



Combined landscape of single-nucleotide variants and copy number alterations in clonal hematopoiesis

Ryunosuke Saiki¹, Yukihide Momozawa², Yasuhito Nannya¹, Masahiro M. Nakagawa^{1,3}, Yotaro Ochi¹, Tetsuichi Yoshizato¹, Chikashi Terao⁴, Yutaka Kuroda⁵, Yuichi Shiraishi⁶, Kenichi Chiba⁶, Hiroko Tanaka⁷, Atsushi Niida⁸, Seiya Imoto⁹, Koichi Matsuda¹⁰, Takayuki Morisaki¹¹, Yoshinori Murakami¹¹, Yoichiro Kamatani^{4,10}, Shuichi Matsuda⁵, Michiaki Kubo¹², Satoru Miyano⁷, Hideki Makishima¹ and Seishi Ogawa^{1,3,13} ✉

Clonal hematopoiesis (CH) in apparently healthy individuals is implicated in the development of hematological malignancies (HM) and cardiovascular diseases. Previous studies of CH analyzed either single-nucleotide variants and indels (SNVs/indels) or copy number alterations (CNAs), but not both. Here, using a combination of targeted sequencing of 23 CH-related genes and array-based CNA detection of blood-derived DNA, we have delineated the landscape of CH-related SNVs/indels and CNAs in 11,234 individuals without HM from the BioBank Japan cohort, including 672 individuals with subsequent HM development, and studied the effects of these somatic alterations on mortality from HM and cardiovascular disease, as well as on hematological and cardiovascular phenotypes. The total number of both types of CH-related lesions and their clone size positively correlated with blood count abnormalities and mortality from HM. CH-related SNVs/indels and CNAs exhibited statistically significant co-occurrence in the same individuals. In particular, co-occurrence of SNVs/indels and CNAs affecting *DNMT3A*, *TET2*, *JAK2* and *TP53* resulted in biallelic alterations of these genes and was associated with higher HM mortality. Co-occurrence of SNVs/indels and CNAs also modulated risks for cardiovascular mortality. These findings highlight the importance of detecting both SNVs/indels and CNAs in the evaluation of CH.

The presence of clonal components in an apparently normal hematopoietic compartment, or CH, has drawn increasing attention in recent years^{1,2}. Although suggested only indirectly by skewed chromosome X inactivation in early studies^{3–7}, CH has recently been demonstrated by detection of CNAs in peripheral blood samples from large cohorts of individuals without blood cancer using single-nucleotide polymorphism (SNP) array data from genome-wide association studies (GWAS)^{8–11}. Showing a substantial overlap with those characteristic of HM, CNAs were shown to be associated with an elevated risk of developing HM^{8,9}. More recently, CH has also been detected by the presence of somatic SNVs/indels in the peripheral blood of apparently healthy individuals^{12–15} and cancer patients^{16,17} using next-generation sequencing. In addition to its link to HM, CH as detected by SNVs/indels has been highlighted by its unexpected association with a significantly increased risk for cardiovascular diseases (CVD)^{12,13,18,19}.

Regardless of the type of genetic lesions by which it is detected, CH is strongly age related with an increasing frequency in the elderly^{8–13}. With substantially improved technologies available to

identify CNAs and somatic SNVs/indels, a complete registry of CNAs and SNVs/indels associated with CH has been elucidated, and these are thought to involve virtually every individual among the most elderly^{20,21}. However, to date, no studies have evaluated CNAs and SNVs/indels together at comparable sensitivity in a large cohort of a general population, although they have recently been investigated in a cancer population in which many patients had been treated with chemo-/radiotherapy²². What is the landscape of CH recognized by combining both CNAs and SNVs/indels in a general population? Are there any interactions between SNVs/indels and CNAs that shape the landscape of CH? How are hematological phenotypes affected by both CH-related lesions? How does it affect HM and CVD risks? These are the key questions to be answered for a better understanding of CH and its implication in HM and CVD.

In the present study, for the purpose of delineating the combined landscape of driver SNVs/indels and CNAs in CH, we performed SNP array-based copy number analysis and targeted sequencing of major CH-related genes on blood-derived DNA from BioBank Japan (BBJ)²³ and which had been SNP-typed for GWAS studies for

¹Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan. ²Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ³Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan. ⁴Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ⁵Department of Orthopaedic Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan. ⁶Division of Cellular Signaling, National Cancer Center Research Institute, Tokyo, Japan. ⁷Department of Integrated Data Science, M&D Data Science Center, Tokyo Medical and Dental University, Tokyo, Japan.

⁸Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ⁹Division of Health Medical Intelligence, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ¹⁰Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. ¹¹Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ¹²RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ¹³Department of Medicine, Centre for Haematology and Regenerative Medicine, Karolinska Institute, Stockholm, Sweden. ✉e-mail: sogawa-tyk@umin.ac.jp

common diseases, including hypertension, diabetes, autoimmune diseases and several solid cancers²³. We then investigated the combined effect of both CH-related lesions on clinical phenotypes and outcomes, particularly that on the mortality from HM and CVD.

Results

Identification of CH-related SNVs/indels and CNAs. We enrolled a total of 11,234 subjects from the BBJ cohort ($n=179,417$) in which SNP array analysis of peripheral blood-derived DNA had been performed for large-scale GWAS studies for common diseases (Supplementary Tables 1 and 2) (https://biobankjp.org/info/pdf/sample_collection.pdf)²³. Among these, 10,623 were randomly selected from 60,787 cases aged ≥ 60 years at the time of sample collection and were confirmed not to have solid cancers as of March 2013. This randomly selected set included 61 cases who were known to have developed and/or died from HM as of March 2017. The remaining 611 consisted of all cases from the entire BBJ cohort who were confirmed as having developed and/or died from HM as of the same date but were not included in the 10,623 cases randomly selected. In total, 672 cases in the entire BBJ cohort were reported to have HM: 215 myeloid, 420 lymphoid and 37 lineage-unknown tumors (Extended Data Fig. 1a). For these 11,234 cases, SNVs/indels in blood were investigated using multiplex PCR-based amplification of exons of 23 CH-related genes, followed by high-throughput sequencing (Methods).²⁴ The sensitivity of SNV detection according to in silico simulations using known SNPs was $>94\%$ for 3% variant allele frequency (VAF) and $>74\%$ for 2% VAF, but $<20\%$ for 1% VAF with a mean depth of $\sim 800\times$ (Supplementary Fig. 1a,b).

In total, we called 4,056 SNVs/indels (2,750 SNVs and 1,306 indels) in 3,071 (27.3%) subjects, of which 2,312 (20.6%) had one, 586 (5.2%) two and 173 (1.5%) at least three SNVs/indels (Fig. 1a). VAF values were widely distributed, from 0.5 to 85.6% with a median of 3.0% (Supplementary Fig. 1c). Age dependence of CH-related SNVs/indels was evident (Fig. 1b). In accordance with previous reports, *DNMT3A* (13.5%), *TET2* (9.5%), *ASXL1* (2.2%) and *PPM1D* (1.4%) were most frequently mutated (Extended Data Fig. 2a,c). Several combinations of genes, including *TET2/DNMT3A*, *ASXL1/TET2*, *ASXL1/CBL*, *SRSF2/TET2* and *SRSF2/ASXL1*, were more frequently comutated than what would be expected by chance only (odds ratio (OR) = 1.53–6.53, $q < 0.05$; Extended Data Fig. 2d). Of interest, many of these combinations are also comutated in myeloid neoplasms with high VAF values^{25–27}, suggesting the presence of these combinations of SNVs/indels in the same cell fraction. This was also expected for some cases having a large ($>50\%$) sum of VAFs of relevant SNVs/indels ('pigeonhole principle'),²⁸ although it was not determined whether these combinations of SNVs/indels affected the same cell populations in the vast majority of cases (Extended Data Fig. 2e–i).

Copy number alteration data were available from a previous study²¹ in which SNP array-based copy number detection in blood-derived DNA was performed for a larger cohort of BBJ cases ($n=179,417$), including all cases enrolled in the current study ($n=11,234$). In total, 2,797 CNA-positive regions/segments were identified in 2,254 (20.1%) cases (Extended Data Fig. 3 and Methods), of which 413 (3.7%) had multiple CNAs (Fig. 1a). Reflecting a higher age distribution of the current cohort, the frequency of CNAs was greater than that in the entire BBJ cohort²¹, even though age-stratified frequencies were almost equivalent between both cohorts (Fig. 1b). The size of detected CNAs ranged from 0.01 to 248 Mb (median, 34.4), depending on the density of informative SNPs and their haplotype configuration, tumor contents and the performance of SNP probes (Supplementary Fig. 1d). Estimated mutant cell fractions for CNAs ranged from 0.2 to 93.2%, with a median of 2.0% and false discovery rate (FDR) < 0.05 , where a substantial number ($n=461$) of CNAs was seen in a cell fraction

of $\leq 1\%$, which was below the limit of detection for SNVs/indels. Thus, smaller clones were detected through CNAs, particularly copy-neutral loss of heterozygosity (LOH) or uniparental disomy (UPD), compared with those detected through SNVs/indels (Supplementary Fig. 1c).

We found 27 significantly recurrent CNAs, many of which are also commonly seen in HM, supporting a pathogenic link between CH and leukemogenesis (Extended Data Fig. 4a–c). In accordance with previous reports^{8–11}, 14qUPD, +21q, del(20q) and +15q were among the most frequent CNA lesions (Extended Data Fig. 2b,c) while del(20q), 16pUPD and 17pUPD showed the largest mean clone size (Supplementary Fig. 2). Several CNAs, including 14qUPD and +21, showed higher frequencies than those reported in Western populations, which is probably due to a higher sensitivity for detection of CNAs in this study compared with that in previous studies in Western populations^{8–11}; when confined to lesions with $\geq 5\%$ cell fractions, the difference across studies becomes less conspicuous for many CNA targets (Extended Data Fig. 4d,e). Nevertheless, even considering varying sensitivity, several CNAs, including +15, del(14q), del(9q), del(20q) and del(13q), still showed a different frequency across studies in both populations²¹, suggesting an ethnic difference in positive selection of CH-related CNAs (Extended Data Fig. 4e), although the exact genetic basis of this difference is largely unclear for most CNAs.

Combined landscape of SNVs/indels and CNAs. When SNVs/indels and CNAs were combined, CH was demonstrated in 4,242 (40%) of 10,623 randomly selected cases ≥ 60 years of age and with no reported cancer history, and in 376 (56%) of 672 cases who developed HM, where 38 of the 376 were < 60 years of age. Combining both lesions, more subjects ($n=1,503$) had two or more lesions than judged by SNVs/indels ($n=759$) or by CNA alone ($n=413$) (Fig. 1a). The frequency of CH and the total number of CH-related lesions, as well as the maximum estimate of clone size in CH⁺ cases, were significantly higher in individuals with abnormal blood counts, particularly in those with cytopenias, compared with those with completely normal blood counts, depending on the number of blood lineages involved (Fig. 1c,d). A similar landscape of combined CH lesions was observed in an independent cohort of 8,023 patients with solid cancer from The Cancer Genome Atlas (TCGA; <https://portal.gdc.cancer.gov/>), although the sensitivity of CH lesions, particularly CNAs, was substantially lower than in the current study due to a lower coverage of exome sequencing and less accurate haplotype phasing required for sensitive CNA detection (Extended Data Fig. 5a–c).

Accounting for 7% of the total cohort and 16% of all CH⁺ cases, 740 individuals harbored both types of lesion, which was significantly more frequent than expected by chance only (Extended Data Fig. 2j), even after adjustment for age (OR = 1.3, $P=0.0003$, age-stratified permutation test; Supplementary Fig. 3 and Methods). SNVs/indels in *TP53*, *TET2*, *JAK2*, *SF3B1* and *U2AF1*—and less significantly in *DNMT3A*, *CBL* and *SRSF2*—were accompanied by significantly more CNAs (Supplementary Fig. 4). The number of cases with multiple CH-related lesions was also significantly higher than expected from the number of all CH-related lesions ($P=0.0067$). The significantly higher frequency of cases with both SNVs/indels and CNAs ($P < 0.0001$) and those with multiple lesions ($P < 0.0001$) was confirmed in the TCGA cohort. These observations raise the possibility that it might be the total number of lesions, rather than the combination of SNVs/indels and CNAs, that is relevant to the positive selection in CH in which multiple CH-related lesions in the same cell contributed to positive selection in a substantial number of cases with multiple CH lesions. In support of this, the maximum clone size in CH⁺ cases significantly correlated with the total number of CH-related SNVs/indels and CNAs, but not with their combinations (Fig. 1e).

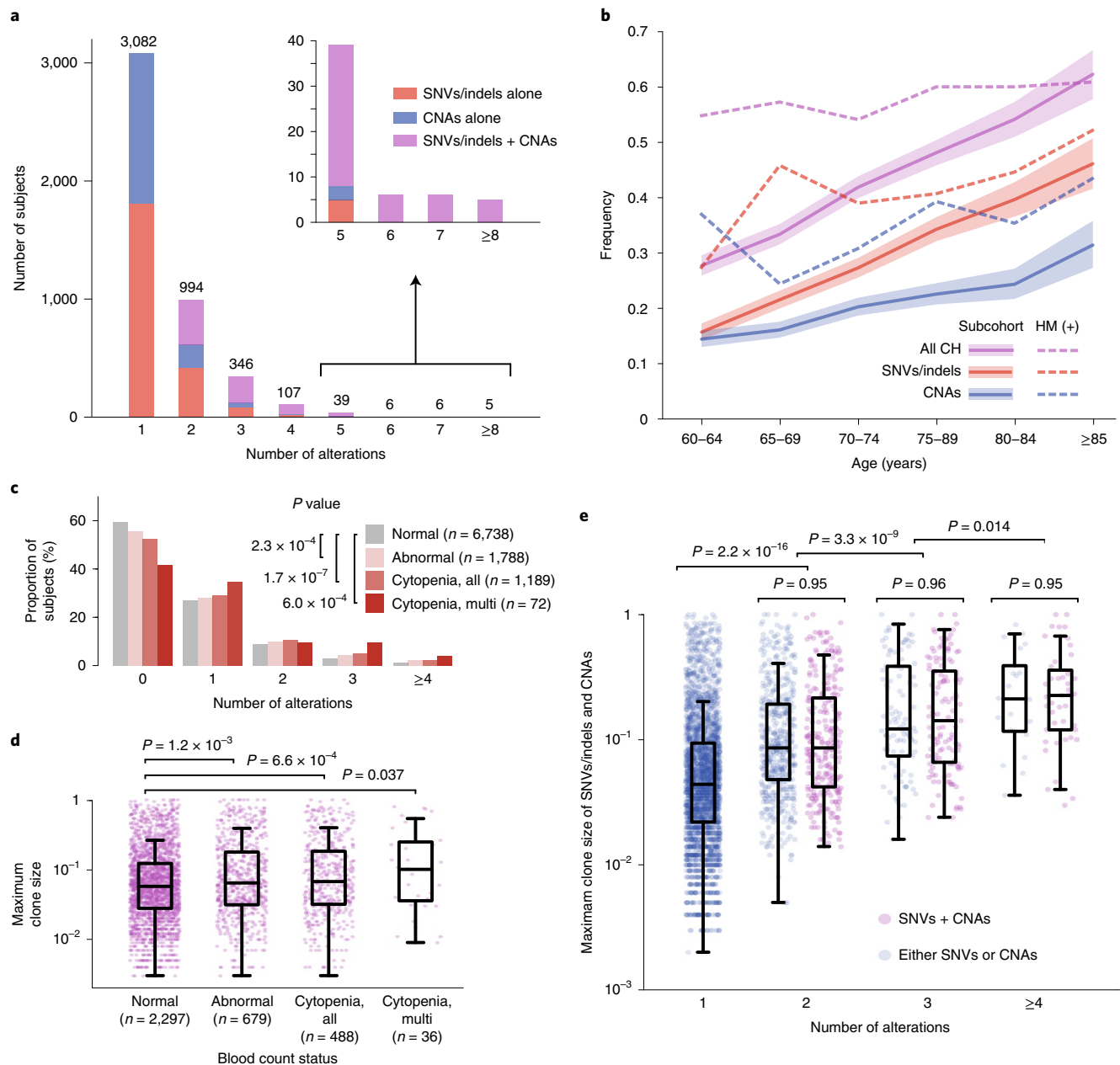


Fig. 1 | Landscape of SNVs/indels and CNAs in clonal hematopoiesis. a, Distribution of the number of genetic alterations in each subject. Subjects with SNVs/indels alone, with CNAs alone or with both are illustrated by different colors. **b**, The prevalence of CH-related SNVs/indels and CNAs according to age. Solid and broken lines indicate frequencies in a randomly extracted subcohort ($n = 10,623$) and in subjects with HM events ($n = 672$), respectively. Colored bands represent 95% confidence intervals (CIs). **c**, Number of co-occurring alterations in subjects with abnormalities in blood cell counts, or with cytopenia. **d**, Maximum cell fraction of CH-related alterations in CH⁺ subjects with or without abnormalities in blood cell counts. **e**, Dot plot of maximum cell fractions of SNVs/indels or CNAs across different numbers of co-occurring alterations. Cell fractions of SNVs/indels are defined as twofold VAF. Those with both SNVs/indels and CNAs are shown in purple, while those with either are shown in blue. **d,e**, Unclassifiable CNAs were excluded because we cannot calculate their precise cell fractions. Box plots indicate the median, first and third quartiles (Q1 and Q3) and whiskers extend to the furthest value between Q1 – 1.5x the interquartile range (IQR) and Q3 + 1.5x IQR. **c–e**, P values were calculated by two-sided Wilcoxon rank-sum test and not adjusted for multiple comparison.

Co-occurring multiple lesions were judged to be present in the same cell in 73 cases on the basis of their large (>1.0) clone size sum²⁸, of which eight were combinations of SNVs/indels and CNAs (Extended Data Fig. 2k). In the vast majority of cases we could not determine the cellular compartment of multiple lesions due to the small clone size of both lesions, which would be better addressed using single-cell-based sequencing. A representative case is shown in Supplementary Fig. 5, in which the presence of both

del(13q) and a *TET2*-involving SNV in the same cell compartment of myeloid lineages was demonstrated using single-cell sequencing (Supplementary Fig. 5a–d). Some combinations of SNVs/indels and CNAs were significantly more frequently observed than expected by chance only (Fig. 2a). Of particular interest among these were co-occurring SNVs/indels and CNAs affecting the same gene/locus. Overall, we found 88 cases having co-occurring SNVs/indels and CNAs affecting eight genes/loci (Extended Data Fig. 6a), among

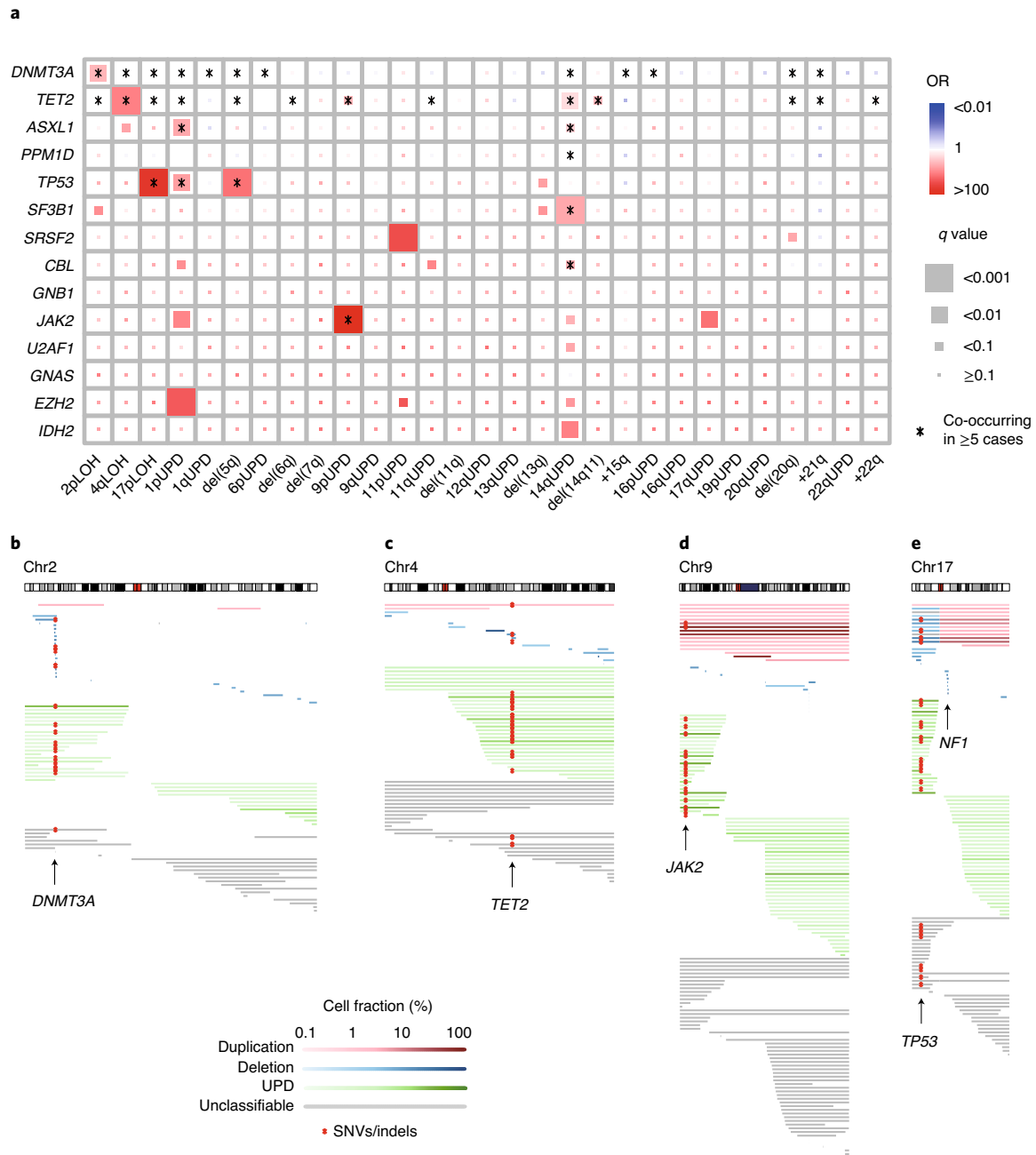


Fig. 2 | Co-occurrences of SNVs/indels and CNAs in clonal hematopoiesis. a, Correlations between individual SNVs/indels and CNAs. Rectangle size indicates the significance of correlations; red and blue rectangles represent positive and negative correlations, respectively. Combinations of SNVs/indels and CNAs seen in five or more subjects are indicated by asterisks. **b–e**, Distributions of CNAs on chromosomes 2 (**b**), 4 (**c**), 9 (**d**) and 17 (**e**). Horizontal bars represent CNAs, and co-occurring SNVs/indels in *DNMT3A*, *TET2*, *JAK2* and *TP53* are indicated by red asterisks. Colors of horizontal bars represent the types and cell fractions of CNAs. Allele imbalances that could not be classified as UPD, deletion or duplication are indicated as unclassifiable CNAs (gray).

which the most frequently involved were *TP53* (with 17pLOH) ($n=24$, $OR=60.6$, $q<0.001$), *TET2* (with 4qLOH; $n=22$, $OR=10.8$, $q<0.001$), *JAK2* (with 9pLOH/gain; $n=18$, $OR=414$, $q<0.001$) and *DNMT3A* (with 2pLOH; $n=16$, $OR=4.02$, $q=0.001$), which were also found in the TCGA cases (Fig. 2a–e and Extended Data Fig. 5d). These co-occurrences were still statistically significant when the increase in VAF caused by LOH was taken into account (Supplementary Fig. 6). In reality, more cases are expected to have these combinations because there were many ‘isolated’ LOH lesions or allelic imbalances affecting these loci that lacked accompanying SNVs/indels ($n=64$) (Fig. 2b–e), which were thought to have

escaped detection due to the lower detection sensitivity for SNVs/indels than CNAs (Supplementary Fig. 1a,c). In fact, using highly sensitive digital droplet PCR (ddPCR) assay targeting mutational hotspots, SNVs in *JAK2* and *TP53* were confirmed in eight out of 44 and 22 out of 37 samples with isolated LOH at 9p and 17p, respectively (Supplementary Fig. 7). Representing well-known mechanisms of biallelic alterations of the relevant driver genes in myeloid malignancies, these combinations of lesions in CH are predicted to affect the same cell, being involved even in the very early stages of positive selection in myeloid leukemogenesis^{29–31}. SNVs/indels were most frequently associated with LOH when they affected *TP53* and

JAK2 in both myeloid malignancies^{32,33} and CH (Extended Data Fig. 6b), also supporting their role in the mechanism of biallelic alterations. Unfortunately, either none of these cases satisfied the pigeonhole principle or no samples were available for single-cell sequencing analysis to directly confirm this at the single-cell level. However, in the case of SNVs/indels associated with UPD, their presence in the same cell compartments in many cases was supported by a highly skewed distribution of mutant cell fractions of both lesions (Supplementary Fig. 8 and Methods).

Besides SNVs/indels and CNAs affecting the same gene/locus, we also detected a significant combination between SNVs/indels in *TET2* and microdeletions of the *TCRA* (involving 14q11.2) locus ($n=7$, $OR=3.53$, $q=0.059$), of which one case was reported to develop T-cell lymphoma (Fig. 2a and Extended Data Fig. 2l). This combination is of potential interest, given that *TET2* is frequently mutated in mature T-cell lymphomas³⁴, particularly in follicular-helper-T-cell-derived lymphomas such as angio-immunoblastic T-cell lymphoma, which are also seen in *Tet2* knockdown mice³⁵. Other potentially relevant combinations included *SF3B1*/14qUPD, *TET2*/14qUPD, *ASXL1*/1pUPD, *TP53*/1pUPD and *TP53*/del(5q) (Fig. 2a), whose biological significance, however, is largely unclear except for the interplay between del(5q) and mutated *TP53* intensively studied in myelodysplastic syndromes (MDS)^{36,37}.

Clinical associations with CH. Next, we investigated common demographic factors that might influence CH-related SNVs/indels and CNAs and the effect of both CH lesions on clinical features and outcomes. In addition to the large effect of age, several factors impacting on CNAs and/or SNVs/indels were observed. Male gender and smoking were significantly associated with SNVs/indels in *ASXL1*, *PPM1D*, splicing factors and *TP53*; and with CNAs, particularly +15, del(20q), +21 (with male gender) and 14qUPD (with smoking), many of which remained significant in multivariate analysis (Fig. 3a). The effect of alcohol consumption was less prominent and mostly confined to an increased incidence of del(20q). Although none of the subjects in our cohort had been diagnosed with HM at the time of sample collection, 1,314 cases had varying degrees of abnormal blood counts (Supplementary Table 3). Even though the landscape of CH in these cytopenic individuals at a glance was largely similar to that in noncytopenic individuals (Extended Data Fig. 7a), cytopenic cases exhibited a significantly high frequency of CH where the frequency significantly correlated with the severity of cytopenia (Fig. 1c). In particular, individuals with abnormally high platelet count had a higher frequency of *JAK2*-involving SNVs/indels and 9pUPD ($OR=50.5$, $q<0.001$ and $OR=26.0$, $q=0.0017$, respectively), while *U2AF1*-involving SNVs/indels and del(20q) were more common in those with cytopenia of any type ($OR=7.39$, $q<0.001$ and $OR=3.10$, $q=0.015$, respectively) (Extended Data Fig. 7b). Individuals with CH-related SNVs/indels had a higher frequency of cytopenia and exhibited lower hemoglobin and mean corpuscular hemoglobin concentration (MCHC) values, while CNAs were associated with lower white blood cell (WBC) and platelet counts and higher mean corpuscular volume (MCV) values (Fig. 3b). The number of all co-occurring alterations, SNVs/indels or CNAs, and VAF of SNVs/indels predicted significantly lower hemoglobin values, while mutant cell fractions of CNAs predicted higher MCV and lower MCHC values (Fig. 3b,c and Extended Data Fig. 7c). In regard to individual alterations, SNVs/indels in *JAK2* were significantly correlated with high platelet count ($q<0.001$), even when the analysis was confined to individuals with normal blood count. Moreover, we found significant associations of lower hemoglobin values with SNVs/indels in *TP53*, *PPM1D*, *SF3B1* and *U2AF1*, and in 4qUPD and del(20q), while SNVs/indels in *PPM1D*, *U2AF1*, 6pUPD and del(20q) were associated with lower platelet counts and SNVs/indels in *TP53* and *SF3B1*, and 11qUPD

correlated with higher MCV (Fig. 3b and Extended Data Fig. 7c). VAF or cell fractions of SNVs/indels and CNAs were also predictive of changes in hemoglobin, platelet counts or MCV (Fig. 3b and Extended Data Fig. 7d), while VAF of *JAK2*-involving SNVs/indels did not correlate with platelet counts (Supplementary Fig. 9). SNVs/indels in *TET2* alone were not associated with reduced hemoglobin values (Fig. 3b). However, interestingly, we observed a significant association of lower hemoglobin values with multiple SNVs/indels in *TET2* and any allelic imbalance affecting 4q, which is most probably attributable to biallelic *TET2* alterations (Fig. 3b,d). We also tested the relationships between CH and values of other blood tests, revealing a negative correlation between *GNB1*-involving SNVs and uric acid concentration with $FDR<0.1$ (Supplementary Fig. 10).

Effect of SNVs/indels and CNAs on HM mortality. Among the major interests in the current study is the effect of SNVs/indels and CNAs on the risk of HM, particularly the combined effect of both CH-related lesions. To determine this, we investigated the effect of CH on the cumulative mortality from HM using Fine-Gray regression modeling in a case-cohort design³⁸, where 7,937 of the 10,623 cases were regarded as a subcohort randomly selected from 43,662 cases who had been followed up for survival and cause of death on the basis of vital statistics from Japan³⁹ (Extended Data Fig. 1b). The median follow-up of these cases was 10.4 years (range, 0.01–13.5), during which 401 HM deaths were confirmed (Extended Data Fig. 1b). Age, sex and versions of SNP array were adjusted, and deaths from any causes other than HM were analyzed as competing risks.

In accordance with previous reports^{8,9,12,13,22}, both SNVs/indels and CNAs were significantly associated with a higher mortality from HM than observed in CH[−] cases (Supplementary Fig. 11), with an estimated cumulative 10-year mortality of 1.28 and 1.32%, respectively (Fig. 4a). The difference in HM mortality between CH⁺ and CH[−] subjects was mostly explained by CH itself, although age and gender made a smaller contribution (Extended Data Fig. 8a). Although lymphoid neoplasms accounted for two-thirds of all HM mortality in the cohort of 43,662 elderly cases attributable mortality in CH⁺ versus CH[−] cases was ~twofold higher from myeloid (0.39%) than from lymphoid (0.21%) neoplasms (Fig. 4b), and the hazard ratio (HR) between CH⁺ and CH[−] cases was >2.5-fold higher for myeloid (3.64) than for lymphoid (1.36) neoplasms. This suggests the predominant effects of CH on myeloid neoplasms, which is in line with the fact that most CH-related lesions target driver genes in myeloid neoplasms. The number of SNVs/indels and CNAs and the total number of CH-related lesions all significantly correlated with higher HM mortality (Fig. 4c and Extended Data Fig. 8b,c). While the maximum clone size of CH-related lesions correlated with the number of CH-related lesions (Fig. 1e), the former was also significantly associated with higher HM mortality independently of the latter (Figs. 4d and Fig. 5a), which is in line with a previous observation that SNVs/indels correlated with development of HM only when they exhibited sufficiently high VAF values ($\geq 1\%$)⁴⁰. Using univariate analysis, the highest risk of HM mortality was conferred by SNVs/indels of *U2AF1*, *EZH2*, *RUNX1*, *SRSF2*, *TP53* and +1q^{11,14,21} (Fig. 5b–d and Supplementary Figs. 12 and 13). As expected from a ~twofold higher attributable mortality for myeloid than lymphoid malignancy, HR and OR were higher in myeloid than in lymphoid HM for most lesions with the exception of trisomy 12, which was associated with lymphoid, but not myeloid, neoplasms (Extended Data Fig. 9a). The impact of CH on HM mortality was more prominent when present in combination with abnormal blood counts, particularly cytopenia. A significantly higher HM mortality associated with CH was observed in subjects with abnormality in blood counts than in those without (Fig. 4e), depending on both the number of CH-related lesions and the severity of cytopenia: a 10-year HM mortality as high as 3.4%

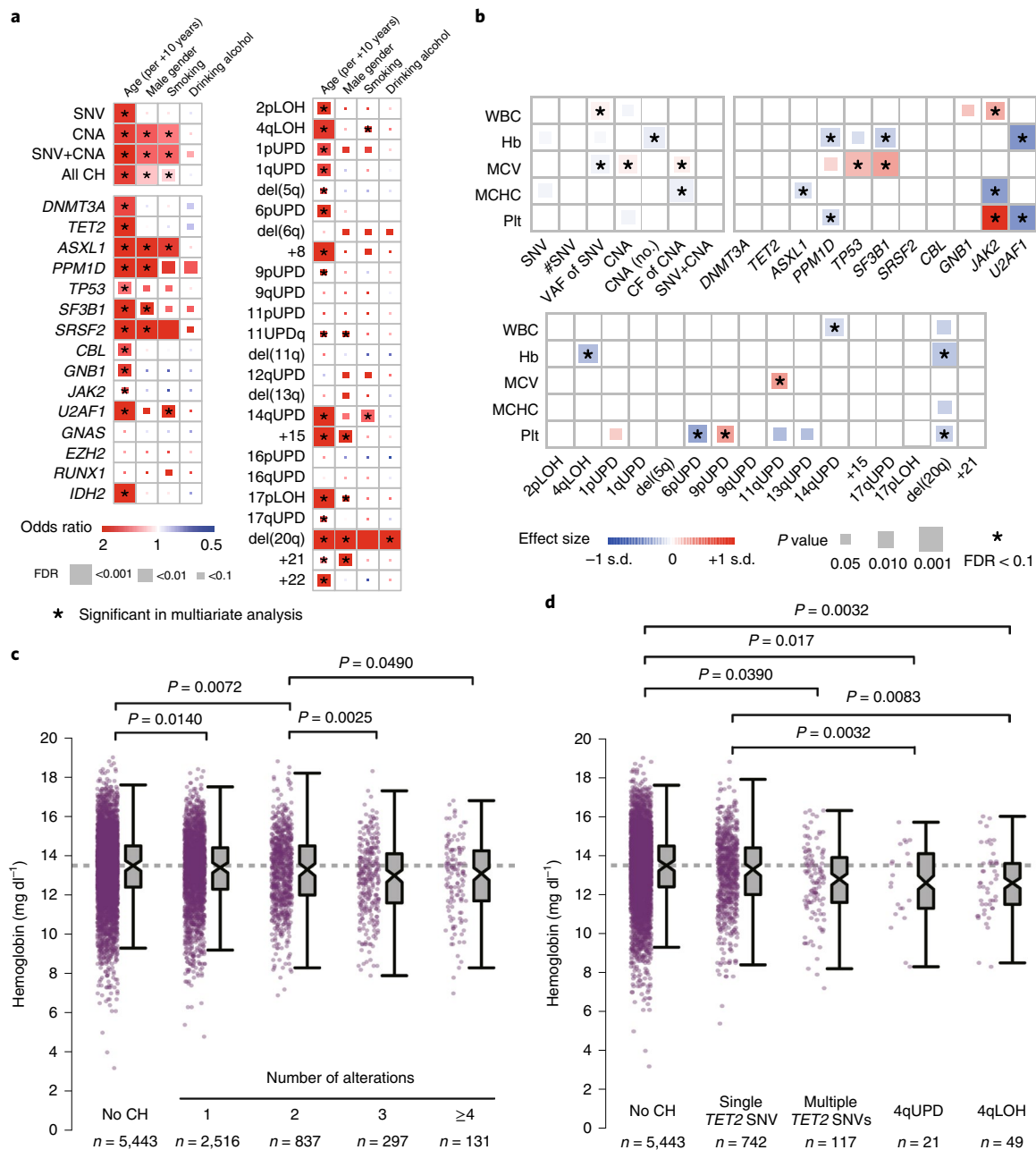


Fig. 3 | Risk factors for CH and effects on blood counts. a, Correlations of genetic alterations with age, male gender, history of smoking and drinking alcohol. Sizes and colors of rectangles represent, respectively, the significance and effect size, as calculated by two-sided Wald test. Asterisks indicate the clinical factors significantly correlated with each alteration in multivariate logistic regression ($P < 0.05$). Age (per +10 years) indicates for every additional 10 years of age. **b**, Correlations between genetic alterations and blood counts. The sizes and colors of rectangles indicate, respectively, the significance and effect size of correlation. P values were calculated by two-sided t -test based on multivariate models including age and gender as covariates. Correlations significant after correction for multiple testing ($FDR < 0.1$) are indicated by asterisks. Hb, hemoglobin; Plt, platelets. **c**, Distribution of hemoglobin in subjects with different numbers of alterations. **d**, Distribution of hemoglobin in subjects with no alterations, with single SNV/indel in *TET2* (single *TET2* SNV), multiple SNVs/indels in *TET2* (multiple *TET2* SNVs), with 4qUPD or with any loss of heterozygosity in 4q are illustrated in dot plots and box plots. P values were calculated by two-sided t -test based on multivariate linear regression models including age and gender as covariates (**b,d**) or by two-sided Wilcoxon rank-sum test (**c**), and not adjusted for multiple comparison. In all box plots, the median, Q1 and Q3 are indicated, and whiskers extend to the furthest value between Q1 - 1.5×IQR and Q3 + 1.5×IQR. The number of subjects in each category is shown below the box plots.

was observed for those with multilineage cytopenia and multiple CH-related SNVs/indels and CNAs, compared with 0.46% for those with normal blood count and lacking CH-related lesions.

The presence of both SNVs/indels and CNAs was associated with significantly increased HM mortality compared with that of SNVs/indels (HR = 2.84, 95%CI: 2.14–3.78) or CNA (HR = 2.64, 95%CI:

1.94–3.60) alone (Fig. 4f). This was observed even when subjects were stratified according to the number of SNVs/indels (Extended Data Fig. 8d–f). However, the combined effect seems to be explained in large part by an increased total number of alterations rather than the type of lesions co-occurring—that is, SNVs/indels versus CNAs. In fact, HM mortality significantly correlated with the total number

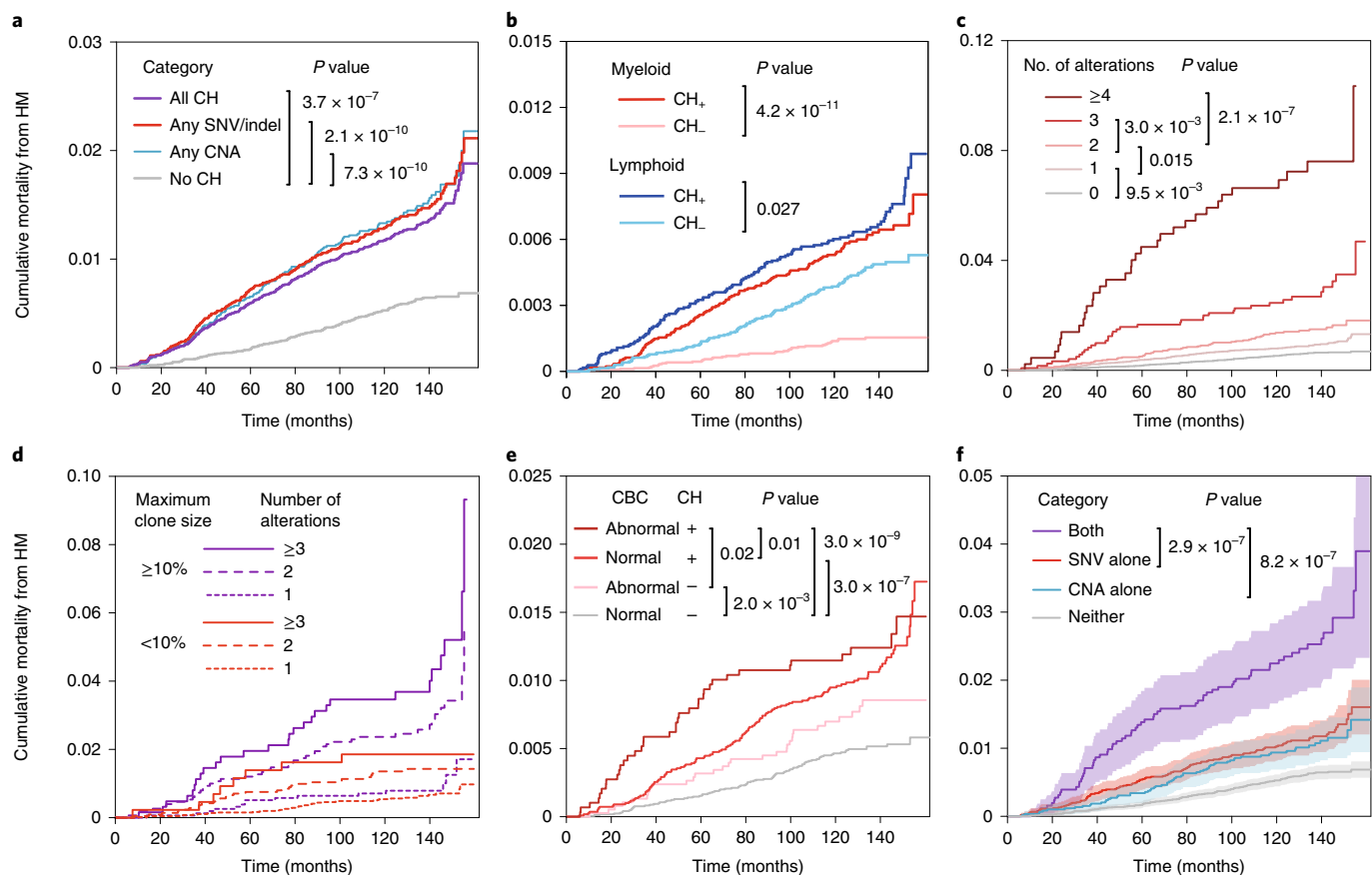


Fig. 4 | Impact of CH on mortality from HM. **a**, Cumulative mortality from HM in subjects with any CH ($n=3,336$), any SNV/indel ($n=2,237$), any CNA ($n=1,613$) or without CH ($n=4,947$). **b**, Cumulative mortality from myeloid and lymphoid malignancies in subjects with or without CH. **c**, Cumulative mortality from HM in subjects with different numbers of CH-related alterations (0, $n=4,947$; 1, $n=2,263$; 2, $n=722$; 3, $n=246$; ≥ 4 , $n=105$). **d**, Cumulative mortality from HM in subjects with different numbers of co-occurring alterations and maximum clone sizes ($<10\%$ or $\geq 10\%$). Cell fractions of unclassifiable CNAs were regarded as $<10\%$. **e**, Cumulative mortality from HM in subjects with CH and abnormalities in complete blood count (CBC) ($n=550$), with CH alone ($n=2,065$), with abnormalities in CBC alone ($n=703$) or with neither ($n=3,094$). **f**, Solid lines indicate cumulative mortality from HM in subjects with both SNVs/indels and CNAs ($n=514$), SNVs/indels alone ($n=1,723$), CNAs alone ($n=1,099$) and with no alterations ($n=4,947$). Colored bands indicate 95% confidence intervals. **a–c, f**, P values were calculated by two-sided Wald test based on multivariate regression models; **e**, P values were calculated by two-sided log-rank test stratified by age (≤ 70 or >70 years) and gender because of nonproportional hazards. P values were not adjusted for multiple comparison.

of CH-related lesions while the co-occurrence of both lesions did not significantly affect the mortality of individuals having the same number of lesions (Extended Data Fig. 8g–i). Many of the SNVs/indels conferring a higher HM mortality, including those affecting *U2AF1*, *SRSR2*, *TP53* and *JAK2*, tended to have a higher total number of CH-related lesions compared with other SNVs/indels (Extended Data Fig. 2m). Nevertheless, the effect on HM mortality was not uniform across different combinations of SNVs/indels and CNAs, regardless of the total number of lesions. In particular, those involving the same gene/locus were associated with higher HM mortality compared with other combinations of SNVs/indels and CNAs (Extended Data Fig. 6c). Increased mortality was largely explained by those affecting *TP53*. However, even with exclusion of *TP53*-involving SNVs/indels and CNAs, the combinations of lesions affecting the same locus showed higher HM mortality than other SNV/indel and CNA combinations. Of interest, *TP53*-involving SNVs/indels also exhibited significant associations with del(5q) and multiple (≥ 3) CNAs mimicking a complex karyotype (Fig. 2a and Extended Data Fig. 6f) which, together with 17pLOH, are among the lesions most commonly associated with *TP53* alterations in a variety of myeloid neoplasms carrying a very poor prognosis,

particularly in MDS^{25,33,41}. In agreement with this, these combinations involving *TP53* alterations were significantly associated with a higher mortality from MDS compared with *TP53*-involving SNVs/indels alone (Extended Data Fig. 6g–h).

An almost identical risk estimation for HM was obtained in a case-control setting including all 672 cases who developed HM (Extended Data Figs. 1a and 9a). A small number of cases in which the onset of HM was recorded due to incomplete follow-up and exclusion of MDS and myeloproliferative neoplasms from the follow-up prevented powered analyses of the effect of CH on cumulative incidence of HM, although a similar trend of the effect of CH was observed with regard to the risk of HM that was seen in the analysis using mortality as an endpoint (Extended Data Fig. 9b–f).

Effect of SNVs/indels and CNAs on cardiovascular mortality. Finally, we investigated the combined effect of SNVs/indels and CNAs on cardiovascular mortality in the cohort of 10,623 individuals, using multivariate models to take into account known risk factors other than CH: age, gender, body mass index (BMI), comorbidities (diabetes mellitus, hypertension and dyslipidemia), history of smoking/drinking alcohol and versions of SNP array. In accordance

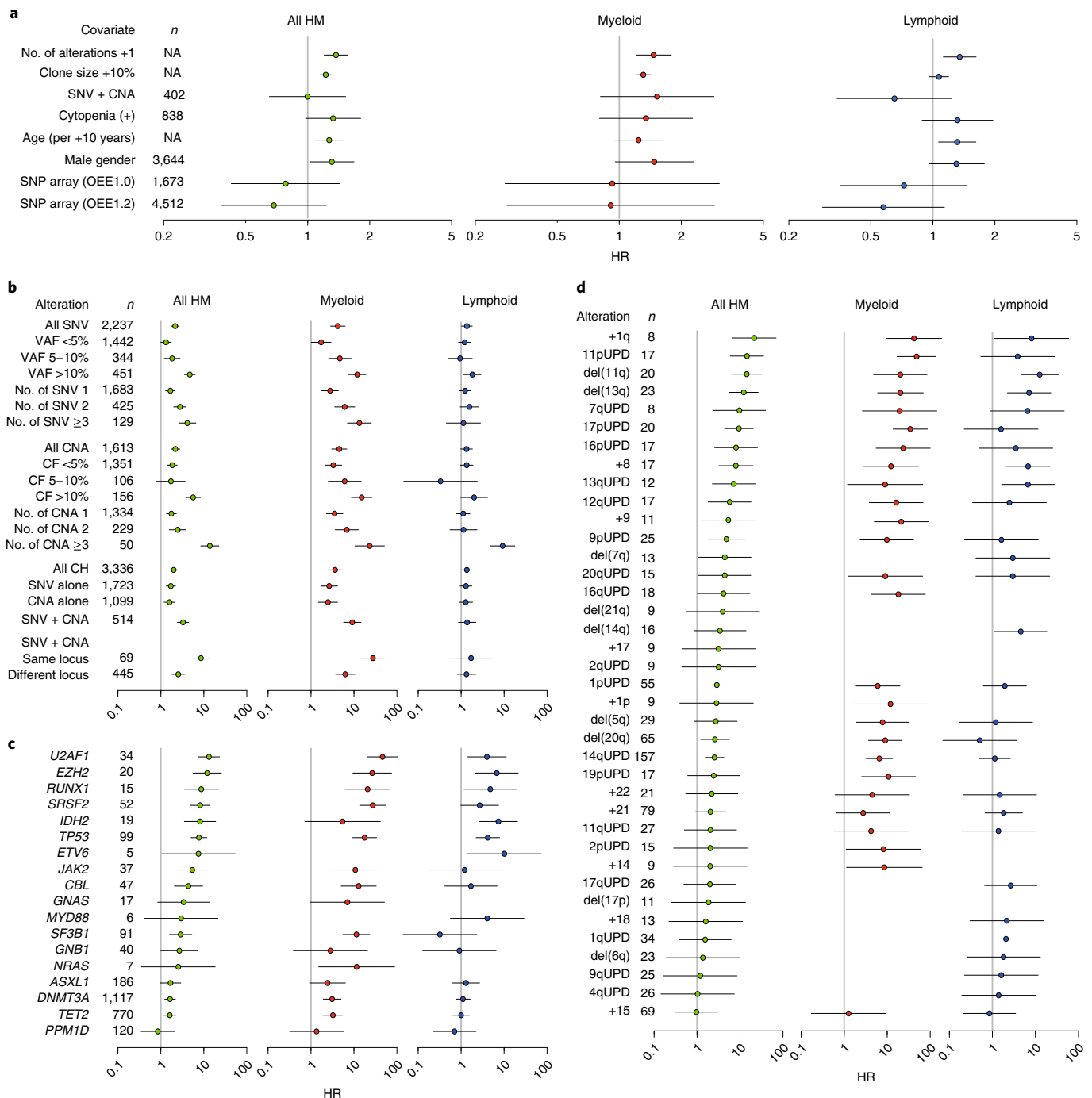


Fig. 5 | Impact of CH-related alterations on mortality from HM. a–d, HRs for mortality from all HM; myeloid neoplasms and lymphoid neoplasms are indicated by green, red and blue dots, respectively, and error bars indicate 95% CIs. **a**, HRs of the indicated covariates including the total number and clone size of CH-related alterations were calculated by multivariate Fine-Gray regression within subjects with available blood cell counts within the case-cohort design (Extended Data Fig.1b; $n=6,412$). **b–d**, HRs of the indicated alterations were calculated within the case-cohort design (Extended Data Fig.1b; $n=8,283$) in comparison with CH⁻ cases. HRs of **(b)** summarized statuses of CH-related alterations, **(c)**, individual CH-related SNVs/indels, **(d)** individual CH-related CNAs. HRs are not shown for alterations with no associated event. Cell fractions of unclassifiable CNAs are regarded as 0 **(a)** or <5% **(b)**. n , number of cases with the indicated alterations; NA, not applicable; no. of alterations +1, one additional alteration; clone size +10%, 10% increase in cell fraction; SNV + CNA, co-occurrence of both SNVs/indels and CNAs; age (per +10 years), for every additional 10 years of age; OEE1.0, OmniExpressExome v.1.0; OEE1.2, OmniExpressExome v.1.2; no. of SNV, number of SNVs/indels; CF, cell fraction of CNAs.

with a previous report¹³, the presence of CH-related SNVs/indels with large clone size ($\text{VAF} \geq 5\%$) was associated with elevated cardiovascular and all-cause mortality ($\text{HR}=1.36$, 95%CI: 1.09–1.71 for cardiovascular mortality; $\text{HR}=1.41$, 95%CI: 1.24–1.60 for

all-cause mortality) (Fig. 6a and Extended Data Fig. 10a). In support of this, we observed a significant association of SNVs/indels with hypertension (Fig. 6b), which was independent of known risk factors for hypertension including older age, higher BMI and diabetes.

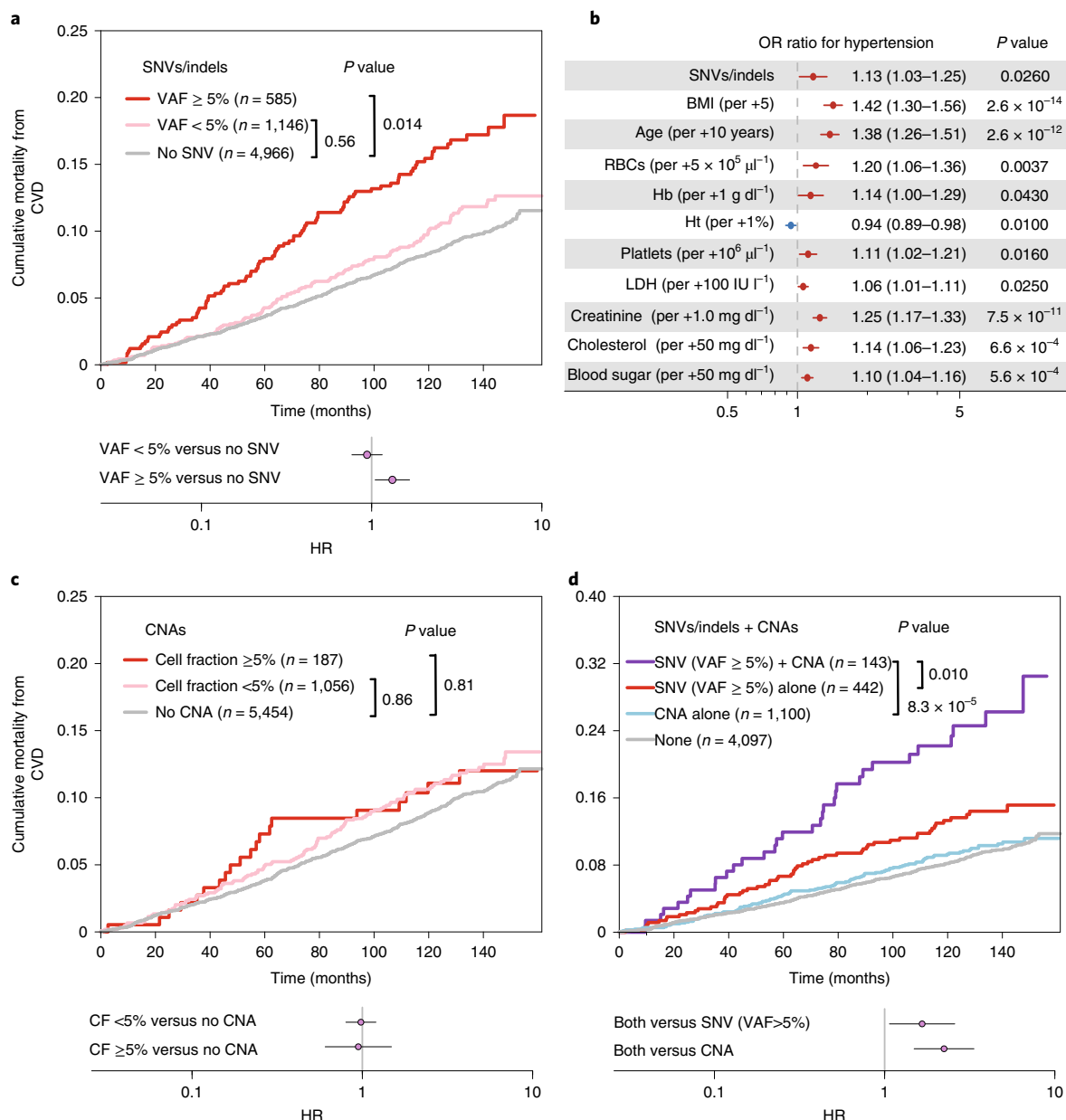


Fig. 6 | Effect of SNV/indels and CNAs on cardiovascular mortality. a, Cardiovascular mortality in subjects with SNV/indels (VAF $\geq 5\%$ or $< 5\%$), and those without SNV/indels. Hazard ratios and P values were calculated in comparison with those without SNV/indels by two-sided Wald test. **b**, Results of multivariate logistic regression for the presence of hypertension among 4,660 subjects with available information for covariates. Explanatory variables were selected by a stepwise method from the following factors: presence of SNV/indels, CNAs, age, gender, BMI, history of drinking alcohol and smoking, presence of diabetes mellitus, hyperlipidemia, hypertension and 13 blood parameter values available in $\geq 70\%$ of subjects. Only the remaining variables are shown. Ht, hematocrit; RBCs, red blood cells. **c**, Cardiovascular mortality in subjects with CNAs (cell fraction (CF) $\geq 5\%$ or $< 5\%$) and without. Hazard ratios are calculated in comparison with those without CNAs. **d**, Cardiovascular mortality in subjects with both SNV/indels (VAF $\geq 5\%$) and CNAs (purple), with SNV/indels (VAF $\geq 5\%$) alone (red), with CNAs alone (blue) and with no SNV/indels (VAF $\geq 5\%$) or CNAs (gray). Hazard ratios were calculated by comparing those with both SNV/indels (VAF $\geq 5\%$) and CNAs with those with SNV/indels (VAF $\geq 5\%$) alone or CNAs alone. **a,c,d**, All comparisons were performed with multivariate models including age, gender, BMI, comorbidities (diabetes mellitus, hypertension and dyslipidemia), history of smoking/drinking alcohol and versions of SNP array among 6,697 subjects with available clinical information. Error bars indicate 95% CIs and P values were calculated by two-sided Wald test and not adjusted for multiple comparison.

By contrast, regardless of their clone size, CNAs alone did not seem to affect cardiovascular or all-cause mortality (Fig. 6c and Extended Data Fig. 10b). However, CNAs in combination with SNVs/indels with VAF $\geq 5\%$ were significantly associated with elevated cardiovascular mortality and all-cause mortality compared with CNAs alone, SNVs/indels alone and either SNVs/indels or CNAs (Fig. 6d

and Extended Data Fig. 10c), although there was no significant difference in cardiovascular mortality or overall survival depending on whether they involved the same locus (Extended Data Fig. 6d,e). In multivariate analysis, the combined effect of both lesions was independent of both the number of co-occurring SNVs/indels (HR = 1.77, $P = 0.012$; Extended Data Fig. 10d,e) and the total number

of alterations (Extended Data Fig. f–h). Given the absence of impact of CNAs alone, the combined effect on cardiovascular and all-cause mortality does not seem to be explained by an increased total number of CH-related lesions; in fact, this number did not correlate with cardiovascular and all-cause mortality, except for a significantly higher mortality for three or more CH-related lesions (Extended Data Fig. 10i), probably involving both SNVs/indels and CNAs. Collectively, these observations suggested that the presence of both SNVs/indels and CNAs increased the cardiovascular and all-cause mortality compared with either lesion alone.

Discussion

Combining targeted deep sequencing of major CH-related genes and SNP array-based copy number analysis of blood-derived DNA from >10,000 individuals aged ≥ 60 years, we have delineated a comprehensive registry of CH in a general population of elderly individuals in terms of both SNVs/indels and CNAs. A case-cohort study design enabled accurate estimation of CH-associated cumulative HM mortality in a large general cohort of elderly individuals (>43,000), including >400 cases who developed HM and substantially saving the cost and effort of sequencing, where only ~8,300 (~18%) individuals per subcohort were fully genotyped. It should be noted that, with a much larger number of cases with HM mortality ($n=401$) compared with previous cohort studies (16 and 37 cases per cohort)^{12,13}, the estimation of HM mortality in individuals with CH-related SNVs/indels was substantially more accurate and with a much narrower CI for both myeloid and lymphoid malignancies, where mortality attributable to CH was mostly explained by myeloid malignancies regardless of the type of CH-related lesions. Estimation of OR for CH⁺ versus CH⁻ cases was even more accurate, with a total of 672 HM events in a case-control study setting.

Including both types of lesion, CH was found in as many as 40% of a general population ≥ 60 years of age, of which 11% had $\geq 10\%$ clone size. As a whole, SNVs/indels and CNAs co-occurred more frequently than expected by chance only. In particular, as repeatedly highlighted in myeloid neoplasms^{29,33,42}, SNVs/indels in *DNTM3A*, *TET2*, *JAK2* and *TP53* significantly co-occurred with LOH at each locus in CH, suggesting the role of biallelic alterations of these genes even at an early stage during leukemogenic evolution. Co-occurrence of *TET2*-related SNVs/indels and deletions involving the *TCRA* locus that are suggestive of evolution of *TET2*-mutated T-cell clones is also of interest. However, even with exclusion of subjects having these combinations affecting the same gene, SNVs/indels and CNAs significantly co-occurred ($P=0.0042$). Given that most CNAs in CH are recurrently seen in myeloid neoplasms, this suggests the presence of functional interactions between CH-related SNVs/indels and CNAs for positive selection, although we cannot exclude the possibility that CNAs may simply represent chromosomal instability induced by one or more CH-related SNVs/indels.

Compared with those having SNVs/indels or CNAs alone, CH⁺ individuals with both lesions showed a higher clone size, more abnormal blood counts and a higher mortality from HM, particularly of myeloid lineages. The combined effect of SNVs/indels and CNAs⁴⁰ is typically exemplified by biallelic alterations in *DNTM3A*, *TET2*, *JAK2* and *TP53* resulting from LOH affecting the mutated locus. However, the effect of combining SNVs/indels and CNAs is largely explained by an increased total number of CH-related lesions. Given that the size of CH clones correlated with the number of CH-related lesions, the increasing number of mutations is thought to promote expansion of clones, contributing to an earlier onset and progression of HM. This underscores the importance of measuring both lesions for accurate estimation of HM mortality, which is expected to increase the number of CH-related lesions evaluated only for SNVs/indels and CNAs alone by 0.25 and 0.36%, respectively, on average, revising 10-year expected HM mortality by 0.14 and 0.19%, respectively. The combined effect of both SNVs/

indels and CNAs was also observed for cardiovascular and all-cause mortality. Of interest, the effect was seen despite the fact that CNAs alone did not affect mortality. Because the effect of SNVs/indels on cardiovascular mortality depended on their VAF values, which increased with the presence of CNAs, the combined effect seems to be mediated in part by an increased size of clones having SNVs/indels, although CNA still remained significant after adjustment for the effects of clone size.

Potential caveats in the current study include a limited number of CH-related genes analyzed ($n=23$), compromised sensitivity in the detection of focal CNAs and a study population exclusively including individuals >60 years of age. However, these 23 genes, which are estimated to capture ~90% of CH-related SNVs/indels^{12,13}, were analyzed using deep sequencing to sensitively detect lesions in very small fractions (~1%), which would not have been possible with a more unbiased sequencing and a larger target size. In addition, CH and related HM and CVD are highly enriched in, and mostly confined, to this age group, respectively. Thus, the limited number of genes and age grouping may not necessarily be limitations, but rather contributed to the efficient and comprehensive analysis of CH-related alterations in a large number of cases initiated to investigate their effects on clinical outcomes and at acceptable cost. However, it is clear that more comprehensive studies with unbiased sequencing and improved copy number detection, including all age groups, are warranted to elucidate the full spectrum of CH-related alterations in future studies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-021-01411-9>.

Received: 19 August 2020; Accepted: 26 May 2021;

Published online: 8 July 2021

References

1. Steensma, D. P. et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
2. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
3. Busque, L. et al. Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. *Blood* **113**, 3472–3474 (2009).
4. Gale, R. E., Wheadon, H. & Linch, D. C. X-chromosome inactivation patterns using HPRT and PGK polymorphisms in haematologically normal and post-chemotherapy females. *Br. J. Haematol.* **79**, 193–197 (1991).
5. Fey, M. F. et al. Clonality and X-inactivation patterns in hematopoietic cell populations detected by the highly informative M27 beta DNA probe. *Blood* **83**, 931–938 (1994).
6. Champion, K. M., Gilbert, J. G., Asimakopoulou, F. A., Hinshelwood, S. & Green, A. R. Clonal haemopoiesis in normal elderly women: implications for the myeloproliferative disorders and myelodysplastic syndromes. *Br. J. Haematol.* **97**, 920–926 (1997).
7. Busque, L. et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).
8. Jacobs, K. B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
9. Laurie, C. C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
10. Loh, P. R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
11. Loh, P. R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
12. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
13. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
14. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).

15. Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015–1023 (2018).
16. Coombs, C. C. et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell Stem Cell* **21**, 374–382 (2017).
17. Bolton, K. L. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat. Genet.* **52**, 1219–1226 (2020).
18. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
19. Fuster, J. J. et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
20. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).
21. Terao, C. et al. Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130–135 (2020).
22. Gao, T. et al. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. *Nat. Commun.* **12**, 338 (2021).
23. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
24. Momozawa, Y. et al. Low-frequency coding variants in CETP and CFB are associated with susceptibility of exudative age-related macular degeneration in the Japanese population. *Hum. Mol. Genet.* **25**, 5027–5034 (2016).
25. Ogawa, S. Genetics of MDS. *Blood* **133**, 1049–1059 (2019).
26. Ochi, Y. et al. Combined cohesin-RUNX1 deficiency synergistically perturbs chromatin looping and causes myelodysplastic syndromes. *Cancer Discov.* **10**, 836–853 (2020).
27. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
28. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
29. Kralovics, R. et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N. Engl. J. Med.* **352**, 1779–1790 (2005).
30. Langemeijer, S. M. et al. Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat. Genet.* **41**, 838–842 (2009).
31. Jasek, M. et al. TP53 mutations in myeloid malignancies are either homozygous or hemizygous due to copy number-neutral loss of heterozygosity or deletion of 17p. *Leukemia* **24**, 216–219 (2010).
32. Thoenissen, N. H. et al. Prevalence and prognostic impact of allelic imbalances associated with leukemic transformation of Philadelphia chromosome-negative myeloproliferative neoplasms. *Blood* **115**, 2882–2890 (2010).
33. Yoshizato, T. et al. Genetic abnormalities in myelodysplasia and secondary acute myeloid leukemia: impact on outcome of stem cell transplantation. *Blood* **129**, 2347–2358 (2017).
34. Watatani, Y. et al. Molecular heterogeneity in peripheral T-cell lymphoma, not otherwise specified revealed by comprehensive genetic profiling. *Leukemia* **33**, 2867–2883 (2019).
35. Muto, H. et al. Reduced TET2 function leads to T-cell lymphoma with follicular helper T-cell-like features in mice. *Blood Cancer J.* **4**, e264 (2014).
36. Schneider, R. K. et al. Rps14 haploinsufficiency causes a block in erythroid differentiation mediated by S100A8 and S100A9. *Nat. Med.* **22**, 288–297 (2016).
37. Stoddart, A. et al. Haploinsufficiency of del(5q) genes, Egr1 and Apc, cooperate with Tp53 loss to induce acute myeloid leukemia in mice. *Blood* **123**, 1069–1078 (2014).
38. Wolkewitz, M., Palomar-Martinez, M., Olaechea-Astigarraga, P., Alvarez-Lerma, F. & Schumacher, M. A full competing risk analysis of hospital-acquired infections can easily be performed by a case-cohort approach. *J. Clin. Epidemiol.* **74**, 187–193 (2016).
39. Hirata, M. et al. Overview of BioBank Japan follow-up data in 32 diseases. *J. Epidemiol.* **27**, S22–S28 (2017).
40. Young, A. L., Tong, R. S., Birmann, B. M. & Druley, T. E. Clonal hematopoiesis and risk of acute myeloid leukemia. *Haematologica* **104**, 2410–2417 (2019).
41. Bernard, E. et al. Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes. *Nat. Med.* **26**, 1549–1556 (2020).
42. Tefferi, A., Lim, K.-H. & Levine, R. Mutation in TET2 in myeloid cancers. *N. Engl. J. Med.* **361**, 1117 (2009).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Sample ascertainment. All subjects in this study were derived from the BBJ project, a multi-hospital-based-registry²³. The BBJ project enrolled approximately 200,000 individuals with at least one of 47 target diseases between the fiscal years 2003 and 2007. From 179,417 participants in the BBJ project in which SNP array analysis of peripheral blood-derived DNA had been performed, we enrolled a total of 11,234 subjects. Among these, 10,623 were randomly selected from 60,787 cases aged ≥ 60 years at the time of sample collection and confirmed not to have solid cancers as of March 2013. Out of the randomly selected 10,623 cases, 61 were recorded as having either developed or died from HM. The remaining 611 subjects, all of whom were recorded as having HM events, were additionally enrolled to maximize the statistical power in survival analysis. In total, we enrolled 672 subjects with any HM events, 138 and 589 of whom, respectively, were recorded as having developed or died from HM. A demographic summary of these subjects is presented in Supplementary Table 1, and the numbers with individual targeted diseases are listed in Supplementary Table 2. Written informed consent was obtained from all participants. The protocol of this study was approved by the following ethics committees:

Kyoto University Graduate School and Faculty of Medicine, Ethics Committee, RIKEN Yokohama Branch Research Ethics Committee and Ethical Review Board of the Institute of Medical Science, The University of Tokyo.

Multiplex PCR-based targeted sequencing. For detection of CH-associated driver SNVs/indels, we performed multiplex PCR-based targeted sequencing as previously described²⁴. Primers were designed to cover coding regions of 23 driver genes commonly mutated in clonal hematopoiesis or myeloid neoplasms: *ASXL1*, *CBL*, *CEBPA*, *DDX41*, *DNMT3A*, *ETV6*, *EZH2*, *GATA2*, *GNAS*, *GNB1*, *IDH1*, *IDH2*, *JAK2*, *KRAS*, *MYD88*, *NRAS*, *PPM1D*, *RUNX1*, *SF3B1*, *SRSF2*, *TET2*, *TP53* and *U2AF1*. PCR product sizes were designed to be 180–300 base pairs (bp) to cover the amplicon by the sequencing reads. We added CGCTCTTCCGATCTCTG to the 5' end of the forward primers and CGCTCTTCCGATCTGAC to the 5' end of the reverse primers to perform a second PCR^{43,44}. We performed multiplex PCR using different primer pools to cover all coding regions of the 23 genes. Next we carried out a second PCR with primer sequences 5'-AATGATACGGCGACACCGAGATCTACACxxxxxxx-A-CACTCTTCCCTACACGACGCTCTTCCGATCTCTG-3' and 5'-CAAGCAG AAGACGGCAGATGAGTxxxxxxxGTGACTGGAGTTTCAGACGTGTGCTCT TCCGATCTGAC-3', where xxxxxxxx represents 8-bp barcodes. All second PCR products were pooled for one sequencing run. After purification of each library using Agencourt AMPure XP (Beckman Coulter), we obtained 2×150-bp paired-end reads with dual 8-bp barcode sequences on a HiSeq2500 instrument.

Calling CH-related SNVs/indels. Sequencing reads were aligned to the human genome reference (hg19) using Burrows–Wheeler Aligner v.0.7.8 (<https://sourceforge.net/projects/bio-bwa/>), with default parameter settings. Mutation calling was performed with Genomon2 pipeline v.2.6.2 (<https://genomon.readthedocs.io/ja/latest>), picard-tools v.1.39 (<http://picard.sourceforge.net/>) and GenomonMutationFilter v.0.2.1 (<https://github.com/Genomon-Project/GenomonMutationFilter>), as previously reported^{33,45–47}. Extracted SNVs/indels were annotated with ANNOVAR (<https://annovar.openbioinformatics.org/en/latest/>). Next, we adopted variants fulfilling the following criteria:

1. Number of variant reads ≥ 10 (≥ 5 for TCGA dataset)
2. VAF $\geq 0.5\%$
3. Nonsynonymous variants within coding sequence or splice-site variants

For calculation of read counts and VAF, we counted only base calls with mapping quality score ≥ 40 and base quality score ≥ 20 .

To further exclude false-positive calls due to sequencing artifacts, we modeled site-specific error rates as beta-binomial distribution using the R package VGAM 1.1.3 (<https://cran.r-project.org/web/packages/VGAM/index.html>). Parameters for beta-binomial distribution were determined by the maximum likelihood method⁴⁸ based on the read counts in all samples. Mutation calls whose VAFs were significantly deviated from background-error distribution ($P_{\text{beta-binomial}} \leq 10^{-6}$) were regarded as true SNVs/indels.

Additionally, those variants that always appeared within similar ranges of VAF (especially <1 or $>40\%$) were likely to be sequencing artifacts or germline polymorphisms rather than true somatic mutations. Based on this assumption, we excluded candidates fulfilling both of the following criteria from the remaining candidates:

1. Candidates observed in at least five samples
2. Mean VAF <1 or $>40\%$, or coefficient of variation in VAF <0.5

Candidates meeting the quality filter noted above were included in the subsequent analyses if they fulfilled one of the following criteria for driver SNVs/indels³⁵:

1. Candidates resulting in amino acid substitutions registered in the Catalog of Somatic Mutations in Cancer v.91 database (<https://cancer.sanger.ac.uk/cosmic>) for at least five counts

2. Candidates fulfilling criterion 1 and at least one of the items in criterion 2 (below).

Criterion 1

Candidates not registered in public databases, including dbSNP138 (<https://www.ncbi.nlm.nih.gov/snp/>), the 1000 Genomes Project as of October 2014 (<https://www.internationalgenome.org/>), the Human Genome Variation Database (<https://www.hgvd.genome.med.kyoto-u.ac.jp/>) and The Exome Aggregation Consortium (<https://gnomad.broadinstitute.org/>).

Criterion 2

- (1) Candidates located on the nonrepeat region with VAF ≥ 4 , <40 , ≥ 60 or $<96\%$
- (2) Nonsense, frameshift or splice-site candidates
- (3) Candidates computationally predicted to have negative consequences: SIFT score <0.05 (<https://sift.bii.a-star.edu.sg/>), damaging by PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>) and high or medium by MutationAssessor (<http://mutationassessor.org/>)

Finally, the resulting set of driver mutations were manually reviewed in Integrated Genome Viewer 2.4.6 (<http://software.broadinstitute.org/software/igv/>).

In silico simulation of SNVs/indels calling. To benchmark the performance in detection of low-VAF SNVs/indels, we performed in silico simulation. Mixing two bam files of variable proportions, we diluted 750 heterozygous SNPs and artificially created low-VAF SNVs/indels (ranging 0.5–5%). Each diluted SNP was classified into six bins according to sequencing depth (100–300×, 300–500×, 500–750×, 750–1,000×, 1,000–1,500× and $>1,500\times$), and sensitivities were calculated separately for the six bins. We calculated sensitivity as a fraction of detected variants within all simulated variants:

$$SN_{VAF=x\%} = TP_{VAF=x\%} / (TP_{VAF=x\%} + FN_{VAF=x\%})$$

where $SN_{VAF=x\%}$ denotes sensitivity for variants with $x\%$ VAFs; $TP_{VAF=x\%}$ denotes number of detected SNPs whose VAFs were diluted to $x\%$; and $FN_{VAF=x\%}$ denotes number of missed SNPs whose VAFs were diluted to $x\%$.

Together with sensitivity, we calculated specificity by sampling genomic positions without known SNPs ($n=5,000$ per simulation). We counted SNVs/indels calls on these positions as false positives, and calculated the specificity as follows:

$$SP = 1 - FP/N$$

where SP is specificity, FP is number of false-positive calls and N is number of sampled genomic positions.

To draw receiver operator characteristic curves, we calculated sensitivities and specificities for nine different cutoffs on beta-binomial P -values (10^{-2} , 10^{-3} , 10^{-4} , 10^{-5} , 10^{-6} , 10^{-7} , 10^{-8} , 10^{-9} and 10^{-10}).

Copy number analysis. Our analyses of CNAs are based on results from a previous publication²¹ in which blood-derived DNA samples from 11,234 subjects were examined by one of three different versions of microarray: Illumina Infinium OmniExpress ($n=708$), Infinium OmniExpressExome v.1.0 ($n=3,152$) or v.1.2 ($n=7,374$). For detection of CNAs, we analyzed allele-specific hybridization intensities for the polymorphisms examined by all versions of arrays ($n=515,355$). Haplotype phasing was performed by Eagle2 (<https://alkesgroup.broadinstitute.org/Eagle/>), and log R ratio (LRR) and B-allele frequency (BAF) were calculated as previously described²¹. Based on long-range haplotype information and LRR/BAF values, we detected allelic imbalances and classified these into duplications, deletions and UPDs, with FDR around 5% (refs. ^{10,21}). Meanwhile, copy number analysis of 8,023 samples from the TCGA cohort was performed with a standalone software, MoChA (<https://github.com/freeseek/mocha/>). Because the power required to detect allelic imbalances exceeded that to distinguish UPD from copy number gain or loss, CNAs were designated as 'unclassifiable' when we could not assign them into specific types of CNA. In the analyses where the exact discrimination between UPD, duplication or deletion (for example, lesion-specific analysis in Figs. 2a and 3) was relevant, we excluded unclassifiable CNAs from the analysis. Although we cannot calculate precise cell fractions for unclassifiable CNAs, their cell fractions are basically expected to be quite small. Therefore, when we classified CNAs by their cell fractions (for example, Figs. 4d, 5b and 6c), unclassifiable CNAs were regarded as below threshold. When we analyzed CNAs in terms of their cell fractions (for example, Figs. 1d,e and 5a), unclassifiable CNAs were excluded. Otherwise, we did include those unclassifiable CNAs in the analysis (for example, Figs. 1a–c, 2b–e, 3c,d, 4 and 6b,d). Based on CNAs detected, we determined chromosomal regions significantly affected with CNAs by parametric aberration recurrence test⁴⁹.

Definition of abnormalities in blood counts. Subjects fulfilling at least one of the following criteria were considered to have abnormalities in blood counts.

1. White blood cells (μl^{-1}) $\geq 10,000$ or $<3,000$
2. Hemoglobin (g dl^{-1}) ≥ 16.5 (male), ≥ 16 (female) or <10
3. Hematocrit (%) ≥ 50
4. Platelets ($10,000 \mu\text{l}^{-1}$) ≥ 50 or <10

These cutoffs on blood counts were adopted from diagnostic criteria for myelodysplastic syndromes or myeloproliferative neoplasms⁵⁰. Among subjects with available counts for WBC, hemoglobin, hematocrit and platelets ($n = 8,345$), 7,031 (84.3%) had normal levels.

Analysis of lineage-sorted samples. Frozen bone marrow was thawed in DMEM (Sigma-Aldrich) containing 10% fetal bovine serum (FBS, Biosera) and 1% penicillin/streptomycin solution (ThermoFisher). After washing cell pellets with PBS containing 2% FBS, cells were stained with antibody mix for 20 min followed by washing with PBS containing 2% FBS and filtered with a 5-ml round-bottom polystyrene test tube with cell strainer snap-cap (ThermoFisher). We then mixed 500 μ l of the filtered cell suspension in PBS containing 2% FBS with 5 μ l of propidium iodide staining solution (BD Bioscience), which was then sorted with a FACSAria III cell sorter (BD Bioscience). The antibodies used in flow cytometry are listed in Supplementary Table 4. For ddPCR and amplicon sequencing we sorted myeloid, erythroid, T-cell and B-cell fractions, and genomic DNA was extracted from sorted cells. To detect allelic imbalances in the region of del(13q), amplicon sequencing was performed with custom primers targeting heterozygous SNPs within the deleted region (ThermoFisher; Supplementary Table 5). To detect the A1153V substitution in *TET2*, ddPCR was performed as described below. For single-cell analysis, CD34⁺ cells were sorted. Cells were resuspended in StemSpan serum-free expansion medium (STEMCELL Technologies) at 400–1,600 cells μ l⁻¹, which was then applied to a Fluidigm C1 platform for combined single-cell gene expression analysis and SNV detection. Detailed methods for single-cell analysis are in preparation for publication (shared upon request, M. M. Nakagawa, R. Inagaki et al.).

ddPCR. For ddPCR, predesigned probes were purchased from Bio-Rad. We mixed 50 ng of gDNA with enzymes (ddPCR Supermix for Probes (no dUTP), Bio-Rad) and the probe mix, followed by droplet generation and PCR amplification according to the manufacturer's protocol. Annealing temperatures was set at 55 °C. We measured amplified droplets using the QX200 system and QuantaSoft 1.7 (Bio-Rad, <https://www.bio-rad.com/webroot/web/pdf/lsr/literature/QuantaSoft-Analysis-Pro-v1.0-Manual.pdf>). Catalog numbers for the probe mix are shown in Supplementary Table 6.

Statistical analysis. All statistical analyses were performed using the R statistical platform v.3.6.1 (<https://www.r-project.org/>). All statistical tests were two-sided. Benjamini–Hochberg multiple testing correction was applied where appropriate.

Age-stratified permutation test for co-occurrences of CH-related alterations. We tested the significance of co-occurrences between SNVs/indels and CNAs under the stratification by subjects' age, because age-dependent frequencies of both CH-related alterations can confound their co-occurrences. First, we stratified subjects into 41 bins according to their age (60, 61, ..., 100 years) and calculated frequencies of SNVs/indels, CNAs and their co-occurrences within each bin. In a single iteration of permutation, we randomized the status of SNVs/indels and CNAs in all subjects while retaining their frequencies in each age bin. Then, the numbers of co-occurrences were summed across all age bins. By repeating this process, we obtained null random distribution of the number of subjects with co-occurring SNVs/indels and CNAs. Comparing the null distribution and actual number of co-occurrences, we obtained *P* values for significance of co-occurrences between SNVs/indels and CNAs. Significant co-occurrences of multiple CH-related alterations were also demonstrated in a similar way, where we counted the total number of CH-related alterations within each age bin. In a single iteration, these alterations were randomly reassigned to the subjects retaining the total number of alterations in each bin. Then, the number of subjects to whom multiple alterations was assigned was counted across all bins. *P* values were calculated by comparing the actual number of cases with at least two alterations and null distribution generated by repeating the process above.

Simulation test for cell-level coexistence of SNVs/indels and CNAs involving the same genes. Regarding the combinations of SNVs/indels and UPDs involving the same genes (*DNMT3A*, *TET2*, *TP53* and *JAK2*), we observed higher VAF values of SNVs/indels than cell fractions of CNAs in 51 of the 55 cases, which suggested they were likely to have been acquired in the same cells and resulted in biallelic alterations (Supplementary Fig. 8a,b). To examine how many of the 51 cases could be explained by the cell-level coexistence of SNVs/indels and UPDs, we performed random simulation on their clone size using a null hypothesis, $H_0(x)$: SNVs/indels and UPDs were independently acquired in at least x cases ($x = 5, 4, \dots, 55$). The *P* value for $H_0(x)$ was calculated assuming that VAF of SNVs/indels and cell fractions of UPDs follow independent distributions (Supplementary Fig. 8c–e). We searched for the maximum x for which the *P* value for $H_0(x)$ was < 0.05 , to obtain the minimum estimate of the number of cases in which the cell-level coexistence of SNVs/indels and UPDs was expected (Supplementary Fig. 8f).

Calculation of adjusted VAF. We observed significant co-occurrences of *DNMT3A*/2pLOH, *TET2*/4qLOH, *JAK2*/9pUPD and *TP53*/17pLOH, which suggested that biallelic alterations of these genes were positively selected in

CH. However, the frequencies of these co-occurrences may be overrepresented because underlying LOH inflated VAF values of SNVs/indels and resulted in higher sensitivity in the detection of SNVs/indels with the coexistence of LOH. To exclude the effect of VAF inflation, we tested the significant co-occurrence of these combinations of SNVs/indels and LOH, focusing on those SNVs/indels having $\geq 5\%$ VAF, for which almost 100% detection sensitivity would be expected (Supplementary Fig. 1a), where inflated VAF values due to LOH were adjusted according to the following formula, by calculating the cell fraction with LOH (Supplementary Fig. 6a):

$$\text{VAF}_{\text{adjusted}} = \text{VAF}_{\text{observed}} - \text{CF}_{\text{LOH}}/2 : \text{ with UPD}$$

$$\text{VAF}_{\text{adjusted}} = \text{VAF}_{\text{observed}} (1 - \text{CF}_{\text{LOH}}/2) : \text{ with deletion}$$

$$\text{VAF}_{\text{adjusted}} = \text{VAF}_{\text{observed}} : \text{ without LOH}$$

where CF_{LOH} denotes the cell fractions of LOH co-occurring in the same genes.

Focusing on SNVs/indels with large adjusted VAF ($> 5\%$), we examined the significance of co-occurrences of *DNMT3A*/2pLOH, *TET2*/4qLOH, *JAK2*/9pUPD and *TP53*/17pLOH (Supplementary Fig. 6b).

Risk factors for CH. To extract risk factors for CH, we examined correlations between genetic alterations in CH and subjects' baseline characteristics (age, sex, history of smoking and drinking alcohol). Information regarding the history of smoking and drinking alcohol was based on self-report questionnaires at the time of DNA sampling. First, we performed univariate logistic regressions for the presence of genetic alterations. Based on factors significantly correlated with genetic alterations ($q < 0.1$), we then performed multivariate logistic regressions to extract independent risk factors ($P < 0.05$).

Effect of CH on blood cell counts. To elucidate the effects of genetic alterations on blood cell counts, we examined correlations between genetic alterations and blood cell counts. After Cox–Box transformation of blood counts with the R package 'car' v.3.0.8 (<https://cran.r-project.org/web/packages/car/index.html>), linear regressions were performed. To correct for confounding effects, all regressions were performed with multivariate models including age, gender and versions of SNP array as covariates, in comparison with subjects without detectable CH.

Prediction models for hypertension. To elucidate the relationships between CH and hypertension, we performed multivariate logistic regression. Optimal sets of variables were selected by a stepwise method from known risk factors and blood test values available for $\geq 70\%$ of the subjects: presence of SNVs/indels and CNAs, age, gender, BMI, history of smoking and drinking alcohol (based on self-report questionnaires), white blood cells, red blood cells, hemoglobin, hematocrit, MCHC, platelets, aspartate aminotransferase, alanine aminotransferase, lactate dehydrogenase (LDH), creatinine, blood urea nitrogen, total cholesterol and glucose.

Survival analysis. We evaluated the effects of CH-related SNVs/indels, CNAs and their combinations on mortality from HM, all-cause mortality and cardiovascular mortality. To define mortality from hematologic malignancies, we included diagnoses within the International Classification of Diseases, Tenth Revision code groups C81–C96 (malignant neoplasms of lymphoid, hematopoietic and related tissue), D45 (polycythemia vera), D46 (myelodysplastic syndromes), D47 (other neoplasms of uncertain behavior of lymphoid, hematopoietic and related tissue) and D7581 (myelofibrosis). For CVD, we included I20–I25 (ischemic heart diseases), I48–I49 (arrhythmia), I50 (heart failure), I60–I67, I69 (cerebrovascular diseases), I70–I72 (aortic atherosclerosis, aortic aneurysm and aortic dissection) and I74 (peripheral artery diseases). In the analysis of all-cause mortality, we performed Cox proportional hazards regression using the R package 'survival' v.3.1.12 (<http://cran.r-project.org/web/packages/survival/index.html>). In the analysis of HM events (mortality or development) or mortality from CVD, we performed competing risk regression based on the Fine–Gray model. In the analysis of events of HM (mortality and development), we applied a case-cohort design to maximize the statistical power as previously described³⁸ (Extended Data Figs. 1b and 9b), including all subjects with HM events within the target cohort. The contribution of each factor to increase in HM mortality was estimated by calculating the log (hazard) of the corresponding factor and averaged for all patients, as previously described³³. Concretely, the contribution of factor i for patient j , designated as C_{ij} , is estimated

$$\beta_i \cdot (x_{ij} - x_{i,\text{Median}})$$

where β_i is the coefficient for factor i , x_{ij} is the covariate of patient j for factor i and $x_{i,\text{Median}}$ is the median of x_{ij} across different subjects. The contribution of factor i to the increase in HM mortality associated with CH is given as

$$\sum_{j \in \text{Subjects with CH}} C_{ij}/N_{\text{CH}} - \sum_{j \in \text{Subjects without CH}} C_{ij}/N_{\text{non-CH}}$$

where N_{CH} and N_{non-CH} denote the number of subjects with and without CH, respectively. The relative contribution of each prognostic factor (CH, age and gender) is shown in Extended Data Fig. 8a. Meanwhile, cardiovascular mortality and overall survival were analyzed in a cohort of 10,623 subjects randomly selected. To correct for confounding effects, we included subjects' age, gender and version of SNP array in the multivariate models for events of HM, while age, gender, BMI, presence of diabetes mellitus, hyperlipidemia, hypertension, history of tobacco smoking and alcohol drinking and version of SNP array were included in the models for all-cause and cardiovascular mortalities.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Tables of somatic SNVs/indels and CNAs detected in this study have been deposited in the Japanese Genome-phenotype Archive under accession code JGAS000293 (<https://humandbs.biosciencedbc.jp/en/hum0014-v22>). Clinical data used in this study can be provided by the BBJ project upon request (<https://biobankjp.org/english/index.html>). Source data are provided with this paper.

Code availability

Custom computational codes used to reproduce figures from the paper are available at https://github.com/RSaikiRSaiki/CH_2021.

References

43. Harismendy, O. et al. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol.* **12**, R124 (2011).
44. Forshew, T. et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra168 (2012).
45. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
46. Haferlach, T. et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241–247 (2014).
47. Suzuki, H. et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).
48. Shiraishi, Y. et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
49. Niida, A., Imoto, S., Shimamura, T. & Miyano, S. Statistical model-based testing to evaluate the recurrence of genomic aberrations. *Bioinformatics* **28**, i115–i120 (2012).
50. 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).

Acknowledgements

This work was supported by the Japan Agency for Medical Research and Development (nos. JP15cm0106056h0005, JP19cm0106501h0004, JP16ck0106073h0003 and JP19ck0106250h0003 to S.O.; nos. JP17km0405110h0005 and JP19ck0106470h0001 to H.M.; and no. JP19ck0106353h0003 to Y.N.); the Core Research for Evolutional Science and Technology (no. JP19gm1110011 to S.O.); the Ministry of Education, Culture, Sports, Science and Technology of Japan; the High Performance Computing Infrastructure System Research Project (nos. hp160219, hp170227, hp180198 and hp190158 to S.O. and S. Miyano) (this research used computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research project); the Japan Society for the Promotion of Science; Scientific Research on Innovative Areas (nos. JP15H05909 to S.O. and S. Miyano and JP15H05912 to S. Miyano) and KAKENHI (nos. JP26221308 and JP19H05656 to S.O., JP16H05338 and JP19H01053 to H.M. and JP15H05707 to S. Miyano); and the Takeda Science Foundation (to S.O., H.M. and T.Y.). S.O. is a recipient of the JSPS Core-to-Core Program A: Advanced Research Networks. DNA samples and subjects' clinical data were provided by BBJ, the Institute of Medical Science, the University of Tokyo. The supercomputing resource was provided by the Human Genome Center, the Institute of Medical Science, the University of Tokyo. We thank K. Matsuo at the Aichi Cancer Center Research Institute (Nagoya, Japan), who suggested the design of the case-cohort study for estimation of cumulative mortality from, and incidence of, hematological malignancies. We thank the TCGA Consortium and all its members for making publicly available their invaluable data.

Author contributions

R.S., H.M. and S.O. designed the study. K.M., Y. Kamatani, T.M. and Y. Murakami provided DNA samples and clinical data. Y. Kuroda and S. Matsuda provided bone marrow samples. C.T. and Y. Kamatani performed copy number analysis. Y. Momozawa and M.K. performed sequencing. M.M.N. performed cell sorting and single-cell analysis. R.S., M.M.N., Y.O., T.Y., Y.S., K.C., H.T., A.N., S.I. and S. Miyano performed bioinformatics analysis. R.S., Y.N., M.M.N., Y.O., T.Y., H.M. and S.O. prepared the manuscript. All authors participated in discussions and interpretation of the data and results.

Competing interests

The authors declare no competing interests.

Additional information

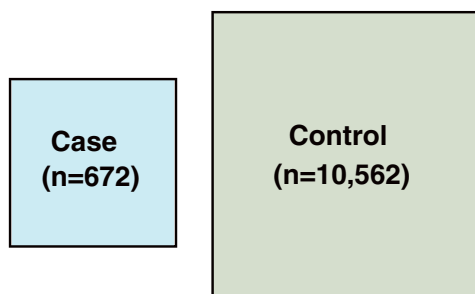
Extended data is available for this paper at <https://doi.org/10.1038/s41591-021-01411-9>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-021-01411-9>.

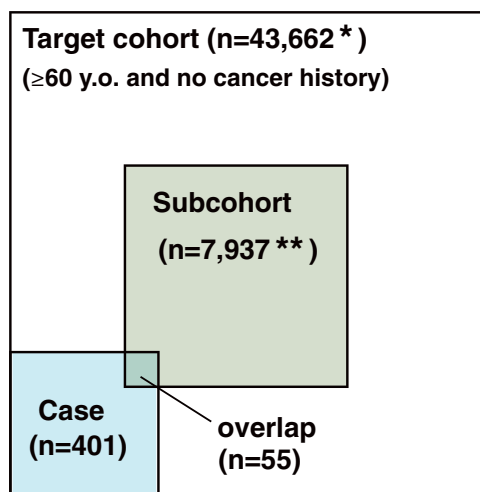
Correspondence and requests for materials should be addressed to S.O.

Peer review information *Nature Medicine* thanks Daniel Link, Duane Hassane, Todd Druley and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

a**Case-control study for all HM**

	CH(+)				CH(-)	Total
	SNV alone	CNA alone	Both	All CH(+)		
Case	154	115	107	376	296	672
Myeloid	53	41	66	160	55	215
AML	32	12	19	63	27	90
MDS	16	25	34	75	25	100
MPN	1	1	2	4	1	5
CML	1	1	5	7	2	9
Others	3	2	6	11	0	11
Lymphoid	90	69	32	191	229	420
B-NHL	61	44	18	123	143	266
T-NHL	4	7	4	15	17	32
CLL	3	2	2	7	0	7
ALL	4	3	0	7	12	19
MM/PCT	17	12	7	36	53	89
Others	1	1	1	3	4	7
Lineage Unknown	11	5	9	25	12	37
Control	2,177	1,399	633	4,209	6,353	10,562
Total	2,331	1,514	740	4,585	6,649	11,234

b**Case-cohort study for HM death**

★ Among 60,787 cases aged ≥60 years and confirmed not to have solid cancers as of March 2013, 43,662 had the follow up data for survival.

★★ Among 10,623 cases randomly selected from the 60,787 cases, 7,937 had the follow up data for survival.

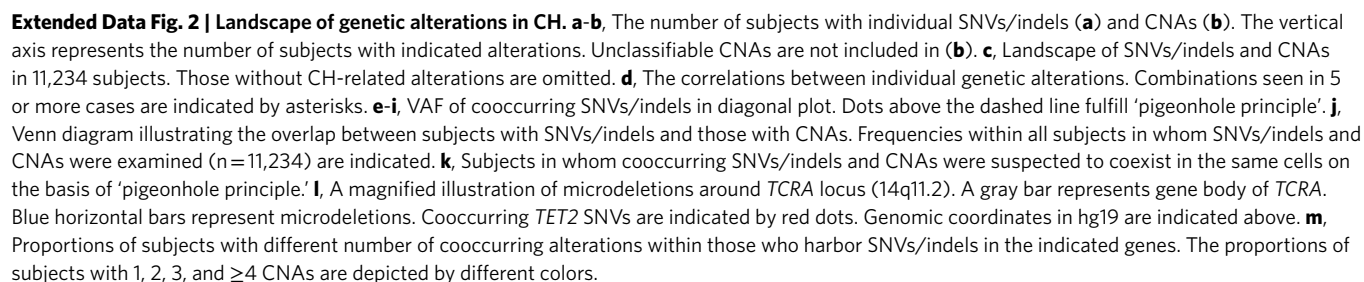
Subcohort

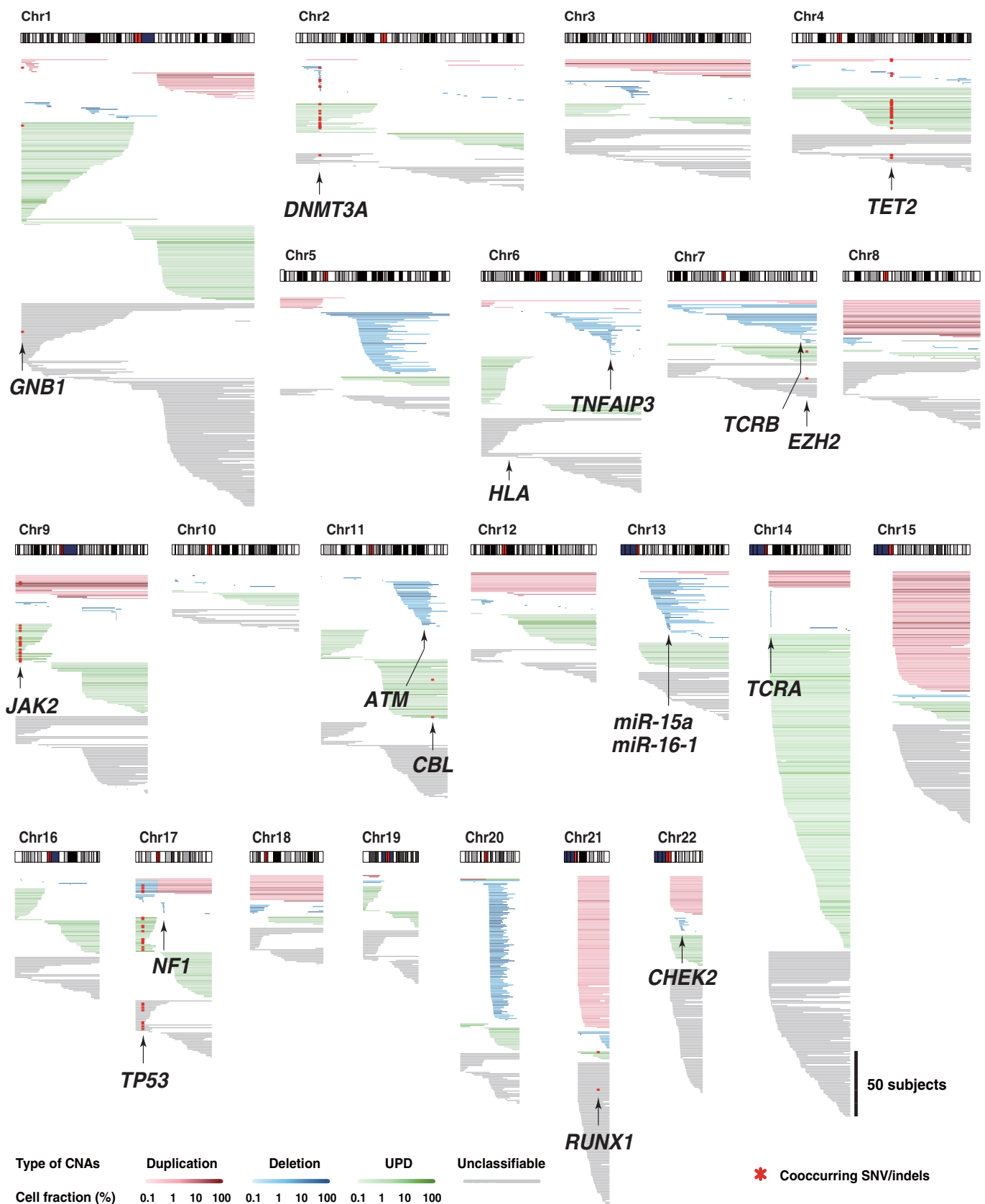
	CH(+)				CH(-)	Total
	SNV alone	CNA alone	Both	All CH(+)		
Hematological malignancy (+)	14	11	7	32	23	55
Myeloid	5	5	4	14	5	19
AML	1	0	1	2	3	5
MDS	3	4	3	10	1	11
MPN	0	0	0	0	0	0
CML	1	0	0	1	1	2
Others	0	1	0	1	0	1
Lymphoid	8	6	3	17	18	35
B-NHL	6	4	3	13	14	27
T-NHL	1	0	0	1	3	4
CLL	0	0	0	0	0	0
ALL	0	0	0	0	0	0
MM/PCT	1	2	0	3	1	4
Others	0	0	0	0	0	0
Lineage Unknown	1	0	0	1	0	1
Hematological malignancy (-)	1,614	1,036	447	3,097	4,785	7,882
Total	1,628	1,047	454	3,129	4,808	7,937

Case (Death from HM)

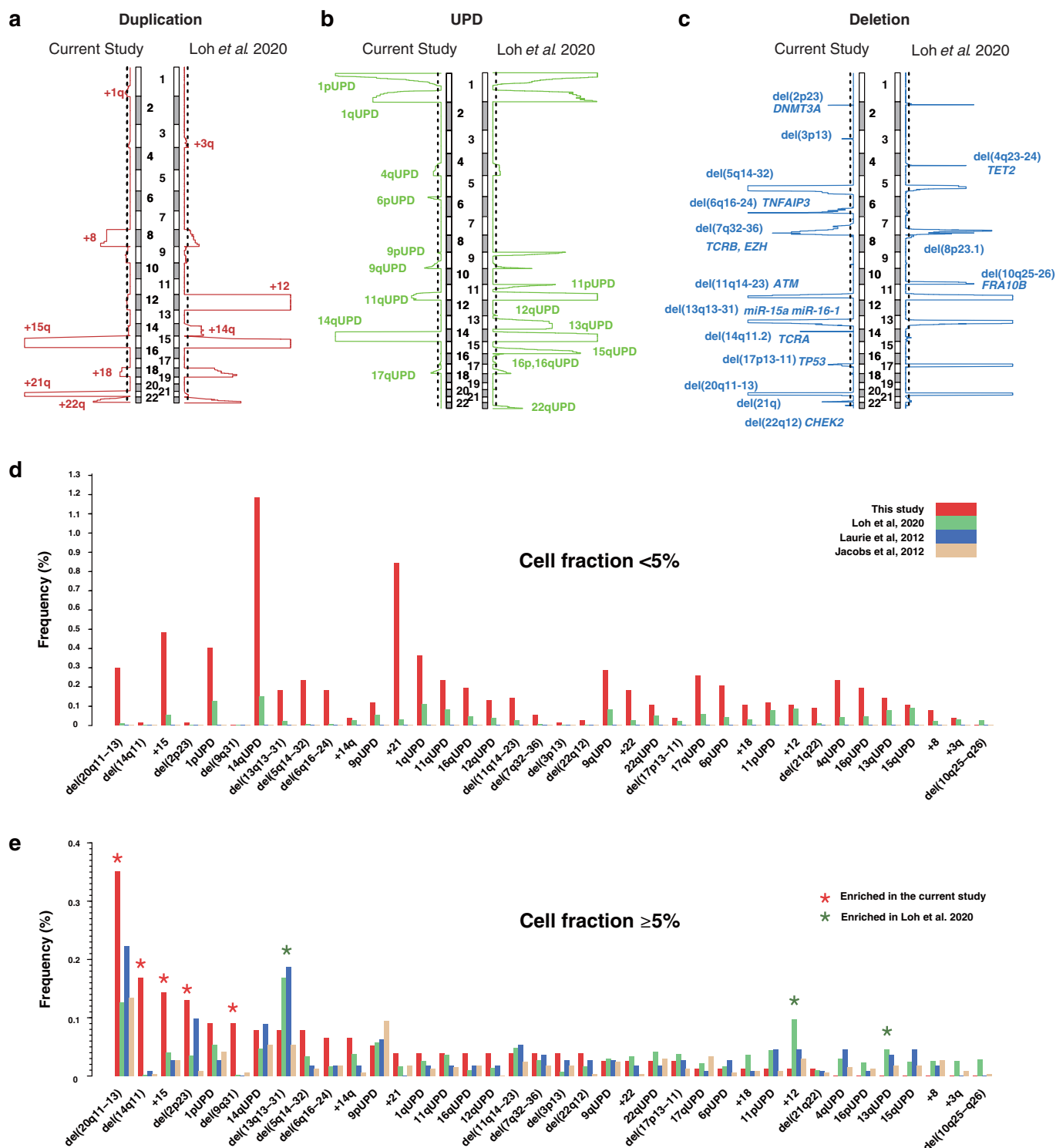
	CH(+)				CH(-)	Total
	SNV alone	CNA alone	Both	All CH(+)		
Hematological malignancy (+)	109	63	67	239	162	401
Myeloid	41	24	42	107	39	146
AML	24	8	8	40	20	60
MDS	14	13	23	50	17	67
MPN	0	1	2	3	0	3
CML	1	1	5	7	2	9
Others	2	1	4	7	0	7
Lymphoid	62	38	22	122	122	244
B-NHL	38	25	11	74	74	148
T-NHL	3	3	3	9	12	21
CLL	3	1	2	6	0	6
ALL	3	1	0	4	5	9
MM/PCT	13	8	4	25	28	53
Others	2	0	2	4	3	7
Lineage Unknown	6	1	3	10	1	11
Hematological malignancy (-)	0	0	0	0	0	0
Total	109	63	67	239	162	401

Extended Data Fig. 1 | Design of case-control and case-cohort study. **a**, Design of case-control study (Left). Diagnosis of hematological malignancies (HM) in subjects with or without CH enrolled in the case-control study (Right). **b**, Design of case-cohort study for death from HM (Left). Diagnosis of HM in subjects with or without CH enrolled in the case-cohort study (Right). AML, acute myeloid leukemia; MDS, myelodysplastic syndromes; MPN, myeloproliferative neoplasms; CML, chronic myeloid leukemia; B-NHL, B-cell non-Hodgkin lymphoma; T-NHL, T-cell non-Hodgkin lymphoma; CLL, chronic lymphoid leukemia; ALL, acute lymphoblastic leukemia; MM, multiple myeloma; PCT, plasma cell tumor.

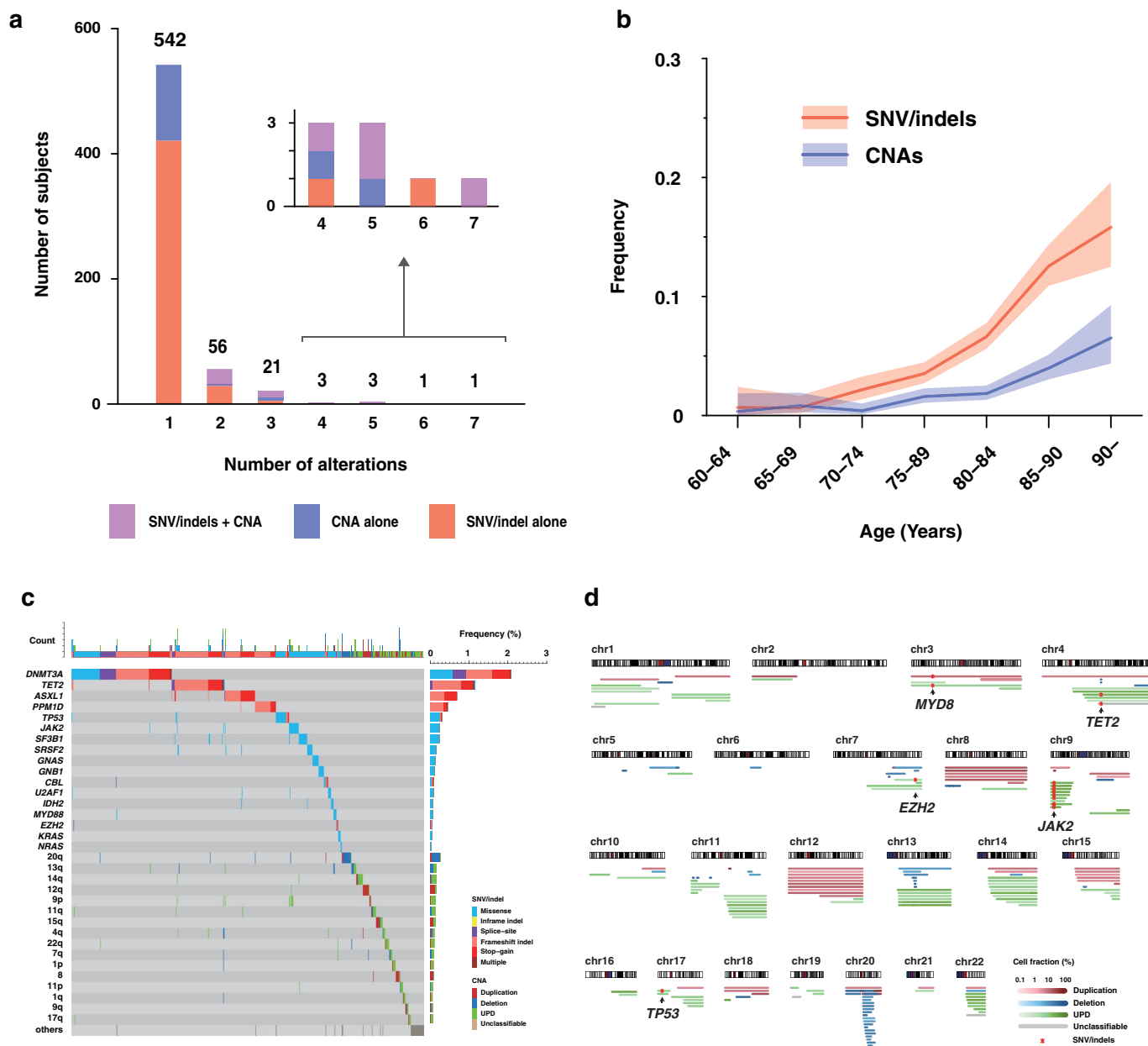




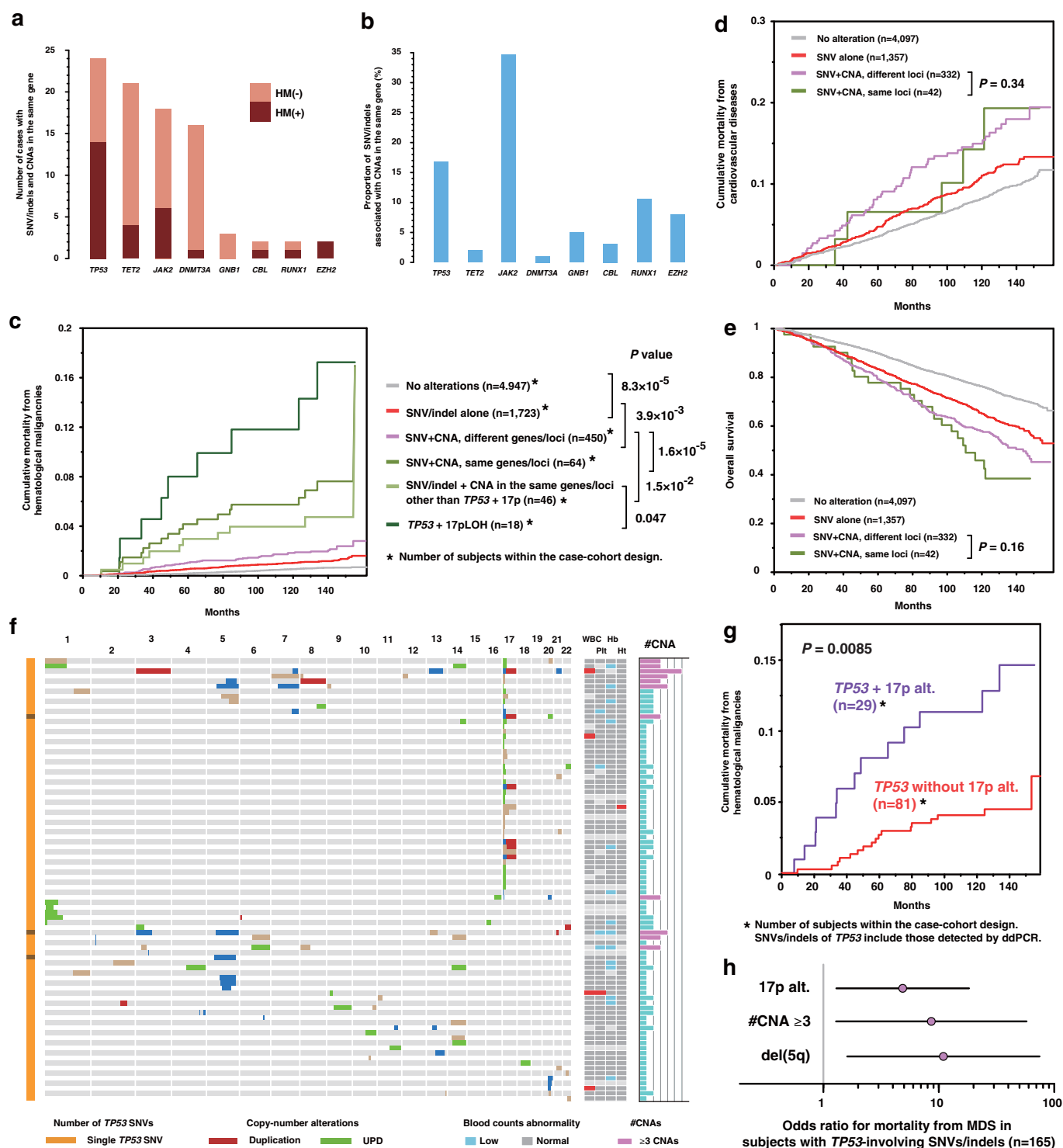
Extended Data Fig. 3 | Distribution of CNAs in all chromosomes. Distributions of CNAs on all chromosomes are illustrated. Loci of known driver genes are indicated by arrows. Each horizontal bar represents one CNA. Cooccurring SNVs/indels are indicated by red dots. Types of CNAs are depicted by different colors as indicated in the annotations.



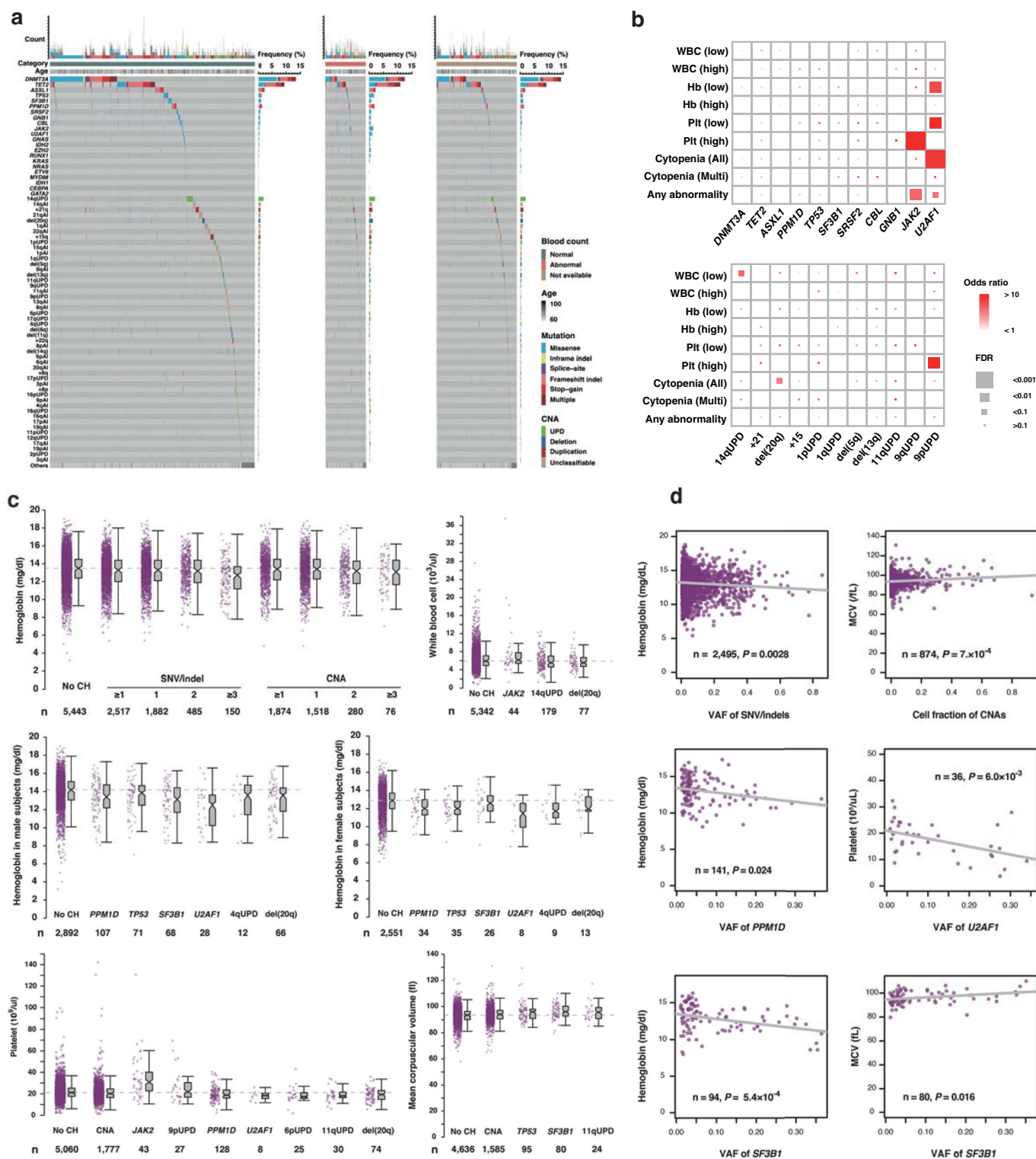
Extended Data Fig. 4 | Chromosomal regions significantly affected by CNAs. **a-c**, Chromosomal regions significantly affected by duplications (**a**), UPDs (**b**), and deletions (**c**) in a Japanese cohort (current study) and in a British cohort¹¹. Statistical significance for recurrence of CNAs were evaluated by PART⁴⁹. Dashed lines indicate thresholds for statistical significance (FDR = 0.25). **d-e**, Comparison of frequencies of individual CNAs between the current and previous studies^{8,9,11}. Comparisons were performed in those aged 60-75 years. In (**d**) or (**e**), CNAs in <5% or ≥5% cell fractions were taken into account, respectively. CNAs significantly enriched in either cohort (FDR < 0.1) were indicated by asterisks in (**e**).



Extended Data Fig. 5 | Analysis of SNVs/indels and CNAs in peripheral blood samples in TCGA cohort. a, Distribution of the number of genetic alterations in each subject. Subjects with SNVs/indels alone, with CNAs alone, or with both of them are illustrated by different colors. **b**, Solid lines indicate the prevalence of CH-related SNVs/indels and CNAs, according to age. Colored bands represent the 95% confidence intervals. **c**, The landscape of CH-related SNVs/indels and CNAs. Each row represents genetic alterations or affected chromosomal arms, and each column represents subjects. Subjects without any alterations are omitted. Types of SNVs/indels and CNAs are depicted by different colors. **d**, Distributions of CNAs on all chromosomes are illustrated. Loci of cooccurring SNVs/indels are indicated by arrows. Each horizontal bar represents one CNA. Cooccurring SNVs/indels are indicated by red asterisks. Types of CNAs are depicted by different colors.

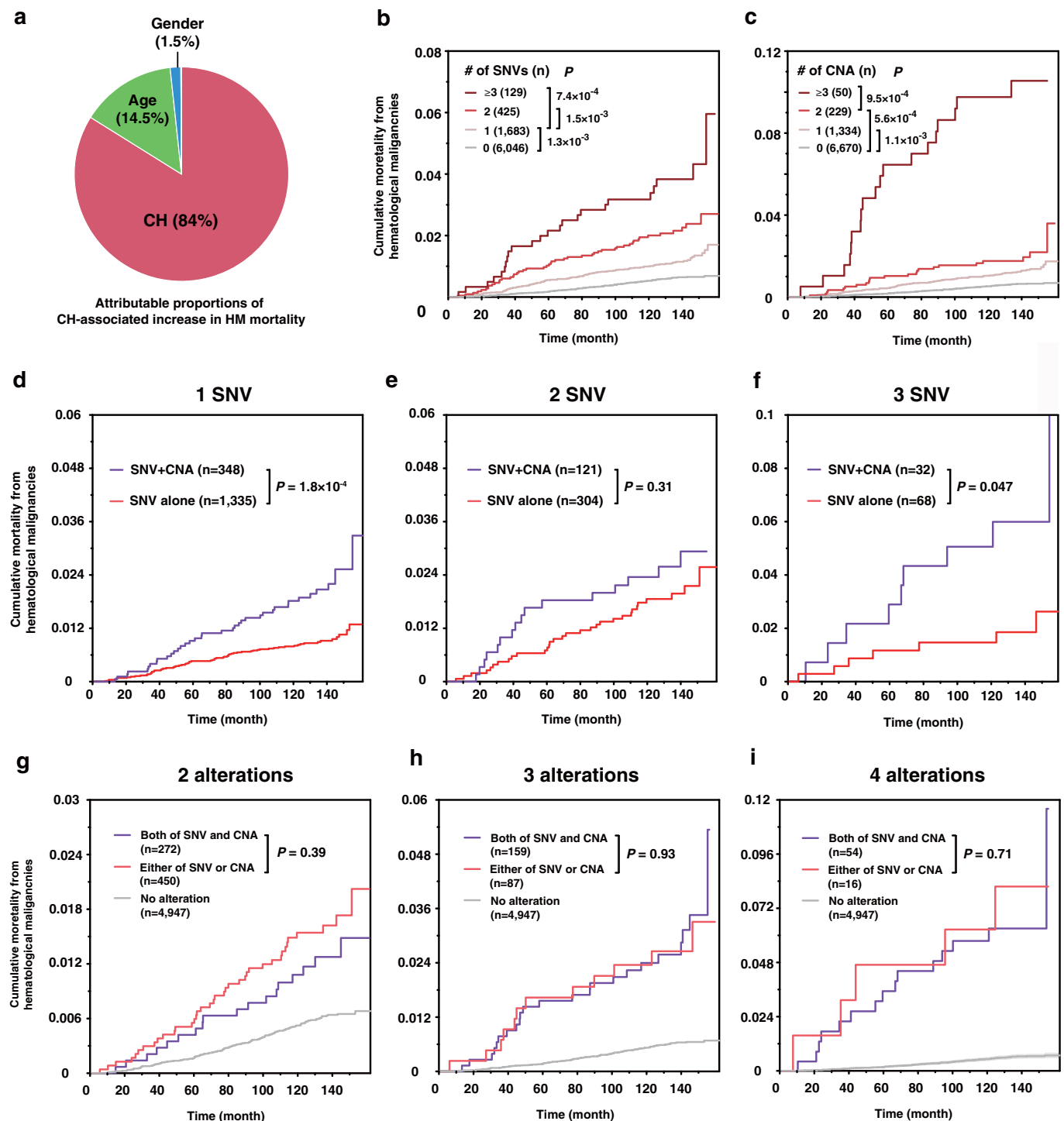


Extended Data Fig. 6 | Interplay between SNVs/indels and CNAs. **a**, Number of subjects with SNVs/indels and CNAs involving the same genes/loci. **b**, Proportion of SNVs/indels associated with CNAs in the same genes/loci. **c**, Cumulative mortality from hematological malignancies. **d**, Cumulative mortality from cardiovascular diseases. **e**, Survival curves for overall survival. **f**, Profiles of CNAs in subjects with SNVs/indels in TP53. Abnormally high or low blood counts (WBC, Platelet, hemoglobin, and hematocrit) are indicated by red or blue, respectively. Numbers of cooccurring CNAs are indicated on the right side (#CNA), where subjects with ≥3 CNAs were highlighted by purple. Subjects without any CNA are abbreviated. **g**, Mortality from hematological malignancies in TP53-mutated cases with or without CNAs in 17p. **h**, Odds ratio for mortality from MDS calculated by multivariate logistic regression in subjects with TP53-involving SNVs/indels. Error bars indicate 95% confidence intervals. We included unclassifiable CNAs involving 17p in 17p alterations (17p alt.) in panel (g-h) because they are most likely to be LOH (UPDs or deletions). TP53-involving SNVs/indels in panel (f-h) included those detected by ddPCR (Supplementary Fig. 3).

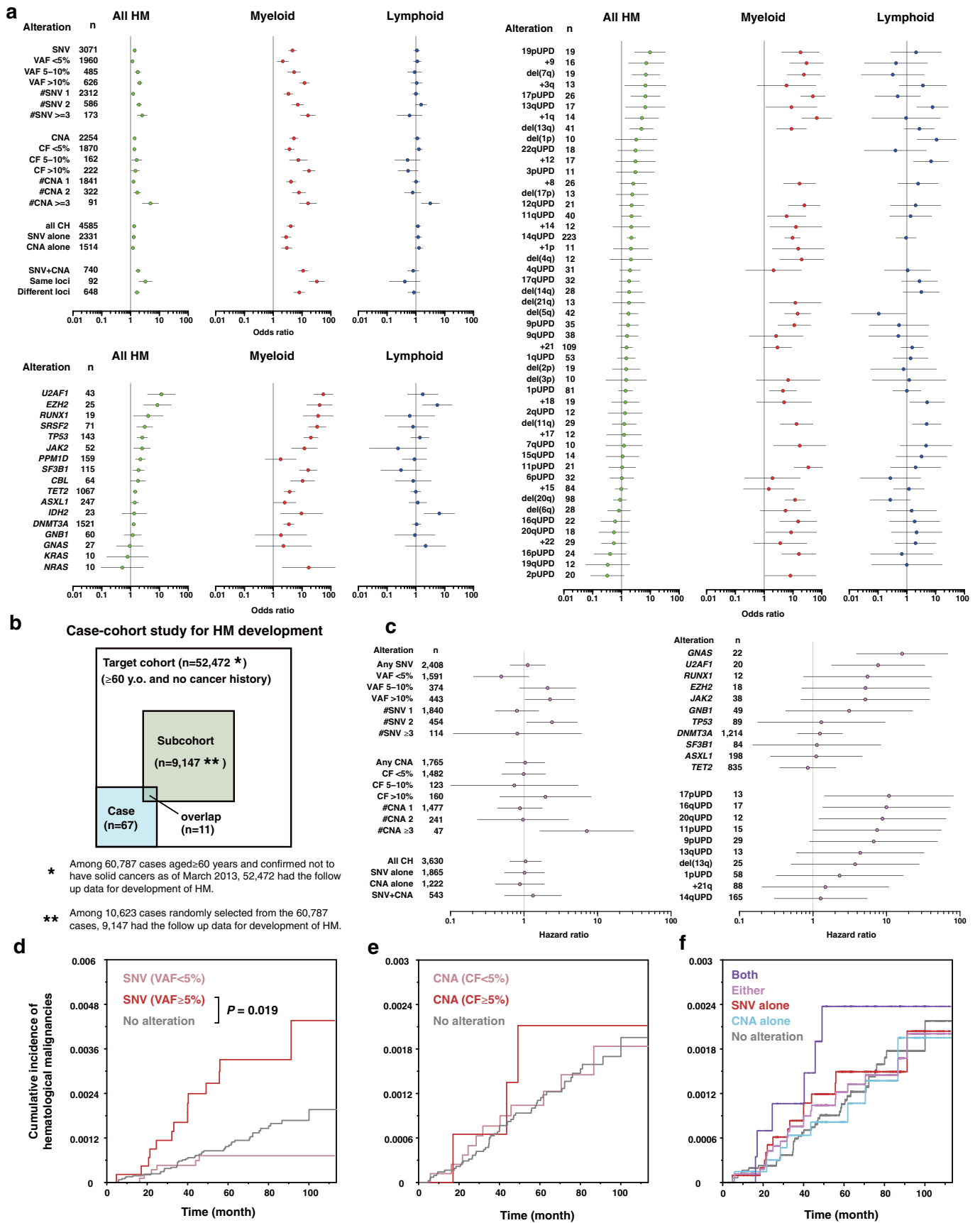


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Genetic alterations in CH and abnormalities in blood counts. **a**, Landscape of SNVs/indels and CNAs in subjects without abnormalities in blood counts (left), in those with any abnormalities in blood counts (middle), and in those with no available blood counts (right). Each row represents a genetic alteration while each column represents a subject. Subjects without any alteration are omitted. Different types of mutations and CNAs are depicted by different colors. **b**, Enrichment of genetic alterations in subjects with abnormalities in blood counts. Sizes of rectangles indicate significance of enrichment. Colors of rectangles indicate odds ratios. The enrichment of alterations was examined by Fisher exact test. Cytopenia (All), subjects with cytopenia in at least one lineage; Cytopenia (Multi), subjects with cytopenia in ≥ 2 lineage. WBC, white blood cell; Hb, hemoglobin; Plt, platelet. **c**, Distribution of blood cell counts in subjects with different CH-related alterations. In all box plots, the median, first and third quartiles (Q1 and Q3) are indicated, and whiskers extend to the furthest value between $Q1 - 1.5 \times \text{the interquartile range (IQR)}$ and $Q3 + 1.5 \times \text{IQR}$. Numbers of subjects (*n*) are indicated below the names of alterations. **d**, Relationships between blood cell counts and VAF of SNVs/indels or cell fractions of CNAs. *P* values are calculated by two-sided *t* test in multivariate linear regression models, taking the effect of age and gender into account. Correction for multiple testing is not performed.

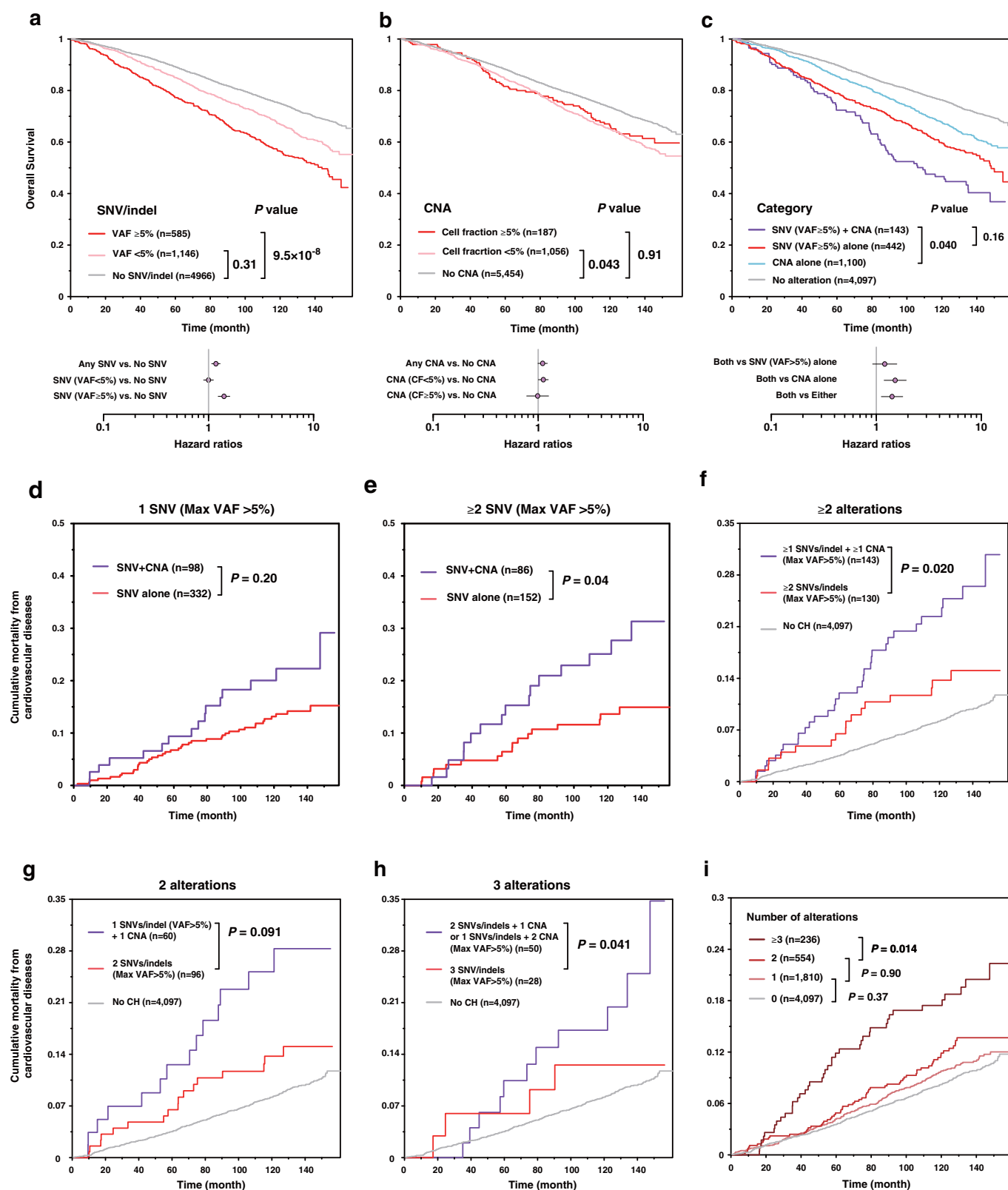


Extended Data Fig. 8 | Impact of CH on mortality from HM stratified by number of alterations. **a**, Pie chart showing the proportions of difference in mortality from hematological malignancies (HM) between subjects with or without CH (Fig. 4a) which are attributable to each prognostic factor (Online methods). **b–c**, Cumulative mortality from HM in subjects with different number of SNVs/indels (**b**), or CNAs (**c**). **d–f**, Cumulative mortality from HM in subjects with both SNVs/indels and CNAs or in those with SNVs/indels alone. Subjects with 1 (**d**), 2 (**e**), or ≥ 3 alterations (**f**) are separately shown. **g–i**, Cumulative mortality from HM in subjects with both SNVs/indels and CNAs or in those with either of them. Subjects with 2 (**g**), 3 (**h**), or 4 alterations (**i**) are separately shown. Throughout the figure, *P* values were calculated by two-sided Wald test and not adjusted for multiple comparison.



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

Extended Data Fig. 9 | Association of CH-related SNVs/indels and CNAs with hematological malignancies. **a**, Odds ratios for the events (death and/or development) of hematological malignancies in case-control study (Extended Data Fig. 1a). Error bars indicate 95% confidence intervals. **b**, Design of case-cohort study for development of hematological malignancies. **c**, Hazard ratios for development of hematological malignancies. Error bars indicate 95% confidence intervals. **d-f**, Effect of SNVs/indels (**d**), CNAs (**e**), and combined SNVs/indels and CNAs (**f**) on the cumulative incidence of development of hematological malignancies. *P* values are calculated by two-sided Wald test. *n*, number of cases with the indicated alterations; SNV + CNA, cooccurrence of both SNVs/indels and CNAs; #SNV, number of SNVs/indels; CF, cell fraction of CNAs; #CNA, number of CNAs.



Extended Data Fig. 10 | Combined effect of SNVs/indels and CNAs on overall survival and cardiovascular mortality. **a–c**, Effect of SNVs/indels (**a**), CNAs (**b**), or combined SNVs/indels and CNAs (**c**) on overall survivals. In the forest plots, error bars indicate 95% confidence intervals. **d–e**, Cumulative mortality from cardiovascular diseases stratified by the number of cooccurring SNVs/indels. **f–h**, Cumulative mortality from cardiovascular diseases in subjects with SNVs/indels (Max VAF $> 5\%$) alone and those with both of SNVs/indels (Max VAF $> 5\%$) and CNAs. Subjects with ≥ 2 (**f**), 2 (**g**), and 3 (**h**) alterations are separately shown. **i**, Cumulative mortality from cardiovascular diseases in subjects with different number of CH-related alterations. Throughout the figure, P values were calculated by two-sided Wald test in (**a–c**, **f–i**), or two-sided Log-rank test stratified by age and gender in (**d–e**), and were not corrected for multiple comparison.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for data collection.

Data analysis

Sequencing data was processed using Genomon v2.6.2. Filtering of mutations was performed as described in Online Methods. "VGAM" package was used to model sequencing error rates as a beta-binominal distribution.

Phasing of long-range haplotype was performed with Eagle2. Copy-number alterations (CNAs) were detected with private C++ codes compiled by gcc-c++v4.8. A standalone software implementation (MoChA) of the algorithm to call CNAs is available at <https://github.com/freeseek/mocha>.

Survival analysis was performed with "survival" package. Events of hematological malignancies were analyzed in a design of "case control study" or "case-cohort study". Cardiovascular mortality and overall survivals were analyzed in a randomly selected 10,623 subjects ("cohort study"). In the analysis of hematological malignancies (mortality or development) and cardiovascular mortality, we performed competing risk regression based on Fine-Gray model.

Custom computational codes used in this study is deposited on GitHub (https://github.com/RSaikiRSaiki/CH_2021).

List of programs and softwares:

Genomon2 pipeline: version 2.6.2 (<https://genomon.readthedocs.io/ja/latest>)
 - Burrows-Wheeler Aligner: version 0.7.8 (<https://sourceforge.net/projects/bio-bwa/>)
 - picard-tools: version 1.39 (<http://picard.sourceforge.net/>)
 - GenomonMutationFilter: v0.2.1 (<https://github.com/Genomon-Project/GenomonMutationFilter>)
 ANNOVAR: 2018-04-16 (<https://annovar.openbioinformatics.org/en/latest/>)
 PolyPhen: v2 (<http://genetics.bwh.harvard.edu/pph2/>)

MutationAssessor: v2.3 (<http://mutationassessor.org/>)
 SIFT: v2.3 (<https://sift.bii.a-star.edu.sg/>)
 Integrative Genomics Viewer (IGV): version 2.4.6 (<http://software.broadinstitute.org/software/igv/>)
 Eagle2: version 2.4.1 (<https://alkesgroup.broadinstitute.org/Eagle/>)
 MoChA: 2020-09-01 (<https://github.com/freeseek/mocha/>)
 R: v.3.6.1 (<https://cran.r-project.org/>)
 car: 3.0.8 (<https://cran.r-project.org/web/packages/car/index.html>)
 VGAM: 1.1.3 (<https://cran.r-project.org/web/packages/VGAM/index.html>)
 survival: 3.1.12 (<http://cran.r-project.org/web/packages/survival/index.html>)
 QuantaSoft: version 1.7 (<https://www.bio-rad.com/webroot/web/pdf/lr/literature/QuantaSoft-Analysis-Pro-v1.0-Manual.pdf>)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We deposited data of both SNVs/indels and CNAs on Japanese Genome-phenotype Archive (JGA; <https://www.ddbj.nig.ac.jp/jga>) under accession code JGAD00000000123, including the information of the cases who did not have SNVs/indels or CNAs. Clinical information cannot be deposited in the public database because of restrictions by Japanese Ministry of Health, Labour, and Welfare and Biobank Japan (BBJ). Clinical information will be provided by BBJ upon request (<https://biobankjp.org/en/index.html>).

In identification of driver SNVs/indels, we used the information from Catalogue Of Somatic Mutations in Cancer (COSMIC) v91 (<https://cancer.sanger.ac.uk/cosmic>), ExAC (<https://gnomad.broadinstitute.org/>), Human Genetic Variation Database (<https://www.hgvd.genome.med.kyoto-u.ac.jp/>), 1000 Genomes Project database (<https://www.internationalgenome.org/>), and dbSNP (<https://www.ncbi.nlm.nih.gov/snp/>).

We also utilized the whole-exome sequencing and SNP-array data of blood-derived DNA taken from 8,023 cancer patients in The Cancer Genome Atlas (TCGA; <https://portal.gdc.cancer.gov/>) to investigate clonal hematopoiesis in the independent dataset.

To compare the frequencies of copy-number alterations (CNAs) between different ethnic groups, we utilized published data:

- Loh, P.R., et al. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* (2020).
- Jacobs, K.B., et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet* 44, 651-658 (2012).
- Laurie, C.C., et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* 44, 642-650 (2012).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No statistical methods were used to predetermine sample size. In determining the number of subjects, we roughly estimated the proportion of subjects who carry both SNVs/indels and CNAs as ~5% at 70 years old based on the frequency in previous publications: Dasai P et al. (2018) and Terao C et al. (2020). Based on this estimation, both alterations will be found ~500 out of 10,000 subjects, which is probably sufficient to examine co-occurring patterns of both alterations.

In total, we enrolled 11,234 subjects.

Among them, 10,623 were randomly selected from the participants of BBJ project, fulfilling the following criteria:

- (i) All subjects were older than 60 years old at the time of DNA sampling.
- (ii) No subjects had history of malignant neoplasms at the time of DNA sampling and during the follow-up periods in BBJ project.

The remaining 611 cases consisted of all cases from the entire BBJ cohort who were confirmed to develop and/or die from HM and were not included in the 10,623 cases above.

Data exclusions

We excluded following subjects based on a pre-established criteria as did in Terao C et al. *Nature* (2020).

- Outliers from East Asian clusters in a plot in which we projected BBJ subjects in combination with 1000 Genomes Project samples in the 2-dimensional principal component space to make the ethnic background of subjects uniform
- Samples genetically identical to another sample
- Samples with call rates less than 0.98 to achieve uniform qualities of SNP array data
- Samples whose reported sex information was not supported by genotypes in the X chromosome.

Replication

We analyzed an independent dataset of 8,023 cancer patients in The Cancer Genome Atlas to reproduce our observation on the combined landscape of SNVs/indels and CNAs in clonal hematopoiesis. In this attempt, we reproduced the statistically significant cooccurrences of SNVs/indels and CNAs, and their recurrent cooccurrences specifically in TET2, JAK2, and TP53.

Randomization

Randomization is not relevant to the study, as there was no experimental group allocation.

Blinding

Blinding was not relevant to the study, as there was no control and treatment arms involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- n/a
- Involved in the study
- ☐ ☒ Antibodies
- ☒ ☐ Eukaryotic cell lines
- ☒ ☐ Palaeontology and archaeology
- ☒ ☐ Animals and other organisms
- ☐ ☒ Human research participants
- ☒ ☐ Clinical data
- ☒ ☐ Dual use research of concern

- n/a
- Involved in the study
- ☒ ☐ ChIP-seq
- ☒ ☐ Flow cytometry
- ☒ ☐ MRI-based neuroimaging

Antibodies

Antibodies used

FITC anti-human CD19 (Catalog number: 560994, Manufacturer: BD Bioscience, Clone: HIB19)
 PE anti-human CD3 (Catalog number: 552127, Manufacturer: BD Bioscience, Clone: SP34-2)
 APC anti-human CD235a (Catalog number: 561775, Manufacturer: BD Bioscience, Clone: HIR2)
 PE-Cy7 anti-human CD34 (Catalog number: 343516, Manufacturer: Biolegend, Clone: 581)
 BV421 anti-human CD33 (Catalog number: 744761, Manufacturer: BD Bioscience, Clone: P67.6)
 BV421 anti-human CD13 (Catalog number: 744862, Manufacturer: BD Bioscience, Clone: L138)

Validation

- anti-human CD19: Nadler LM, Anderson KC, Marti G, et al. B4, a human B lymphocyte-associated antigen expressed on normal, mitogen-activated, and malignant B lymphocytes. *J Immunol.* 1983; 131(1):244-250.
 - anti-human CD3: Blumberg RS, Ley S, Sancho J, et al. Structure of the T-cell antigen receptor: evidence for two CD3 epsilon subunits in the T-cell receptor-CD3 complex. *Proc Natl Acad Sci U S A.* 1990; 87(18):7220-7224.
 - anti-human CD235a: Rogers CE, Bradley MS, Palsson BO, et al. Flow cytometric analysis of human bone marrow perfusion cultures: erythroid development and relationship with burst-forming units-erythroid. *Exp Hematol.* 1996; 24(5):597-604.
 - anti-human CD34: Felschow DM, McVeigh ML, Hoehn GT, et al. The adapter protein CrkL associates with CD34. *Blood.* 2001 Jun 15;97(12):3768-75.
 - anti-human CD33: Andrews RG, Torok-Storb B, Bernstein ID. Myeloid-associated differentiation antigens on stem cells and their progeny identified by monoclonal antibodies. *Blood.* 1983; 62(1):124-32.
 - anti-human CD13: Bradstock KF, Favaloro EJ, Kabral A, et al. Myeloid progenitor surface antigen identified by monoclonal antibody. *Br J Haematol.* 1985; 61(1):11-20.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

BBJ is composed of approximately 200,000 participant who had one of the 47 diseases predefined. This cohort has trend of old age (mean age at enrollment 62.8 years) and relatively high fraction of male subjects (54.1%). Out of them, we further selected 11,234 subjects. The demographic characteristics are summarized in Supplementary Table 1.

Recruitment

BBJ is a multi-hospital-based registry that collected clinical information, DNA, and serum samples from approximately 200,000 patients with one or more of 47 target diseases at a total of 66 hospitals between fiscal years 2003 and 2007. We selected 11,234 subjects from the participants of BBJ project as described above.

Because subjects in BBJ cohort have at least one of the 47 targeted diseases, that can result in selection bias. Nevertheless, the prevalence of the targeted diseases in BBJ cohort was in a good correlation with those in Japanese general population (Nagai, A., et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* 27, S2-S8 (2017).), and the effect of selection bias was, if any, estimated to be small enough to investigate the prevalence of clonal hematopoiesis and effects on clinical covariates.

Ethics oversight

The protocol of this study was approved by
 - Kyoto University Graduate School and Faculty of Medicine, Ethics Committee,
 - RIKEN Yokohama Branch Research Ethics Committee, and

- Ethical review board of the Institute of Medical Science, The University of Tokyo.

Note that full information on the approval of the study protocol must also be provided in the manuscript.