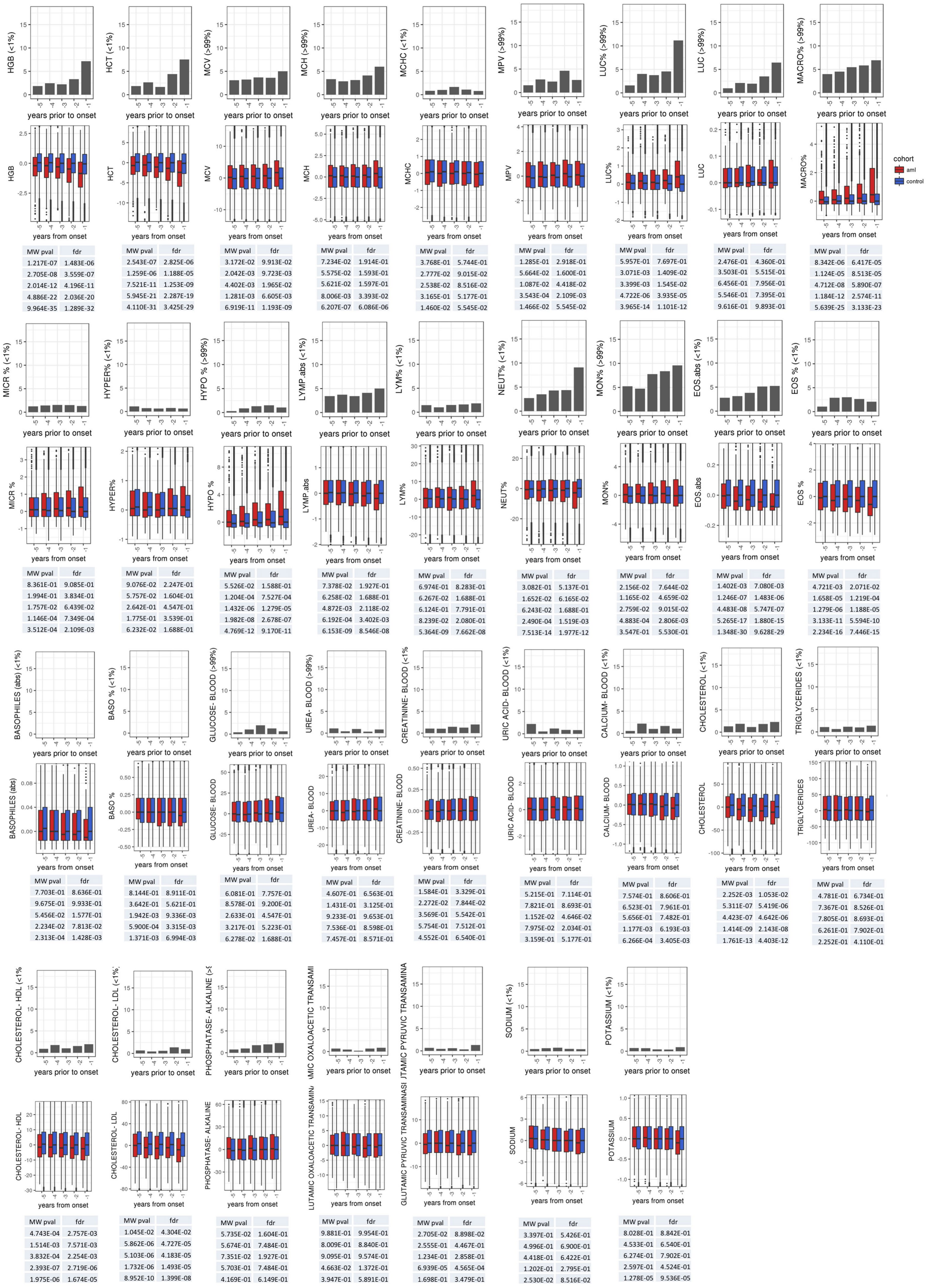


Extended Data Fig. 5 | AML-free survival based on mutation status and RDW. a, Kaplan–Meier curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status in genes mutated in at least three samples across the combined validation and discovery cohorts. $n = 796$ unique individuals. **b**, Kaplan–Meier curve of AML-free survival stratified according to RDW value >14 or ≤14. Plot represents data for $n = 128$ biologically independent individuals who had RDW measurements, including all pre-AML cases regardless of ARCH-PD status, and controls with ARCH-PD (controls without detectable mutations were omitted).



<50	210	148	112	91	79	65
50-60	135	86	62	55	44	36
60-70	211	98	60	35	25	18
70-80	188	44	27	20	13	8
>80	133	14	6	3	1	1

Extended Data Fig. 6 | Description of the cohort and the EHR-derived measurements. a, Kaplan–Meier curves showing age stratified survival rates for 875 individuals who developed AML. **b**, Line plot representation of the number of cases per 100,000 control individuals in the EHR database. The centre values and error bars define the mean and s.d., respectively.



HGB

MW pval	fdr
1.217E-07	1.483E-06
2.705E-08	3.559E-07
2.014E-12	4.196E-11
4.886E-22	2.036E-20
9.964E-35	1.289E-32

HCT

MW pval	fdr
2.543E-07	2.825E-06
1.259E-06	1.188E-05
7.521E-11	1.253E-09
5.945E-21	2.287E-19
4.110E-31	3.425E-29

MCV

MW pval	fdr
3.172E-02	9.913E-02
2.042E-03	9.723E-03
4.402E-03	1.965E-02
1.281E-03	6.605E-03
6.919E-11	1.193E-09

MCH

MW pval	fdr
7.234E-02	1.914E-01
5.575E-02	1.593E-01
5.621E-02	1.597E-01
8.006E-03	3.393E-02
6.207E-07	6.086E-06

MCHC

MW pval	fdr
3.768E-01	5.744E-01
2.777E-02	9.015E-02
2.538E-02	8.516E-02
3.165E-01	5.177E-01
1.460E-02	5.545E-02

MPV

MW pval	fdr
1.285E-01	2.918E-01
5.664E-02	1.600E-01
1.087E-02	4.418E-02
3.543E-04	2.109E-03
1.466E-02	5.545E-02

LUC%

MW pval	fdr
5.957E-01	7.697E-01
3.071E-03	1.409E-02
3.399E-03	1.545E-02
4.722E-06	3.935E-05
3.965E-14	1.101E-12

LUC

MW pval	fdr
2.476E-01	4.360E-01
3.503E-01	5.515E-01
6.456E-01	7.956E-01
5.546E-01	7.395E-01
9.616E-01	9.893E-01

MACRO%

MW pval	fdr
8.342E-06	6.417E-05
1.124E-05	8.513E-05
4.712E-08	5.890E-07
1.184E-12	2.574E-11
5.639E-25	3.133E-23

MICR %

MW pval	fdr
8.361E-01	9.085E-01
1.994E-01	3.834E-01
1.757E-02	6.439E-02
1.146E-04	7.349E-04
3.512E-04	2.109E-03

HYPER%

MW pval	fdr
9.076E-02	2.247E-01
5.757E-02	1.604E-01
2.642E-01	4.547E-01
1.775E-01	3.539E-01
6.232E-02	1.688E-01

HYPO %

MW pval	fdr
5.526E-02	1.588E-01
1.204E-04	7.527E-04
1.432E-06	1.279E-05
1.982E-08	2.678E-07
4.769E-12	9.170E-11

LYMP.abs

MW pval	fdr
7.378E-02	1.927E-01
6.258E-02	1.688E-01
4.872E-03	2.118E-02
6.192E-04	3.402E-03
6.153E-09	8.546E-08

LYM%

MW pval	fdr
6.974E-01	8.283E-01
6.267E-02	1.688E-01
6.124E-01	7.791E-01
8.239E-02	2.080E-01
5.364E-09	7.662E-08

NEUT%

MW pval	fdr
3.082E-01	5.137E-01
1.652E-02	6.165E-02
6.243E-02	1.688E-01
2.490E-04	1.519E-03
7.513E-14	1.977E-12

MON%

MW pval	fdr
2.156E-02	7.644E-02
1.165E-02	4.659E-02
2.759E-02	9.015E-02
4.883E-04	2.806E-03
3.547E-01	5.530E-01

EOS.abs

MW pval	fdr
1.402E-03	7.080E-03
1.246E-07	1.483E-06
4.483E-08	5.747E-07
5.265E-17	1.880E-15
1.348E-30	9.628E-29

EOS %

MW pval	fdr
4.721E-03	2.071E-02
1.658E-05	1.219E-04
1.279E-06	1.188E-05
3.133E-11	5.594E-10
2.234E-16	7.446E-15

BASOPHILES (abs)

MW pval	fdr
7.703E-01	8.636E-01
9.675E-01	9.933E-01
5.456E-02	1.577E-01
2.234E-02	7.813E-02
2.313E-04	1.428E-03

BASO %

MW pval	fdr
8.144E-01	8.911E-01
3.642E-01	5.621E-01
1.942E-03	9.336E-03
5.900E-04	3.315E-03
1.371E-03	6.994E-03

GLUCOSE- BLOOD

MW pval	fdr
6.081E-01	7.757E-01
8.578E-01	9.200E-01
2.633E-01	4.547E-01
3.217E-01	5.223E-01
6.278E-02	1.688E-01

UREA- BLOOD

MW pval	fdr
4.607E-01	6.563E-01
1.431E-01	3.125E-01
9.233E-01	9.653E-01
7.536E-01	8.598E-01
7.457E-01	8.571E-01

CREATININE- BLOOD

MW pval	fdr
1.584E-01	3.329E-01
2.272E-02	7.844E-02
3.569E-01	5.542E-01
5.754E-01	7.512E-01
4.552E-01	6.540E-01

URIC ACID- BLOOD

MW pval	fdr
5.215E-01	7.114E-01
7.821E-01	8.693E-01
1.152E-02	4.646E-02
7.975E-02	2.034E-01
3.159E-01	5.177E-01

CALCIUM- BLOOD

MW pval	fdr
7.574E-01	8.606E-01
6.523E-01	7.961E-01
5.656E-01	7.482E-01
1.177E-03	6.193E-03
6.266E-04	3.405E-03

CHOLESTEROL

MW pval	fdr
2.252E-03	1.053E-02
5.311E-07	5.419E-06
4.423E-07	4.642E-06
1.414E-09	2.143E-08
1.761E-13	4.403E-12

TRIGLYCERIDES

MW pval	fdr
4.781E-01	6.734E-01
7.367E-01	8.526E-01
7.805E-01	8.693E-01
6.261E-01	7.902E-01
2.252E-01	4.110E-01

CHOLESTEROL- HDL

MW pval	fdr
4.743E-04	2.757E-03
1.514E-03	7.571E-03
3.832E-04	2.254E-03
2.393E-07	2.719E-06
1.975E-06	1.674E-05

CHOLESTEROL- LDL

MW pval	fdr
1.045E-02	4.304E-02
5.862E-06	4.727E-05
5.103E-06	4.183E-05
1.732E-06	1.493E-05
8.952E-10	1.399E-08

PHOSPHATASE- ALKALINE

MW pval	fdr
5.735E-02	1.604E-01
5.674E-01	7.484E-01
7.351E-02	1.927E-01
5.703E-01	7.484E-01
4.169E-01	6.149E-01

LUTAMIC OXALOACETIC TRANSAMINA

MW pval	fdr
9.881E-01	9.954E-01
8.009E-01	8.840E-01
9.095E-01	9.574E-01
4.663E-02	1.372E-01
3.947E-01	5.891E-01

GLUTAMIC PYRUVIC TRANSAMINASI

MW pval	fdr
2.705E-02	8.898E-02
2.555E-01	4.467E-01
1.234E-01	2.858E-01
6.939E-05	4.565E-04
1.698E-01	3.479E-01

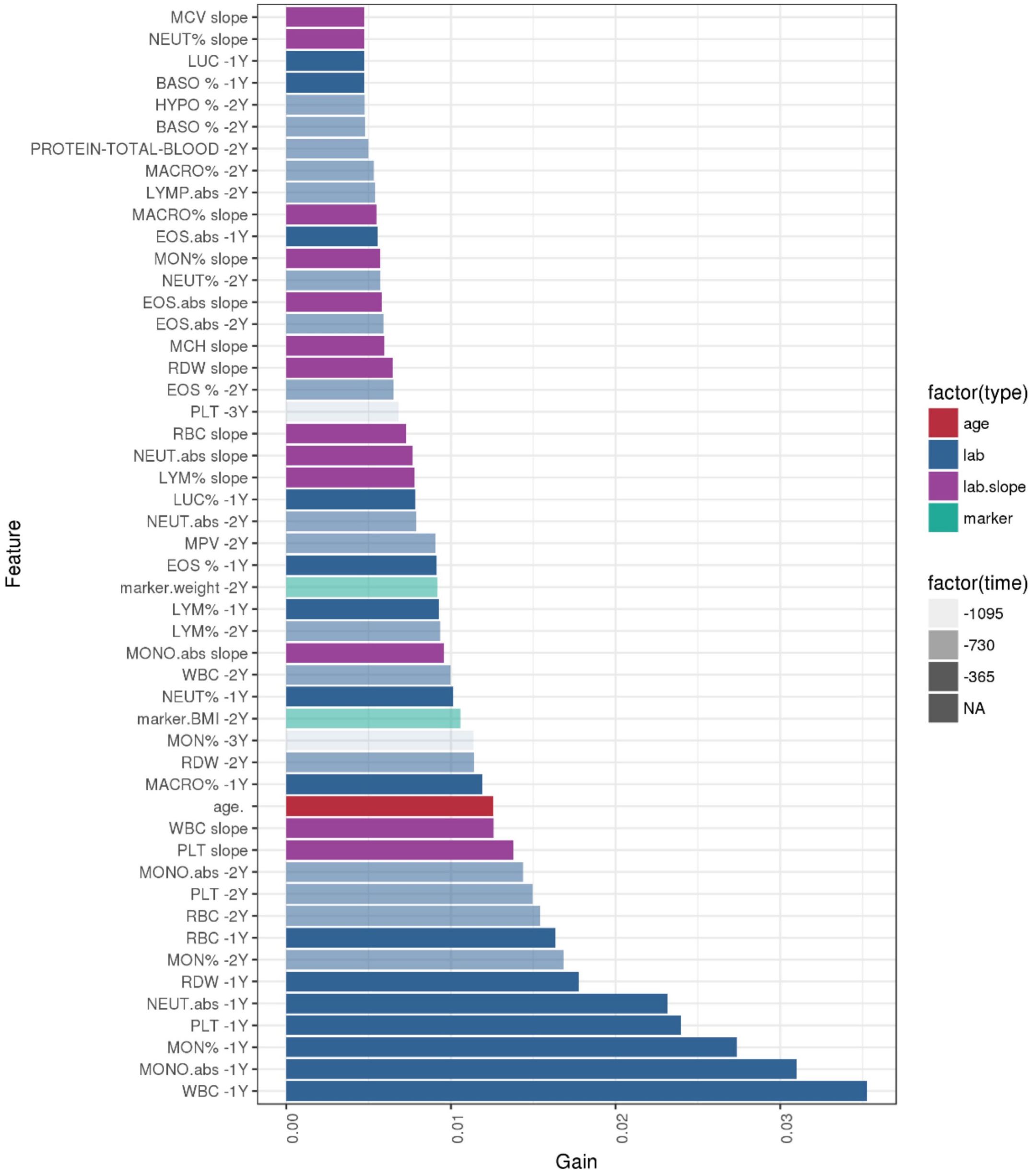
SODIUM

MW pval	fdr
3.397E-01	5.426E-01
4.996E-01	6.900E-01
4.418E-01	6.422E-01
1.202E-01	2.795E-01
2.530E-02	8.516E-02

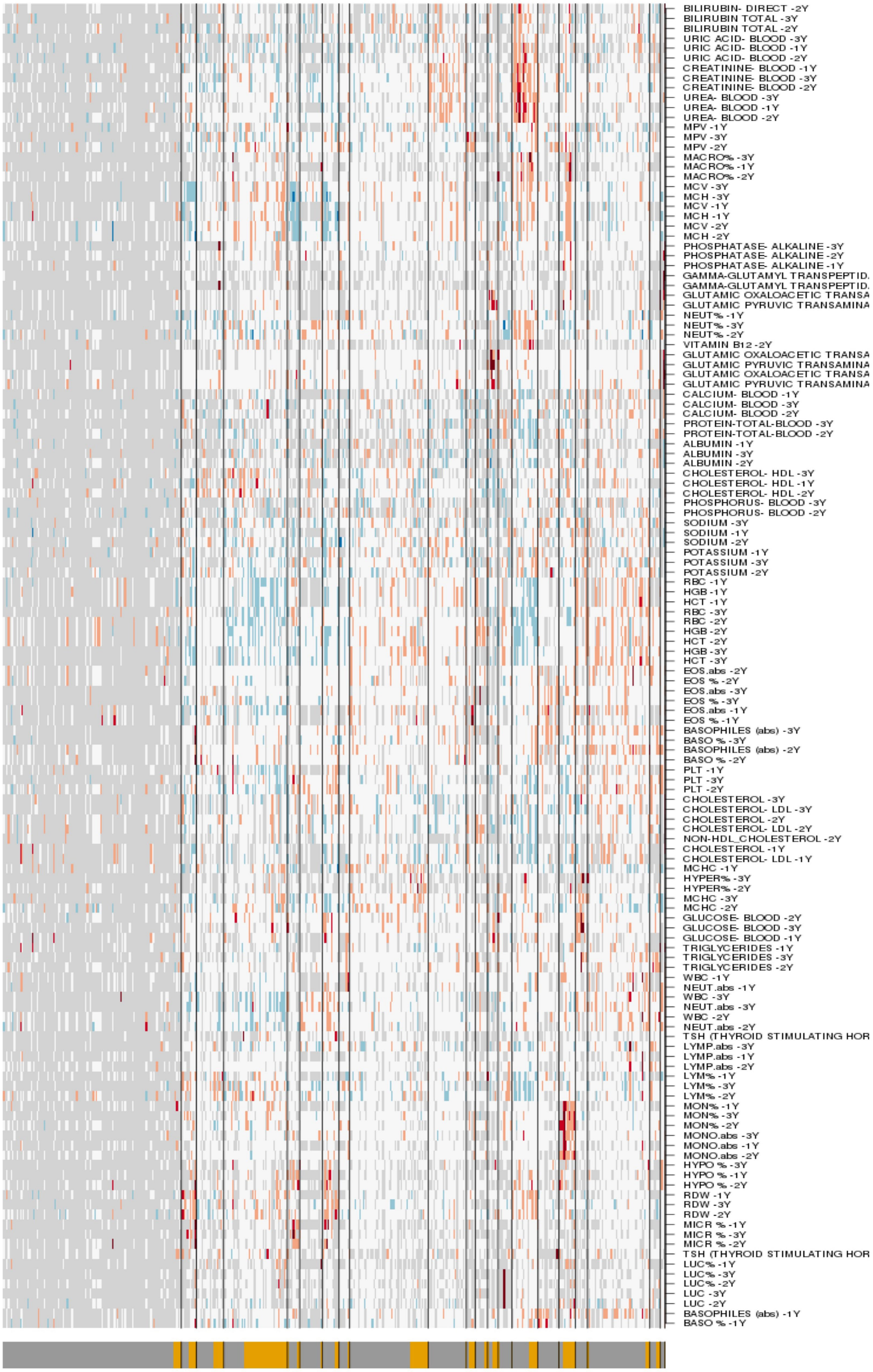
POTASSIUM

MW pval	fdr
8.028E-01	8.842E-01
4.533E-01	6.540E-01
6.274E-01	7.902E-01
2.597E-01	4.524E-01
1.278E-05	9.536E-05

Extended Data Fig. 7 | Laboratory measurements contributing to the EHR model. Normalized laboratory measurements for pre-AMLs (red) and controls (blue) (middle) and their association (bottom) with higher risk of AML are shown. The grey bars indicate the percentage of pre-AML cases with laboratory results either below the 1st percentile or above the 99th percentile. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5 × interquartile range, respectively.



Extended Data Fig. 8 | Top 50 parameters for the EHR model. The relative contribution of the top 50 features incorporated into the EHR prediction model, ranked according to their predictive value (gain). 1Y, one year before AML diagnosis; 2Y, two years before AML diagnosis; 3Y, three years before AML diagnosis; BASO%, percentage of basophils; BMI, body mass index; EOS.abs, absolute eosinophil count; EOS%, percentage of eosinophils; HYPO%, percentage of hypochromia; LUC, large unstained cells; LYM%, percentage of lymphocytes; LYMPH.abs, absolute lymphocyte count; MACRO%, percentage of macrocytosis; MCH, mean corpuscular haemoglobin; MCV, mean corpuscular volume; MON%, percentage of monocytes; MONO.abs, absolute monocyte count; MPV, mean platelet volume; NEUT.abs, absolute neutrophil count; NEUT%, percentage of neutrophils; PLT, platelet count; RBC, red blood cell count; RDW, red cell distributiom width; WBC, white blood cell count.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Distribution of EHR model parameters. Heat map illustrating absolute values of clinical measurements. Blue, white and red indicate low, intermediate and high values, respectively. Light grey indicates missing data. False-negative and true-positive annotations are indicated at the bottom as dark-grey and yellow colour bars, respectively. 1Y, one year before AML diagnosis; 2Y, two years before AML diagnosis; 3Y, three years before AML diagnosis; BASO%, percentage of basophils; EOS%, percentage eosinophils; EOS.abs, absolute eosinophil count; HCT, haematocrit; HDL; high density lipoprotein; HGB, haemoglobin; Hyper%, percentage of hyperchromia; Hypo%, percentage of hypochromia; LDL, low density lipoprotein; LUC, large unstained cells; LYM%, percentage of lymphocytes; LYMP.abs, absolute lymphocyte count; MACRO%, percentage of macrocytosis; MCH, mean corpuscular haemoglobin; MCHC, mean corpuscular haemoglobin concentration; MCV, mean corpuscular volume; MICR%, percentage of microcytosis; MON%, percentage of monocytes; MONO.abs, absolute monocyte count; MPV, mean platelet volume; PLT, platelet count; NEUT%, percentage of neutrophils; NEUT.abs, absolute neutrophil count; RBC, red blood cell count; RDW, red cell distribution width; Transamina, transaminase; Transpeptid., transpeptidase; TSH, thyroid stimulating hormone; WBC, white blood cell count.

Extended Data Table 1 | Genes sequenced by cRNA bait pull-down in the validation cohort

GNB1	FBXW7	CUL2	RAD51	CBLC
CSF3R	IRF1	CDH23	IDH2	U2AF2
MPL	CSF1R	PTEN	CREBBP	ASXL1
NRAS	NPM1	SMC3	SMG1	PTPRT
NOTCH2	PHACTR1	HRAS	CBFB	GNAS
RIT1	DAXX	WT1	CTCF	RUNX1
CACNA1E	PHIP	SF1	SMPD3	U2AF1
FAM5C	MYB	EED	PRPF8	CSF2RB
DNMT3A	FNDC1	CNTN5	TP53	CBX7
ASXL2	CUX1	MLL	NF1	EP300
SF3B1	MLL5	CBL	SUZ12	ZRSR2
IDH1	LUC7L2	ETV6	STAT5B	BCOR
CUL3	BRAF	KRAS	KANSL1	KDM6A
GIGYF2	CUL1	MLL2	DCAF7	GATA1
CBLB	EZH2	PRPF40B	SRSF2	SMC1A
GATA2	MLL3	PPFIA2	ASXL3	PHF8
STAG1	RAD21	SH2B3	SETBP1	MED12
PIK3CA	MYC	PTPN11	DNMT1	ATRX
FRYL	JAK2	FLT3	EPOR	RPS6KA6
KIT	CDKN2A	PDS5B	JAK3	DIAPH2
UGT2A3	HNRNPK	DCLK1	CEBPA	STAG2
TET2	NOTCH1	RB1	ZFP36	PHF6

nature research

Corresponding author(s): Liran I. Shlush

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

Sample size was determined by the availability of the samples from EPIC (European Prospective Investigation into Cancer and Nutrition) while assuring large enough cohort for the reported statistical analyses. (124 pre-AML samples and 676 controls). As detailed in Methods, we used weighting to minimise the biases introduced by the artificial case-control ratio and calculated hazard ratios relative to the (approximate) true cumulative AML incidence of about 1-3/1,000 in the given age range over a follow up of 10-20 years.

2. Data exclusions

Describe any data exclusions.

For the model described in figure 3 we have excluded samples taken less than 6 months prior AML diagnosis from model training and validation in order to avoid skewing the model towards significance (as per reviewer comments). This resulted in the removal of 4 individuals from the discovery cohort, leaving 91 individuals in the discovery cohort pre-AML group.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Experimental replication was not attempted but rather validated in a second independent cohort

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

AML patients were identified based on the following ICD9 codes: 9861/3 9860/3 9801/3 9866/3 9891/3 9867/3 9874/3 9840/3 9872/3 9895/3 9873/3, which included only cases of de novo AML, and no secondary AML. Age and gender matched individuals that were not diagnosed for any hematological disorders during the average follow-up period of 12.6 were selected as the control group,

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

For assessment of statistical significant differences between the cases as controls blinding is impractical as the groups must be defined. Machine learning algorithms were blind in a sense as they been configure to be trained and tested in different set of samples.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	1) Sequencing reads were aligned using BWA-aln and BWA-mem 2) indel-realignment was done using GATK 2) Variant calling was done using deepSNV, Varscan2, Pindel v2.2 and CaVEMan v1.11.2. All of these algorithms are extensively described and validated in the published literature. 3) Code used to generate predictive models will be deposited on GitHub at the time of manuscript publication For full description please refer to the appropriate method sections

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique material were used in this study

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used in this study
b. Describe the method of cell line authentication used.	No cell lines were used in this study
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used in this study
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used in this study

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used in this study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

For the discovery cohort, samples were collected from individuals upon enrollment into the EPIC study between 1993 and 1998.
95 individuals (39 males and 56 females) who developed AML an average of 6.26 years (IQR=4.88 years) after the sample was collected were included in the pre-AML group. 414 age and gender (167 males and 247 females) matched individuals were selected for the control group, as they did not develop any hematological disorders during the average follow-up period of 11.6 years (IQR=2.13 years). The median age at recruitment was 56.75 years (range, 36.08 to 74.42)

For the validation cohort, samples were ascertained from participants in the EPIC-Norfolk longitudinal cohort study enrolled between 1994 and 2010.
37 individuals (15 males and 22 females) who developed AML an average of 10.5 years from first sampling (IQR=8.3 years) were included in the pre-AML group. 262 age and gender (135 males and 127 females) matched individuals were selected for the control group, as they did not develop any hematological disorders during the average follow-up period of 17.5 years from first sampling (IQR=3.8 years). The median age at recruitment was 65.05 years (range, 43.9 to 88.1)