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

Ryunosuke Saiki[1], Yukihide Momozawa[2], Yasuhito Nannya[1], Masahiro M Nakagawa[1,3], Yotaro Ochi[1], Tetsuichi Yoshizato[1], Chikashi Terao[4], Yutaka Kuroda [5], Yuichi Shiraishi[6], Kenichi Chiba[6], Hiroko Tanaka[7], Atsushi Niida[8], Seiya Imoto[9], Koichi Matsuda[10], Takayuki Morisaki[11], Yoshinori Murakami[11], Yoichiro Kamatani[4,10], Shuichi Matsuda[5], Michiaki Kubo[12], Satoru Miyano[7], Hideki Makishima[1], Seishi Ogawa[1,3,13]

[1]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[2]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[3]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan

[4]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[5]Department of Orthopaedic Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[6]Division of Cellular Signaling, National Cancer Center Research Institute, Tokyo, Japan

[7]Department of Integrated Data Science, M&D Data Science Center, Tokyo Medical and Dental University, Tokyo, Japan

[8]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

[9]Division of Health Medical Intelligence, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan

[10]Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan

[11]Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan

[12]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[13]Department of Medicine, Centre for Haematology and Regenerative Medicine, Karolinska Institute, Stockholm, Sweden

Correspondence should be addressed to:

Seishi Ogawa (sogawa-tky@umin.ac.jp).

Conflict of interest disclosure:

The authors declare no conflict of interest.

Text word count: 4686 words

Number of figures: 6; Number of references: 50

Abstract

Clonal hematopoiesis (CH) in apparently healthy individuals is implicated in the development of hematological malignancies (HM) and cardiovascular diseases. Previous studies of CH have analyzed either single-nucleotide variants and indels (SNVs/indels) or copy number alterations (CNAs), but not both. Here, by combining 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 bi-allelic alterations of these genes and were 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.

Introduction

The presence of clonal components in an apparently normal hematopoietic compartment, or clonal hematopoiesis (CH), has been drawing an increasing attention of recent years[1,2]. Although suggested only indirectly by skewed chromosome X inactivation in early studies[3-7], CH has recently been demonstrated by detecting copy number alterations (CNAs) in the peripheral blood samples from large cohorts of individuals without blood cancers using single-nucleotide polymorphism (SNP) array data from genome-wide association studies (GWAS)[8-11]. Showing a substantial overlap to those characteristic of hematological malignancies (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 single-nucleotide variants and indels (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 to identify CNAs and somatic SNVs/indels, a complete registry of CNAs and SNVs/indels associated with CH has been elucidated, which are thought to involve virtually every individual in the extreme elderly[20,21]. However, to date, no studies have evaluated both CNAs and SNVs/indels together at a comparable sensitivity in a large cohort of a general population, although they have recently been investigated in a cancer population, where many had been treated with chemo/radiotherapy[22]. 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 shaped 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 better understanding of CH and its implication in HM and CVD.

In the present study, for the purpose of delineating the combined landscape of common 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 the Biobank Japan (BBJ)[23], which had been SNP-typed for GWAS studies for common diseases, including hypertension, diabetes, autoimmune diseases and several solid cancers[23]. 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 Table 1,2) (https://biobankjp.org/info/pdf/sample_collection.pdf)[23]. Among these 10,623 were randomly selected from 60,787 cases who were 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 develop and/or die from HM as of March 2017. The remaining 611 consisted of all cases from the entire BBJ cohort who were confirmed to develop and/or die from HM as of the same date but were not included in the randomly selected 10,623 cases. In total, 672 cases were reported to have HM in the entire BBJ cohort, which included 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 (Online methods).[24] 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 ~800x (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%) ≥3 SNV/indels (Fig. 1a). Their VAFs 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 co-mutated than expected only by chance (OR: 1.53-6.53, q<0.05) (Extended Data Fig. 2d). Of interest, many of these combinations are also co-mutated in myeloid neoplasms with large 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 or not these combinations of SNVs/indels affected the same cell populations in the vast majority of cases (Extended Data Fig. 2e-i).

CNAs data were available from the previous study[21], 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 the 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, Online 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 higher than that in the entire BBJ cohort[21], even though age-stratified frequencies were almost equivalent between both cohorts (Fig. 1b). The sizes of detected CNAs ranged from 0.01 to 248 Mb (median: 34.4), depending on density of informative SNPs and their haplotype configuration, tumor contents, and performance of SNP probes (Supplementary Fig. 1d). Estimated mutant cell fractions (MCF) for CNAs were ranged from 0.2% to 93.2% with a median of 2.0% with FDR<0.05, where a substantial number (n=461) of CNAs were seen in a cell fraction of ≤1%, which was below the limit of detection for SNVs/indels. Thus, smaller clones were detected through CNAs, particularly copy-neutral loss-of-heterozygosity (CN-LOH) or uniparental-disomy (UPD), compared with 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, such as 14qUPD and +21, showed higher frequencies than reported in western populations, which is likely due to a higher sensitivity for detecting CNAs in this study compared with that in previous studies in western populations[8-11]; when confined to lesions with ≥5% cell fractions, the difference across studies becomes less conspicuous for many CNA targets (Extended Data Fig. 4d,e). Nevertheless, even considering the different sensitivities, several CNAs, including +15, del(14q), del (9q), del(20q) and del(13q), still showed a different frequency across studies in both populations [21], suggesting an ethnic difference in positive selection of CH-related CNAs (Extended Data Fig. 4e), although the exact genetic basis of the ethnic 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 randomly selected 10,623 cases who were ≥60 years of age with no reported cancer history and in 376 (56%) of 672 cases who developed HM, where 38 of the 376 were <60 years old. Combining both lesions, more subjects (n=1,503) had two or more lesions than judged by SNVs/indels (n=759) or 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 larger in individuals with abnormal blood counts, particularly 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 solid cancer patients from The Cancer Genome Atlas (TCGA; https://portal.gdc.cancer.gov/), although the sensitivity of CH-lesions, particularly CNAs, was substantially lower than the current study due to a lower coverage of exome sequencing and a less accurate haplotype phasing required for sensitive CNA detection (Extended Data Fig. 5a,b,c).

Accounting for 7% of the total cohort and 16% of all CH(+) cases, 740 individuals harbored both types of lesions, which were significantly more frequent than expected only by chance (Extended Data Fig. 2j), even after their age was adjusted (odds ratio [OR]=1.3; *P*=0.0003, age-stratified permutation test) (Supplementary Fig. 3, Online 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 larger 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) were confirmed in the TCGA cohort. These observations raise a 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 their combinations per se (Fig. 1e).

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 sum[28], of which 8 were combinations between 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 small clone size of both lesions, which would be better addressed using single cell-based sequencing. A representative case was 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 only by chance (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 8 genes/loci (Extended Data Fig. 6a), of which 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, Extended Data Fig. 5d). These cooccurrences were still statistically significant when the inflation of 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 escape from detection due to lower sensitivity of detecting SNVs/indels than CNAs (Supplementary Fig. 1a,c). In fact, using highly sensitive ddPCR assay targeting mutational hotspots, SNVs in *JAK2* and *TP53* were confirmed in 8 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 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, none of these cases satisfied the pigeonhole principle or no samples were available for single cell-sequencing analysis to directly confirm this at a 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, Online 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* (14q11.2 involving the) 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 lymphomas[34], particularly in follicular-helper T-cell-derived lymphomas, such as angio-immunoblastic T-cell lymphoma (AITL), which are also seen in *Tet2* knockdown mice[35]. 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 MDS[36,37].

Clinical associations with CH

Next, we investigated common demographic factors that may 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 impacted 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 non-cytopenic 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 counts 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 sort (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 was associated with lower white blood cell (WBC) and platelet counts and larger mean corpuscular volume (MCV) value (Fig. 3b). The number of all co-occurring alterations, SNVs/indels or CNAs, and VAF of SNVs/indels predicted significantly lower hemoglobin values, while MCF of CNAs predicted larger MCV and lower MCHC values (Fig. 3b,c, Extended Data Fig. 7c). As for individual alterations, SNVs/indels in *JAK2* were significantly correlated with high platelet counts (q<0.001) even when the analysis was confined to the individuals with normal blood counts. Moreover, we found significant associations of lower hemoglobin values with SNVs/indels in *TP53*, *PPM1D*, *SF3B1*, and *U2AF1*, and 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 larger MCV (Fig. 3b, Extended Data Fig. 7c). VAF or cell fractions of SNVs/indels and CNAs were also predictive of the changes in hemoglobin, platelet counts, or MCV (Fig. 3b, Extended Data Fig. 7d), while VAF of *JAK2*-involving SNVs/indels did not correlated with platelet counts (Supplementary Fig.

9). SNVs/indels in *TET2* alone were not associated with a reduced hemoglobin value (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 likely attributable to biallelic *TET2* alterations (Fig. 3b,d). We also tested the relationships between CH and values of other blood tests to reveal a negative correlation between *GNB1*-involving SNVs and uric acid concentration achieved 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 see this, we investigated the effect of CH on the cumulative mortality from HM using the Fine and Gray regression modeling in a case-cohort design[38], where 7,937 of the 10,623 cases were regarded as a subcohort that were randomly selected from 43,662 cases who had been followed up for survival and cause of deaths on the basis of the vital statistics of Japan[39] (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 of HM mortality between CH-positive and -negative subjects were mostly explained by CH itself, although age and gender made 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(+) vs. CH(－) cases was ~two times higher from myeloid neoplasms (0.39%) than lymphoid (0.21%) neoplasms (Fig. 4b) and the hazard ratio between CH(+) and CH(－) cases was >2.5 times larger for myeloid (3.64) than 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 targeted 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, 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 a higher HM mortality independently of the latter (Fig. 4d, Fig. 5a), which was in line with a previous observation that SNVs/indels correlated with development of HM only when they exhibited sufficiently large VAFs (≥1%)[40]. In univariate analysis, the largest risk of HM mortality was conferred by SNVs/indels of *U2AF1*, *EZH2, RUNX1, SRSF2, TP53*, and +1q[11,14,21] (Fig. 5b-d, Supplementary Fig. 12, 13). As expected from a ~2 times larger attributable mortality for myeloid than lymphoid malignancy, HRs and ORs were higher in myeloid than lymphoid HM for most of the lesions, with an 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 it was 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 the number of CH-related lesions and on the severity of cytopenia; as large as 3.4% 10-year HM mortality was observed for those with multi-lineage cytopenia and multiple CH-related SNVs/indels and CNAs, compared with 0.46% for those with normal blood count lacking CH-related lesions.

The presence of both SNVs/indels and CNAs was associated with a 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). It 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-occurred, i.e., SNVs/indels vs. CNAs. In fact, the HM mortality significantly correlated with the total number of CH-related lesions and 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 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 a higher HM mortality, compared with other combinations of SNVs/indels and CNAs (Extended Data Fig. 6c). The increase mortality was largely explained by those affecting *TP53*. However, even excluding *TP53*-involving SNVs/indels and CNAs, the combinations of lesions affecting the same locus showed a higher HM mortality than other SNVs/indels and CNAs combinations. Of interest, *TP53*-involving SNVs/indels also exhibited significant associations with del(5q) and multiple (≥3) CNAs mimicking a complex karyotype (Fig. 2a, Extended Data Fig. 6f), which together with 17pLOH, are among the most common lesions associated with *TP53* alterations in a variety of myeloid neoplasms with 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 Fig. 1a, 9a). A small number of cases in which the onset of HM was recorded due to incomplete follow-up and exclusion of MDS and MPN 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 were 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, comorbidities (diabetes mellitus, hypertension, and dyslipidemia), history of smoking/drinking, and versions of SNP array. In accordance with the previous reports[13], the presence of CH-related SNVs/indels with large clone size (VAFs ≥ 5%) were associated with an elevated cardiovascular and all-cause mortality (HR=1.36, 95%CI:1.09-1.71 for cardiovascular mortality; HR=1.41, 95%CI:1.24-1.60 for all-cause mortality) (Fig. 6a, Extended Data Fig. 10a). In support of this, we observed significant association of SNVs/indels with hypertension (Fig. 6b), which was independent of known risk factors for hypertension, including older age, a higher BMI, and diabetes. By contrast, regardless of their clone size, CNAs alone did not seem to affect cardiovascular or all-cause mortality (Fig. 6c, Extended Data Fig. 10b). However, CNAs in combination with SNVs/indels with ≥5% VAFs 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 or not they involved the same locus (Extended Data Fig. 6d,e). In multivariate analysis, the combined effect of both lesions was independent of the number of cooccurring SNVs/indels (HR=1.77, *P*=0.012, Extended Data Fig. 10d,e) and the total number of alterations (Extended Data Fig. 10f-h). Given no 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, the total number of CH-related lesions did not correlate with cardiovascular and all-cause mortality, except for a significantly higher mortality for ≥3 CH-related lesions (Extended Data Fig. 10i), likely 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 of both lesions.

Discussion

Combining targeted deep sequencing of major CH-relate genes and SNP array-based copy number analysis of blood-derive 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 SNV/indel and CNA. A case-cohort study design enabled an accurate estimation of CH-associated cumulative HM mortality in a large general cohort of elderly individuals (>43,000) including >400 cases who developed HM, substantially saving the cost and effort of sequencing, where only ~8300 (~18%) individuals/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/cohort)[12,13], the estimation of HM mortality in individuals with CH-related SNV/indels was substantially more accurate with a much smaller confidential interval for both myeloid and lymphoid malignancies, where the mortality attributable to CH was mostly explained by myeloid malignancies regardless of type of CH-related lesions. Estimation of odds ratios for CH(+) vs. CH(−) cases were even more accurate with a total of 672 HM events in a case-control study setting.

Including both types of lesions, CH was found in as many as 40% of a general population of ≥60 years of age, of which 11% had ≥10% clone size. As a whole, SNVs/indels and CNAs co-occurred more frequently than expected only by chance. 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 in an early stage during leukemogenic evolution. Co-occurrence of *TET2*-involving 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 excluding the subjects having these combinations affecting the same gene, SNVs/indels and CNAs significantly co-occurred (*P*=0.0042). Given that most of the 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 a possibility that CNAs might just 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 CNAs[40], is typically exemplified by biallelic alterations in *DNTM3A, TET2, JAK2,* and *TP53*, caused by LOH affecting the mutated locus. However, the effect of combined 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 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 that CNAs alone did not affect the mortality. Because the effect of SNVs/indels on cardiovascular mortality depended on their VAFs, 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 the effects of clone size was adjusted.

Potential caveats in the current study include a limited number of CH-related genes analyzed (n=23), a compromised sensitivity of detecting focal CNAs, and the study population exclusively including individuals over 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 with 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 group might not necessarily be the limitations, but rather contributed to efficient analyses of comprehensive analysis of CH-related alterations in a large number of cases to investigate their effects on clinical outcomes at an acceptable cost. However, clearly more comprehensive studies with unbiased sequencing and improved copy number detection including all age groups should be warranted to elucidate the full spectrum of CH-related alterations in future studies.

Acknowledgement

This work was supported by the Japan Agency for Medical Research and Development (AMED) (JP15cm0106056h0005, JP19cm0106501h0004, JP16ck0106073h0003, JP19ck0106250h0003 to S.O.; JP17km0405110h0005 and JP19ck0106470h0001 to H.M.; JP19ck0106353h0003 to Y.N.) and the Core Research for Evolutional Science and Technology (CREST) (JP19gm1110011 to S.O.); the Ministry of Education, Culture, Sports, Science and Technology of Japan; the High Performance Computing Infrastructure System Research Project (hp160219, hp170227, hp180198 and hp190158 to S.O. and S.M.) (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 (JSPS); Scientific Research on Innovative Areas (JP15H05909 to S.O. and S.M.; JP15H05912 to S.M.) and KAKENHI (JP26221308 and JP19H05656 to S.O.; JP16H05338 and JP19H01053 to H.M.; JP15H05707 to S.M.); the Takeda Science Foundation (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 Biobank Japan, the Institute of Medical Science, the University of Tokyo. The super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo. We gratefully acknowledge Keitaro Matsuo in Aichi Cancer Center Research Institute (Nagoya, Japan) who suggested the design of case-cohort study for the estimation of cumulative mortality from and incidence of hematological malignancies. We are grateful to the TCGA Consortium and all its members for making their invaluable data publicly available.

Author contributions

R.S., H.M., and S.O. designed the study. K.M., Y.K., T.M., and Y.M. provided DNA samples and clinical data. Y.K. and S.M. provided bone marrow samples. T.C., and Y.K. performed copy-number analysis. Y.M. 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., N.A., S.I., and S.M. 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.

Online methods

Sample ascertainment

All subjects in this study were derived from BioBank Japan (BBJ) project, a multi-hospital-based-registry[23]. BBJ project enrolled approximately 200,000 individuals with at least one of 47 target diseases between fiscal years 2003 and 2007. From 179,417 participants of 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 who were aged ≥60 years at the time of sample collection and were confirmed not to have solid cancers as of March 2013. Out of the randomly selected 10,623 cases, 61 were recorded to develop or die from HM. The remaining 611 subjects, all of whom were recorded to have 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 which were recorded to develop and die from HM, respectively. Subjects' demographic summary was presented in Supplementary Table 1. The numbers of subjects with individual targeted diseases were listed in Supplementary Table 2. Written informed consent had been obtained from all participants. The protocol of this study was approved by 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

To detect CH-associated driver mutations, we performed multiplex PCR-based targeted sequencing, as previously described[24]. 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 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 second PCR[43,44]. We performed multiplex PCR using different primer pools to cover all coding regions of the 23 genes. Then we performed second PCR with primer sequences 5'-AATGATACGGCGACCACCGAGATCTACACxxxxxxxxACACTCTTTCCCTACACGACGCTCTTCCGATCTCTG-3' and 5'-CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAC-3', where xxxxxxxx represents 8-bp barcodes. All second PCR products were pooled for one sequencing run. After each library was purified using Agencourt AMPure XP (Beckman Coulter), we obtained 2x150-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 0.7.8 (https://sourceforge.net/projects/bio-bwa/), version 0.7.8, with default parameter settings. Mutation calling was performed with Genomon2 pipeline version 2.6.2 (https://genomon.readthedocs.io/ja/latest), picard-tools version 1.39 (http://picard.sourceforge.net/), and GenomonMutationFilter v0.2.1 (https://github.com/Genomon-Project/GenomonMutationFilter), as previously reported[33,45-47]. Extracted mutations were annotated with ANNOVAR (https://annovar.openbioinformatics.org/en/latest/). Then, we adopted variants fulfilling the following criteria:

(i) Number of variant reads ≥ 10 (≥5 for TCGA dataset) †

(ii) Variant allele frequency (VAF) ≥ 0.5%†

(iii) Non-synonymous variants within coding-sequence or splice-site variants

(† For calculation of read counts and VAFs, we only counted base calls fulfilling Mapping Quality score ≥ 40, and Base Quality score ≥ 20.)

To further exclude false positive calls due to sequencing artifacts, we modeled site-specific err or rates as beta-binominal distribution, using R package, VGAM (1.1.3, https://cran.r-project.org/web/ packages/VGAM/index.html). Parameters for beta-binomial distribution were determined by maximum likelihood method[48] based on the read counts in all samples. Mutation calls whose VAFs were signif icantly deviated from background-error distribution ($P_{\text{beta-binomial}} \leq 10^{-6}$) were regarded as true mutatio ns.

Additionally, variants always appeared within similar ranges of VAFs (especially <1%, or >40%) were likely to be sequencing artefacts 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:

(i) Candidates observed in ≥5 samples

(ii) Mean VAF <1%, or >40%, or coefficient of variation of VAFs < 0.5.

The candidates fulfilling the quality filter noted above were included in the subsequent analyses if they fulfil one of the following criteria for driver mutations[33]:

(i) Candidates resulting in amino-acid substitutions which were registered in the Catalogue of Somatic Mutations in Cancer (COSMIC) v91 databases (https://cancer.sanger.ac.uk/cosmic) for ≥ 5 counts

(ii) Candidates which fulfill the Criteria 1 and at least one of the Criteria 2

Criteria 1

Candidates which were not registered in public databases, including dbSNP138 (https://www.ncbi.nlm.

nih.gov/snp/), the 1000 genomes project as of 2014 Oct (https://www.internationalgenome.org/), Human Genome Variation Database (HGVD; https://www.hgvd.genome.med.kyoto-u.ac.jp/), and The Exome Aggregation Consortium (ExAC; https://gnomad.broadinstitute.org/).

Criteria 2

a) Candidates located on the non-repeat region with VAFs ≥4% <40% or ≥60% <96%

b) Nonsense, frameshift, or splice-site candidates

c) Candidates which were 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 mutation calling**

To benchmark the performance in detection of low-VAF mutations, we performed *in silico* simulation. Mixing 2 bam files with variable proportions, we diluted 750 heterozygous SNPs and artificially created low-VAF mutations (ranging from 0.5% to 5%). Each diluted SNPs were classified into 6 bins according to sequencing depths (x100-x300, x300-x500, x500-x750, x750-x1000, x1000-1500, and x1500-), and sensitivities were calculated separately for the 6 bins. We calculated sensitivity as a fraction of detected variants within all simulated variants:

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

($SN_{VAF = x\%}$: sensitivity for variants with x% VAFs,

$TP_{VAF = x\%}$: number of detected SNPs whose VAFs were diluted to x%,

$FN_{VAF = x\%}$: number of missed SNPs whose VAFs were diluted to x%).

Together with sensitivity, we calculated specificity by sampling genomic positions without known SNPs (n = 5000/simulation). We counted mutation calls on these positions as false positives, and calculated the specificity as follows:

$$SP = 1 – FP / N$$

(SP: specificity,

FP: number of false-positive mutation calls,

N: number of sampled genomic positions).

To draw receiver operator characteristic (ROC) curves, we calculated sensitivities and specificities for 9 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 analysis pertaining CNAs are based on the result in previous publication[21], in which blood derived DNA samples from the 11,234 subjects were examined by either of three different versions of microarrays: 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 described[21]. Based on long-range haplotype information and LRR/BAF values, we detected allelic imbalances and classified them into duplications, deletions, and UPDs, with false discovery rate around 5%.[10,21] Meanwhile, copy-number analysis of 8023 samples from TCGA cohort was performed with A standalone software, MoChA (https://github.com/freeseek/mocha/). Because the power to detect allelic imbalances exceeded the power to distinguish UPD from copy-number gain or loss, CNAs were designated as "unclassifiable" when we could not assign them into specific types of CNAs. In the analyses where the exact discrimination between UPD, duplication, or deletion (e.g., lesion-specific analysis in Fig 2a, 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 (e.g., Fig. 4d, 5b, and 6c), unclassifiable CNAs were regarded to be smaller than the thresholds. When we analyzed CNAs in terms of their cell fractions (e.g., Fig. 1d,e, 5a), unclassifiable CNAs were excluded. Otherwise, we did include those unclassifiable CNAs in the analysis (e.g., Fig 1a-c, 2b-e, 3c-d, 4, 6b,d). Based on the detected CNAs, we determined chromosomal regions significantly affected with CNAs by PART (parametric aberration recurrence test)[49].

Definition of abnormalities in blood counts

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

(i) White blood cells (/µL): ≥10000, or <3000

(ii) Hemoglobin (g/dL): ≥16.5 (male), ≥16 (female), or <10

(iii) Hematocrit (%): ≥50

(iv) Platelet (10000/µL): ≥50, or <10

These cutoffs on blood counts were adopted from diagnostic criteria for myelodysplastic syndromes or myeloproliferative neoplasms[50]. Out of subjects with available counts for all of WBC, hemoglobin, hematocrit, and platelet (n = 8,345), 7,031 subjects (84.3%) had normal blood cell counts.

Analysis of lineage-sorted samples

Frozen bone marrow Frozen bone marrow was thawed in Dulbecco's Modified Eagle Medium (Sigma-Aldrich) containing 10% of foetal bovine serum (FBS, biosera) and 1% of Penicillin-Streptomycin solution

(ThermoFisher). After the cell pellets were washed with PBS containing 2% FBS, the cells were stained with an 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 mixed 500 µL of the filtered cell suspension in PBS containing 2% FBS was mixed with 5 µL of Propidium Iodide Staining Solution (BD Bioscience), which was then sorted with the FACSAria III cell sorter (BD Bioscience). The antibodies used in flow cytometry are listed in Supplementary Table 4. For digital droplet PCR (ddPCR) and amplicon-sequencing, we sorted myeloid, erythroid, T cell, and B cell fractions and gDNA 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 re-suspended in StemSpan Serum-Free Expansion Medium (STEMCELL Technologies) at 400–1,600 cells/µL, which was then applied into 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, Masahiro M Nakagawa, Ryosaku Inagaki, et al.).

ddPCR

For ddPCR, predesigned probes were purchased from BioRad. We mixed 50 ng of gDNA with enzymes (ddPCR Supermix for Probes (no dUTP), BioRad) 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 (BioRad, https://www.bio-rad.com/webroot/web/pdf/lsr/literature/QuantaSoft-Analysis-Pro-v1.0-Manual.pdf). Catalogue numbers of probe mix are shown in Supplemental Table 6.

Statistical analysis

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

Age-stratified permutation test for cooccurrences of CH-related alterations

We tested the significance of cooccurrences between SNVs/indels and CNAs under the stratification by subjects' age, because age-dependent frequencies of both CH-related alterations can confound their cooccurrences. First, we stratified subjects into 41 bins according to their age (60, 61, …, 100 years old) and calculated frequencies of SNVs/indels, CNAs, and their cooccurrences within each bin. In 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 number of cooccurrences were summed up across all age bins. By repeating this process, we obtained null random distribution of the number of subjects with cooccurring SNVs/indels and CNAs. Comparing the null distribution and the actual number of cooccurrences, we obtained *P* value for significance of cooccurrences between SNVs/indels and CNAs. Significant cooccurrences of multiple CH-related alterations was also demonstrated in a similar way, in which we counted the total number of CH-related alterations within each age bin. In single iteration, these alterations were randomly re-assigned to the subjects retaining the total number of alterations in each bin. Then, the number of subjects to whom multiple alterations were assigned was counted across all bins. *P* value was calculated by comparing the actual number of cases with ≥2 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 VAFs of SNVs/indels than cell fractions of CNAs in 51 of the 55 cases, which suggested they were likely to be acquired in the same cells and resulted in biallelic alterations (Supplementary Figure 8a,b). To examine how many of the 51 cases should be explained by cell-level coexistence of SNVs/indels and UPDs, we performed random simulation on their clone sizes putting a null hypothesis, $H_0(x)$: SNVs/indels and UPDs were independently acquired in at least x cases (x=5,4,…,55). *P* value for $H_0(x)$ was calculated assuming VAFs of SNVs/indels and cell fractions of UPDs follows independent distributions (Supplementary Figure 8c-e). We searched for the maximum x with which *P* value for $H_0(x)$ was below 0.05 to obtain minimum estimate of the number of cases in which cell-level coexistence of SNVs/indels and UPDs was expected (Supplementary Figure 8f).

Calculation of adjusted VAF

We observed significant cooccurrences of *DNMT3A*/2pLOH, *TET2*/4qLOH, *JAK2*/9pUPD, and *TP53*/17pLOH, which suggested biallelic alterations of these genes were positively selected in CH. However, the frequencies of these cooccurrences might be overrepresented because underlying LOH inflated VAFs of SNVs/indels and resulted in higher sensitivity to detect SNVs/indels when LOH coexisted. To exclude the effect of VAF inflation, we tested the significant co-occurrence of these combinations of SNVs/indel and LOH, focusing on those SNVs/indels having ≥5% VAFs, for which almost 100% of detection sensitivity would be expected (Supplementary Fig. 1a), where inflated VAFs due to LOH were adjusted according to the following formula by calculating the cell fraction having LOH (Supplementary Fig. 6a):

$VAF_{adjusted} = VAF_{observed} – CF_{LOH} / 2$: with UPD

$VAF_{adjusted} = VAF_{observed} * (1 – CF_{LOH} / 2)$: with deletion

$VAF_{adjusted} = VAF_{observed}$: without LOH).

(CF_{LOH}: cell fractions of LOH cooccurring in the same genes)

Focusing on SNVs/indels with large adjusted VAFs (> 5%), we examined the significance of cooccurrences 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 baseline characteristics of subjects (age, sex, history of smoking and drinking). Information regarding the history of smoking and drinking were based on self-report questionnaires at DNA sampling. First, we performed univariate logistic regressions for 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 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 R package "car" (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 perfumed in 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 stepwise method from known risk factors and blood test values available for ≥70% of the subjects: presence of SNVs/indels and CNAs, age (+10 years), gender, BMI (+5), history of smoking and drinking (based on self-report questionnaires), white blood cells, red blood cells, hemoglobin, hematocrit, MCHC, platelet, aspartate aminotransferase (AST), alanine aminotransferase (ALT), lactate dehydrogenase (LDH), creatinine, blood urea nitrogen, total cholesterol, and glucose.

Survival analysis

We evaluated the effects of CH-related mutations, CNAs, and their combinations on mortality from HM, all-cause mortality, and cardiovascular mortality. To define mortality from hematologic malignancies, we included diagnoses within ICD10 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), I48-49 (arrythmia), I50 (heart failure), I60-I67, I69 (cerebrovascular diseases), I70-I72 (aortic atherosclerosis, aortic aneurysm, aortic dissection), and I74 (peripheral artery diseases). In analysis on all-cause mortality, we performed Cox proportional hazards regression using the R package, "survival" (3.1.12, http://cran.r-project.org/web/packages/survival/index.html). In analysis of HM events (mortality or development) or mortality from CVD, we performed competing risk regression based on 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 described[38] (Extended Data Fig. 1b, 9b), including all subjects with HM events within the target cohort. Contribution of each factor to the increase in HM mortality was estimated by calculating log (hazard) of the corresponding factor and averaged for all patients as previously described[33]. Concretely, the contribution of factor i for patient j, designated as $C_{i,j}$, is estimated

$$\beta_i \cdot (x_{i,j} - x_{i,\mathrm{Median}})$$

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

$$\Sigma_{j\ \in\ \text{subjects with CH}}\ C_{i,j}/N_{\mathrm{CH}}\ - \Sigma_{j\ \in\ \text{subjects without CH}}\ C_{i,j}/N_{\text{non-CH}},$$

where N_{CH} and $N_{non\text{-}CH}$ stands for the number of subjects with or 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 the randomly selected 10,623 subjects. 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, and hypertension, history of tobacco smoking and alcohol drinking, and version of SNP array were included in the models for all-cause and cardiovascular mortalities.

Data availability

Tables of somatic SNVs/indels and CNAs detected in this study are deposited on Japanese Genome-phenotype Archive (JGA) 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).

Code availability

Custom computational codes to reproduce figures from the manuscript is available at https://github.com/RSaikiRSaiki/CH_2021.

Figure Legends

Fig. 1 | Landscape of SNVs/indels and CNAs in clonal hematopoiesis.

a, Distribution of the number of genetic alterations in each subject. Subjects with SNV/indels alone, with CNAs alone, or with both of them 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 subjects with and without HM events, respectively. Colored bands represent the 95% confidence intervals. c, Number of cooccurring alterations in those with subjects with abnormalities in blood cell counts, or cytopenia. d, Maximum cell fraction of CH-related alterations in CH-positive 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 cooccurring alterations. Cell fractions of SNVs/indels are defined as 2 times VAF. Those with both of SNVs/indels and CNAs are shown in purple, while those with either are shown in blue. In panel (d,e), unclassifiable CNAs were excluded because we cannot calculate their precise cell fractions. The box plots indicate the median, first and third quartiles (Q1 and Q3) and whiskers extend to the furthest value between Q1 – 1.5×the interquartile range (IQR) and Q3 + 1.5×IQR. In (c-e), *P* values were calculated by two-sided Wilcoxon rank-sum test and not adjusted for multiple comparison.

Fig. 2 | Cooccurrences of SNVs/indels and CNAs in clonal hematopoiesis.

a, The correlations between individual SNVs/indels and CNAs. The size of rectangles indicates the significance of correlations. Red rectangles represent positive correlations while blue rectangles represent negative correlations. Combinations of SNVs/indels and CNAs seen in 5 or more subjects are indicated by asterisks. b-e, The distributions of CNAs on chromosome 2 (b), 4 (c), 9 (d), and 17 (e). Horizontal bars represent CNAs, and cooccurring 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 which cannot be classified into any of UPD, deletion, or duplication are indicated as unclassifiable CNAs (gray).

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. Sizes and colors of rectangles represent the significance and effect size calculated by two-sided Wald test. Asterisks indicate the clinical factors significantly correlated with each alteration in multivariate logistic regression ($P<0.05$). b, Correlations between genetic alterations and blood counts. The sizes and colors of rectangles indicate the significance, and effect size of correlation. *P* values are 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. WBC: white blood cell, Hb: hemoglobin, MCV: mean corpuscular volume, MCHC: mean corpuscular hemoglobin concentration, Plt: Platelet. c, Distributions of hemoglobin in subjects with different number of alterations. d, Distributions 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 boxplots. *P* values are calculated by two-sided t test based on multivariate linear regression models including age and gender as covariates in (b, d), and by two-sided Wilcoxson rank sum test in (c), and not adjusted for multiple comparison. 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×the interquartile range (IQR) and Q3 + 1.5×IQR. Number of subjects in each category is shown under boxplots.

Fig. 4 | Impact of CH on mortality from hematological malignancies.

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=4947) are shown. b, Cumulative mortality from myeloid and lymphoid malignancies in subjects with or without CH are shown. 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 cooccurring alterations and maximum clone sizes (<10% or ≥10%). Cell fractions of unclassifiable CNAs were regarded to be smaller than 10%. e, Cumulative mortality from HM in subjects with CH and abnormalities in complete blood counts (CBC) (n=550), with CH alone (n=2,065), with abnormalities in CBC alone (n=703), or without either of them (n=3,094). f, Solid lines indicate cumulative mortality from HM in subjects with both SNV/indels and CNA (n=514), SNV/indels alone (n=1,723), CNAs alone (n=1,099), and without any alterations (n=4,947). Colored bands indicate 95% confidence intervals. In (a-c,f), *P* values were calculated by two-sided Wald test based on multivariate regression models. In (e), *P* values are calculated by two-sided log-rank test stratified by age (≤70 or >70 years old) and gender because of non-proportional hazards. *P* values are not adjusted for multiple comparison throughout the figure.

Fig. 5 | Impact of CH-related alterations on mortality from HM.

a-d, Hazard ratios for mortality from All hematological malignancies (All HM), myeloid neoplasms, and lymphoid neoplasms are indicated by green, red, and blue dots, respectively. Error bars indicate 95% confidence intervals. In (a), hazard ratios of the indicated covariates are 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). In (b-d), hazard ratios of the indicated alterations are calculated within the case-cohort design (Extended Data Fig.1b, n=8,283) in comparison with CH-negative cases. Hazard ratios are not shown for alterations without any event. Cell fractions of unclassifiable CNAs are regarded to be zero in (a), and smaller than 5% in (b). n, number of cases with the indicated alterations; N.A., not applicable; #Alteration, additional one alteration; Clone size +10%, 10% increase in cell fraction; SNV+CNA, cooccurrence of both SNVs/indels and

CNAs; #SNV, number of SNVs/indels; CF, cell fraction of CNAs; #CNA, number of CNAs.

Fig. 6 | Effect of SNV/indels and CNAs on cardiovascular mortality.

a, Cardiovascular mortality in subjects with SNV/indels (VAF ≥5% or <5%), and those without SNV/indels. Hazard ratios and *P* values are calculated in comparison with those without SNV/indels by two-sided Wald test. b, Results of multivariate logistic regressions for the presence of hypertension within 4,660 subjects with available information for covariates. Explanatory variables were selected by stepwise method from following factors: presence of SNV/indels, CNAs, age (+10 years), gender, BMI (+5), history of drinking and smoking, presence of diabetes mellitus, hyperlipidemia, hypertension, and 13 blood test values available in ≥70% of the subjects. Only remaining variables are shown. c, Cardiovascular mortality in subjects with CNAs (cell fraction ≥5% or <5%), and those without CNAs. Hazard ratios are calculated in comparison with those without CNAs. d, Cardiovascular mortality in subjects with both SNV/indels (VAF≥5%) and CNAs (purple), with SNV/indels (VAF≥5%) alone (red), with CNAs alone (blue), and without SNV/indels (VAF≥5%) or CNAs (gray). Hazard ratios are calculated by comparing those with both SNV/indels (VAF≥5%) and CNAs with those with SNV/indels (VAF≥5%) alone, or CNAs alone. In (a), (c) and (d), all comparisons were performed with multivariate models including age, gender, body-mass index, comorbidities (diabetes mellitus, hypertension, and dyslipidemia), history of smoking/drinking, and the versions of SNP array within 6,697 subjects with available clinical information. Throughout the figure, error bars indicate the 95% confidence intervals, and *P* values are calculated by two-sided Wald test and not adjusted for multiple comparison.

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., Asimakopoulos, 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* (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 e374 (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* (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. Thoennissen, 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. Mutation in TET2 in Myeloid Cancers. (2009).

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**, i115-120 (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).

Fig. 1

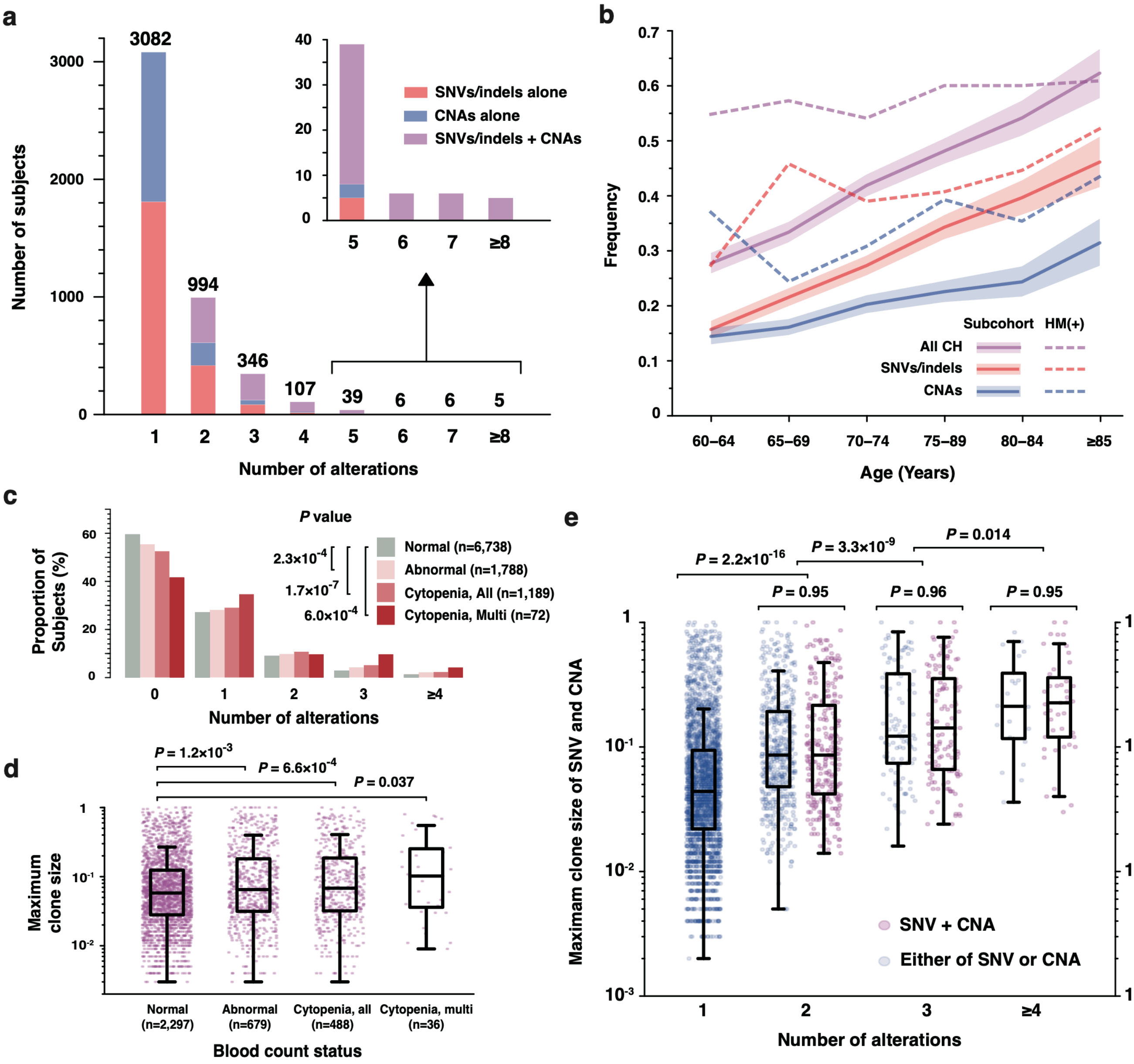


Fig. 1 | Landscape of SNVs/indels and CNAs in clonal hematopoiesis.

a, Distribution of the number of genetic alterations in each subject. Subjects with SNV/indels alone, with CNAs alone, or with both of them 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 subjects with and without HM events, respectively. Colored bands represent the 95% confidence intervals. c, Number of cooccurring alterations in those with subjects with abnormalities in blood cell counts, or cytopenia. d, Maximum cell fraction of CH-related alterations in CH-positive 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 cooccurring alterations. Cell fractions of SNVs/indels are defined as 2 times VAF. Those with both of SNVs/indels and CNAs are shown in purple, while those with either are shown in blue. In panel (d,e), unclassifiable CNAs were excluded because we cannot calculate their precise cell fractions. The box plots indicate the median, first and third quartiles (Q1 and Q3) and whiskers extend to the furthest value between Q1 – 1.5×the interquartile range (IQR) and Q3 + 1.5×IQR. In (c-e), *P* values were calculated by two-sided Wilcoxon rank-sum test and not adjusted for multiple comparison.

Fig. 2

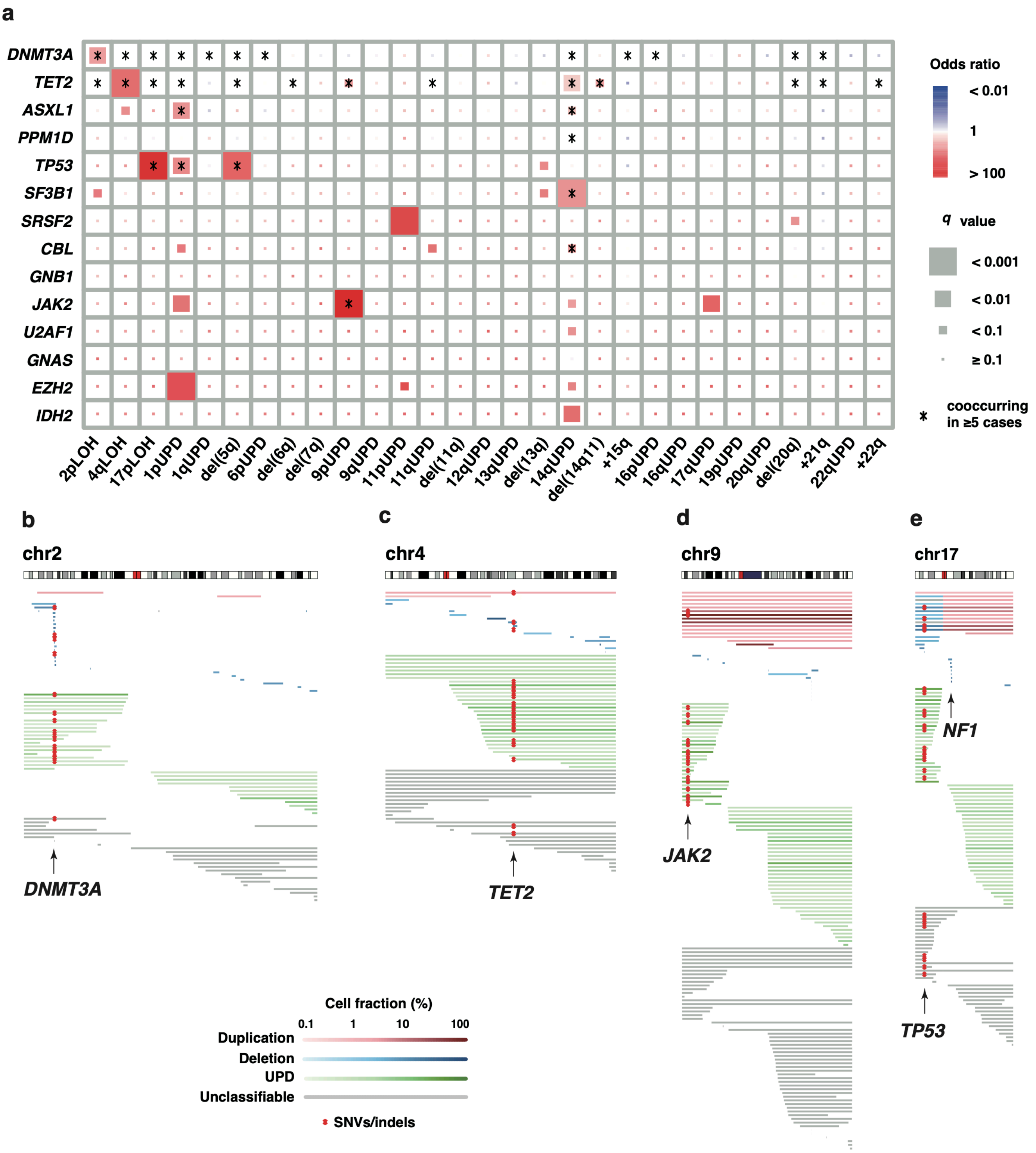


Fig. 2 | Cooccurrences of SNVs/indels and CNAs in clonal hematopoiesis.

a, The correlations between individual SNVs/indels and CNAs. The size of rectangles indicates the significance of correlations. Red rectangles represent positive correlations while blue rectangles represent negative correlations. Combinations of SNVs/indels and CNAs seen in 5 or more subjects are indicated by asterisks. b-e, The distributions of CNAs on chromosome 2 (b), 4 (c), 9 (d), and 17 (e). Horizontal bars represent CNAs, and cooccurring 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 which cannot be classified into any of UPD, deletion, or duplication are indicated as unclassifiable CNAs (gray).

Fig. 3

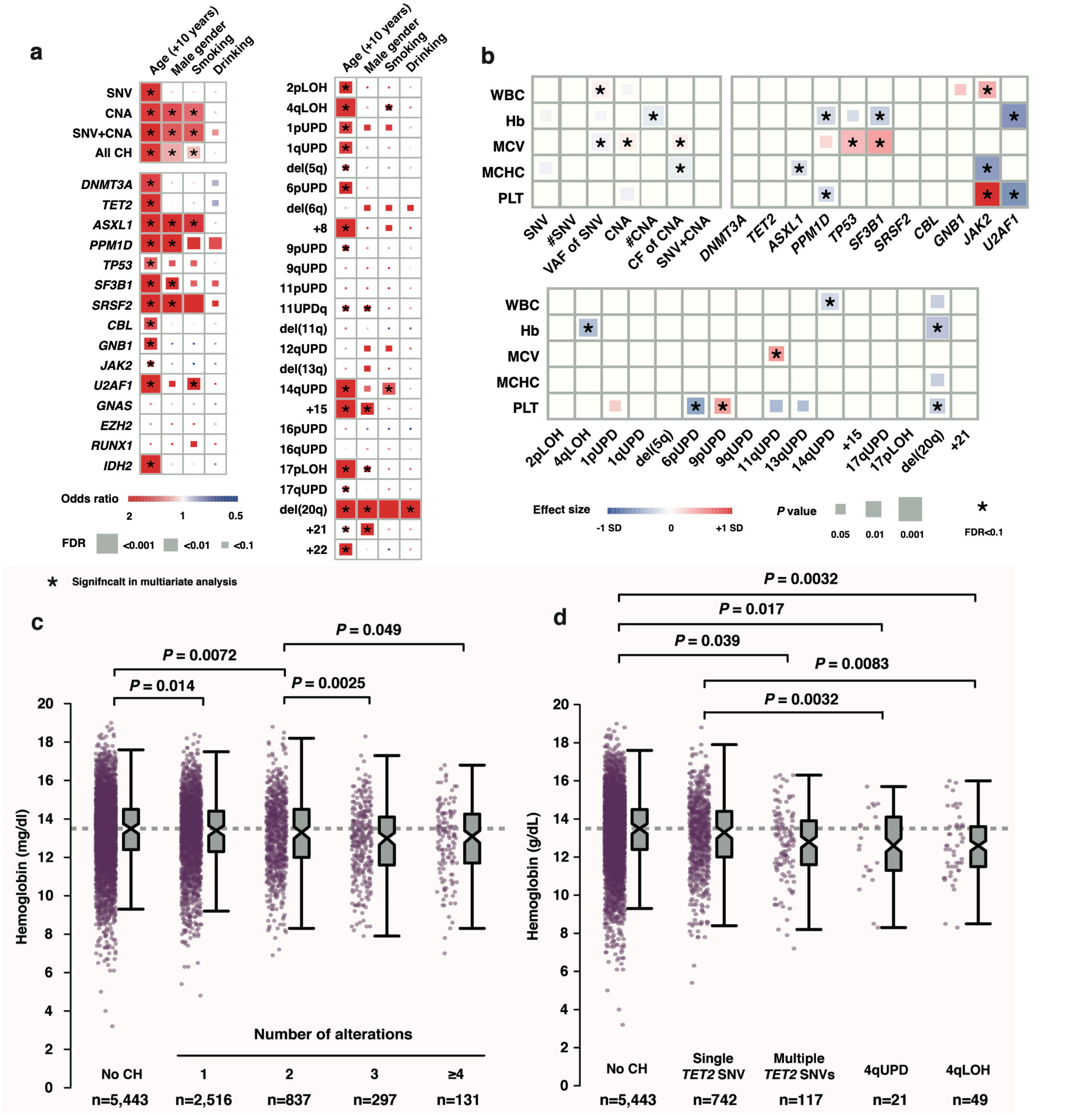


Fig. 3 | Risk factors for CH and effects on blood counts.

a, Correlations of genetic alterations with age, male gender, history of smoking and drinking. Sizes and colors of rectangles represent the significance and effect size calculated by two-sided Wald test. Asterisks indicate the clinical factors significantly correlated with each alteration in multivariate logistic regression (P<0.05). b, Correlations between genetic alterations and blood counts. The sizes and colors of rectangles indicate the significance, and effect size of correlation. P values are 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. WBC: white blood cell, Hb: hemoglobin, MCV: mean corpuscular volume, MCHC: mean corpuscular hemoglobin concentration, Plt: Platelet. c, Distributions of hemoglobin in subjects with different number of alterations. d, Distributions 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 boxplots. *P* values are calculated by two-sided t test based on multivariate linear regression models including age and gender as covariates in (b, d), and by two-sided Wilcoxson rank sum test in (c), and not adjusted for multiple comparison. 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×the interquartile range (IQR) and Q3 + 1.5×IQR. Number of subjects in each category is shown under boxplots.

Fig. 4

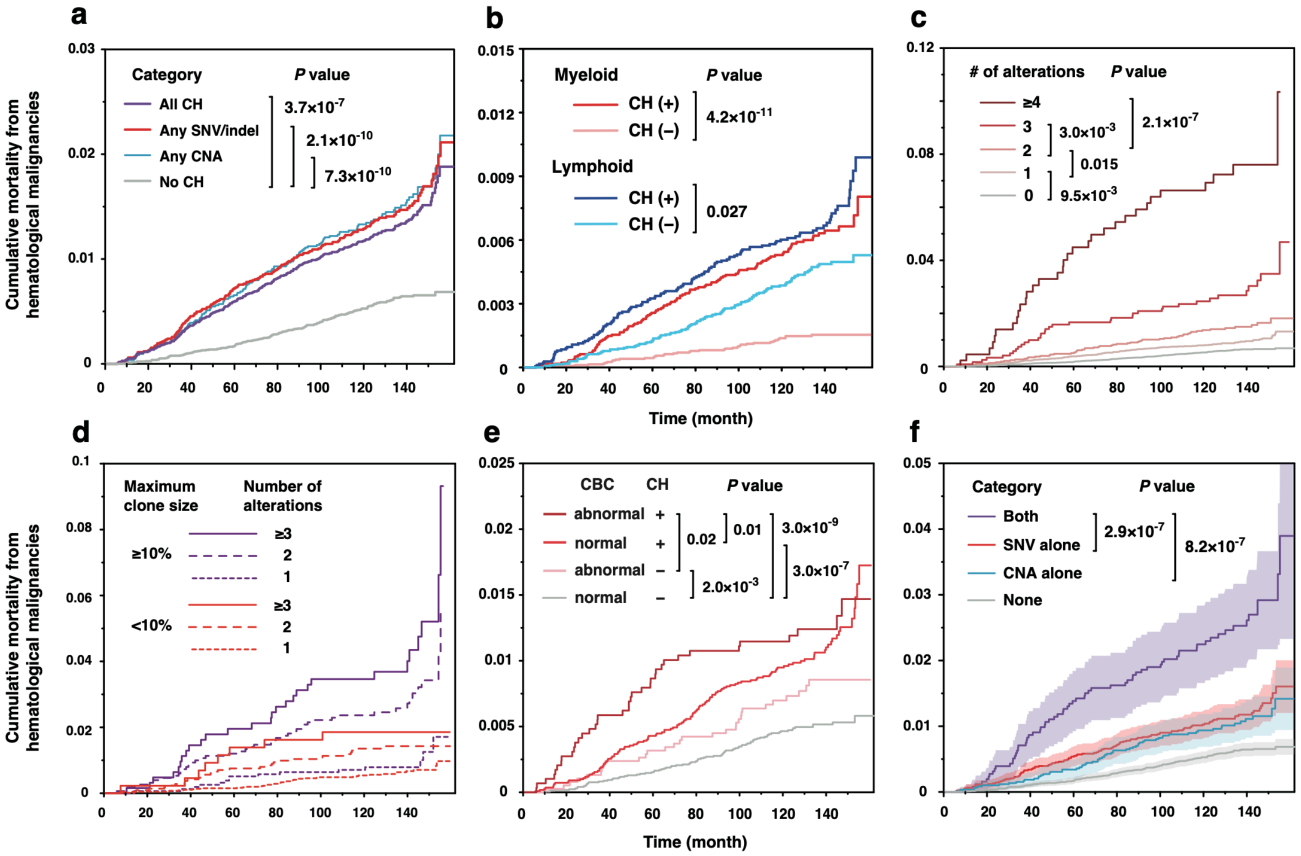


Fig. 4 I Impact of CH on mortality from hematological malignancies.

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=4947) are shown. b, Cumulative mortality from myeloid and lymphoid malignancies in subjects with or without CH are shown. 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 cooccurring alterations and maximum clone sizes (<10% or ≥10%). Cell fractions of unclassifiable CNAs were regarded to be smaller than 10%. e, Cumulative mortality from HM in subjects with CH and abnormalities in complete blood counts (CBC) (n=550), with CH alone (n=2,065), with abnormalities in CBC alone (n=703), or without either of them (n=3,094). f, Solid lines indicate cumulative mortality from HM in subjects with both SNV/indels and CNA (n=514), SNV/indels alone (n=1,723), CNAs alone (n=1,099), and without any alterations (n=4,947). Colored bands indicate 95% confidence intervals. In (a-c,f), *P* values were calculated by two-sided Wald test based on multivariate regression models. In (e), *P* values are calculated by two-sided log-rank test stratified by age (≤70 or >70 years old) and gender because of non-proportional hazards. *P* values are not adjusted for multiple comparison throughout the figure.

Fig. 5

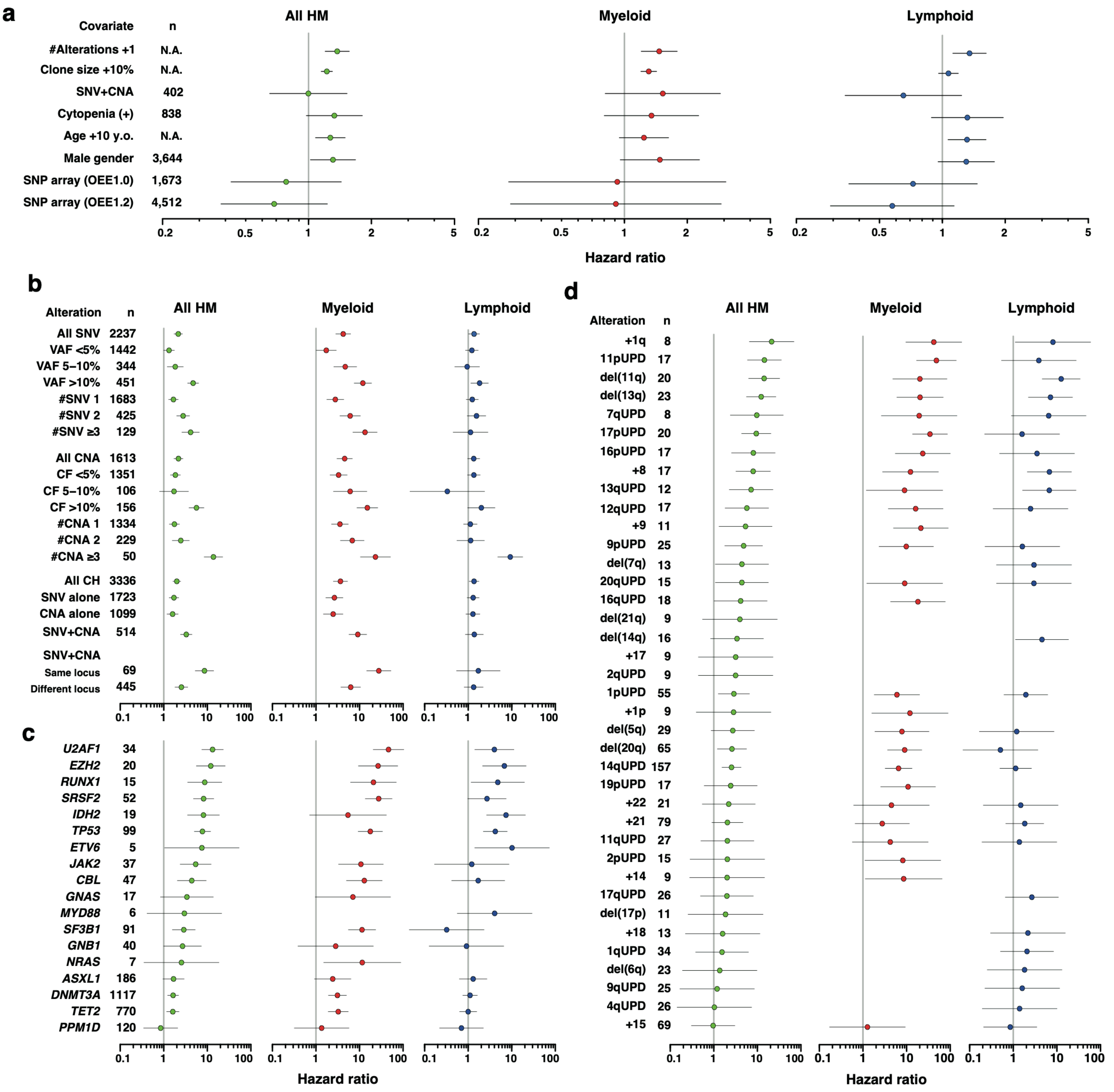


Fig. 5 | Impact of CH-related alterations on mortality from HM.

a-d, Hazard ratios for mortality from All hematological malignancies (All HM), myeloid neoplasms, and lymphoid neoplasms are indicated by green, red, and blue dots, respectively. Error bars indicate 95% confidence intervals. In (a), hazard ratios of the indicated covariates are 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). In (b-d), hazard ratios of the indicated alterations are calculated within the case-cohort design (Extended Data Fig.1b, n=8,283) in comparison with CH-negative cases. Hazard ratios are not shown for alterations without any event. Cell fractions of unclassifiable CNAs are regarded to be zero in (a), and smaller than 5% in (b). n, number of cases with the indicated alterations; N.A., not applicable; #Alteration, additional one alteration; Clone size +10%, 10% increase in cell fraction; SNV+CNA, cooccurrence of both SNVs/indels and CNAs; #SNV, number of SNVs/indels; CF, cell fraction of CNAs; #CNA, number of CNAs.

Fig. 6

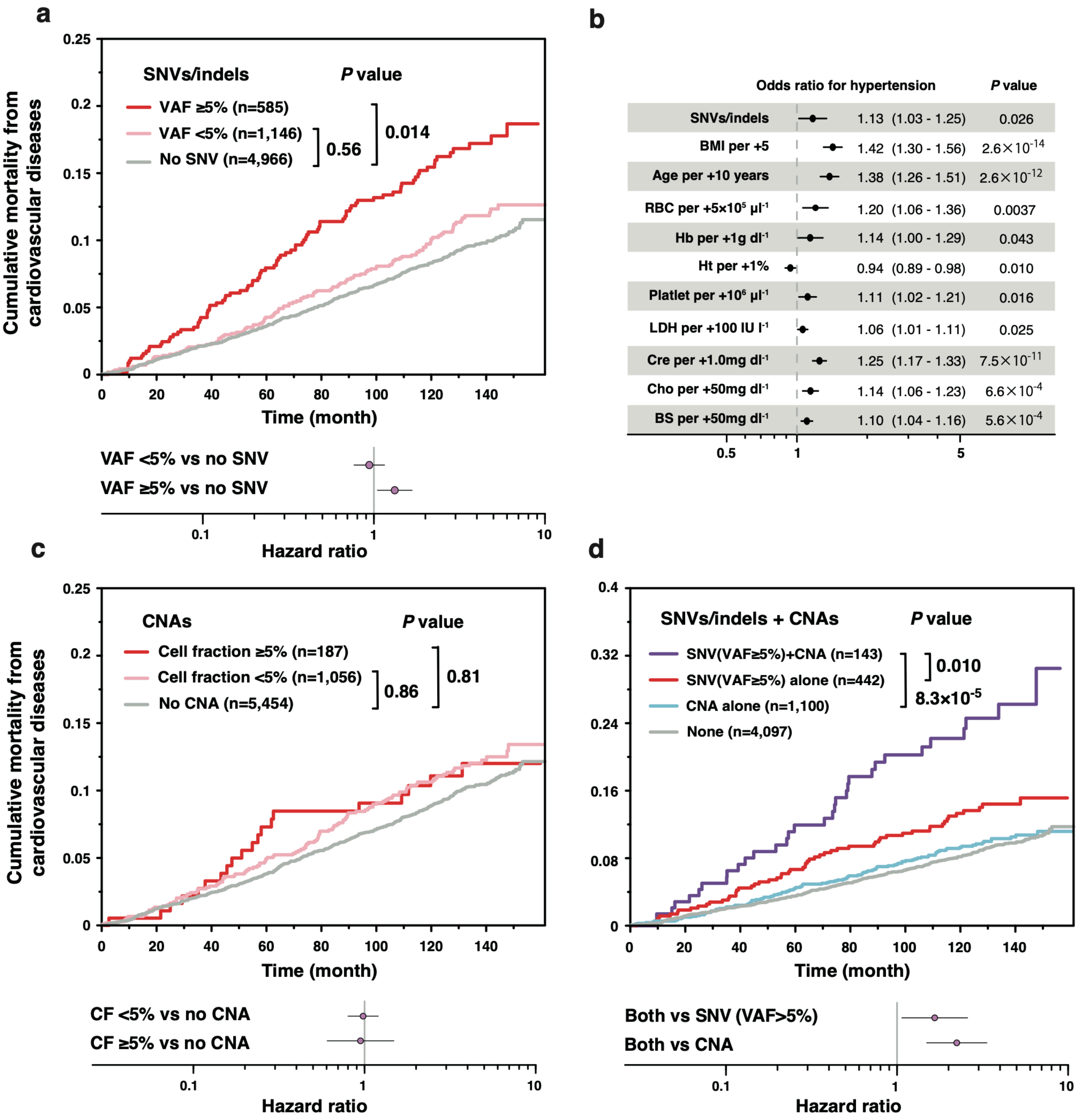
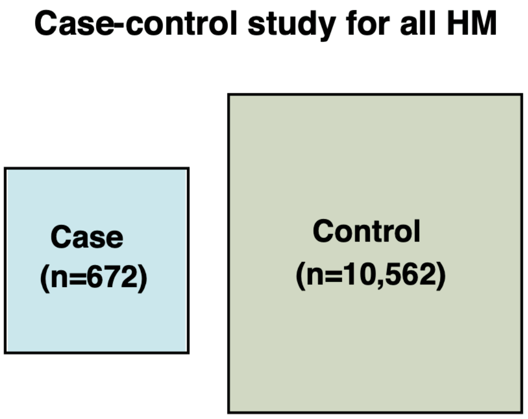


Fig. 6 l Effect of SNV/indels and CNAs on cardiovascular mortality.

a, Cardiovascular mortality in subjects with SNV/indels (VAF ≥5% or <5%), and those without SNV/indels. Hazard ratios and *P* values are calculated in comparison with those without SNV/indels by two-sided Wald test. b, Results of multivariate logistic regressions for the presence of hypertension within 4,660 subjects with available information for covariates. Explanatory variables were selected by stepwise method from following factors: presence of SNV/indels, CNAs, age (+10 years), gender, BMI (+5), history of drinking and smoking, presence of diabetes mellitus, hyperlipidemia, hypertension, and 13 blood test values available in ≥70% of the subjects. Only remaining variables are shown. c, Cardiovascular mortality in subjects with CNAs (cell fraction ≥5% or <5%), and those without CNAs. Hazard ratios are calculated in comparison with those without CNAs. d, Cardiovascular mortality in subjects with both SNV/indels (VAF≥5%) and CNAs (purple), with SNV/indels (VAF≥5%) alone (red), with CNAs alone (blue), and without SNV/indels (VAF≥5%) or CNAs (gray). Hazard ratios are calculated by comparing those with both SNV/indels (VAF≥5%) and CNAs with those with SNV/indels (VAF≥5%) alone, or CNAs alone. In (a), (c) and (d), all comparisons were performed with multivariate models including age, gender, body-mass index, comorbidities (diabetes mellitus, hypertension, and dyslipidemia), history of smoking/drinking, and the versions of SNP array within 6,697 subjects with available clinical information. Throughout the figure, error bars indicate the 95% confidence intervals, and *P* values are calculated by two-sided Wald test and not adjusted for multiple comparison.

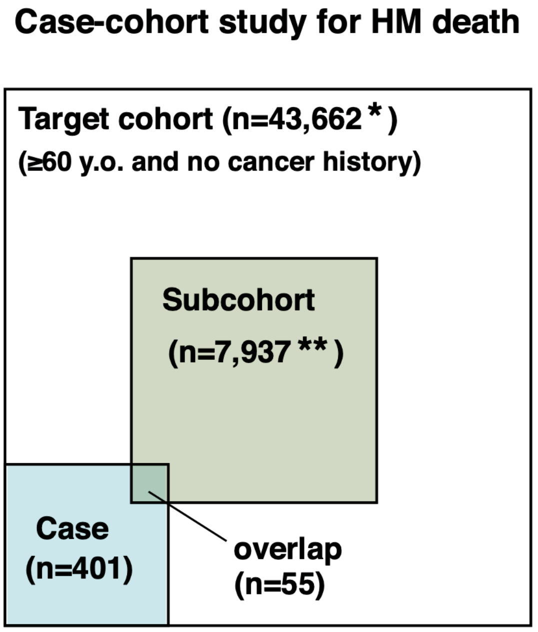
Extended Data Fig. 1

a



	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
Linage 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	6649	11,234

b


Case-cohort study for HM death
Target cohort (n=43,662 *)
(≥60 y.o. and no cancer history)
Subcohort
(n=7,937 **)
Case
(n=401)
overlap
(n=55)

* 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(+)		
Hematogical 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
Linage 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

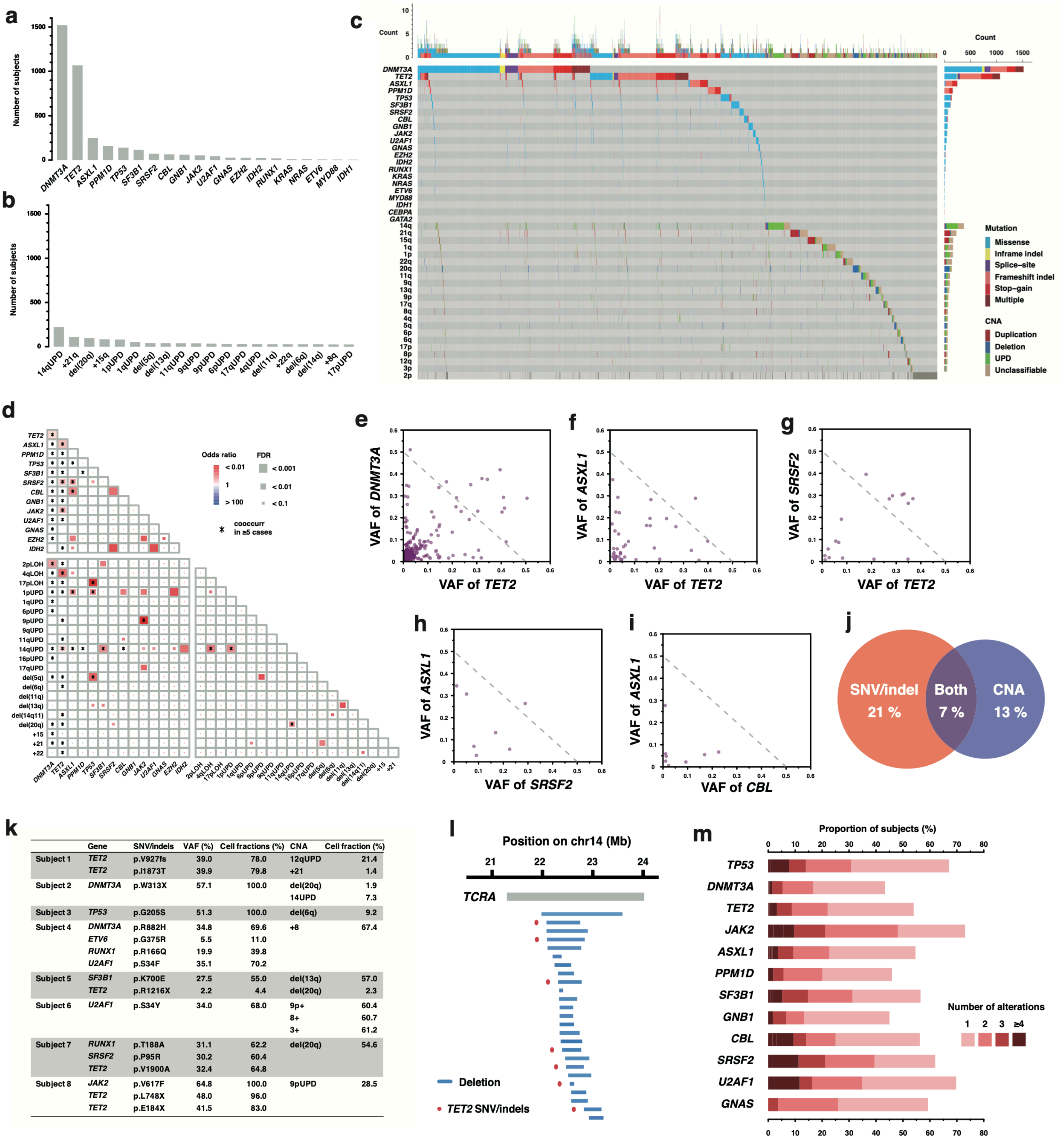
Case (Death from HM)

	CH(+)				CH(−)	Total
	SNV alone	CNA alone	Both	All CH(+)		
Hematogical 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
Linage 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. 2

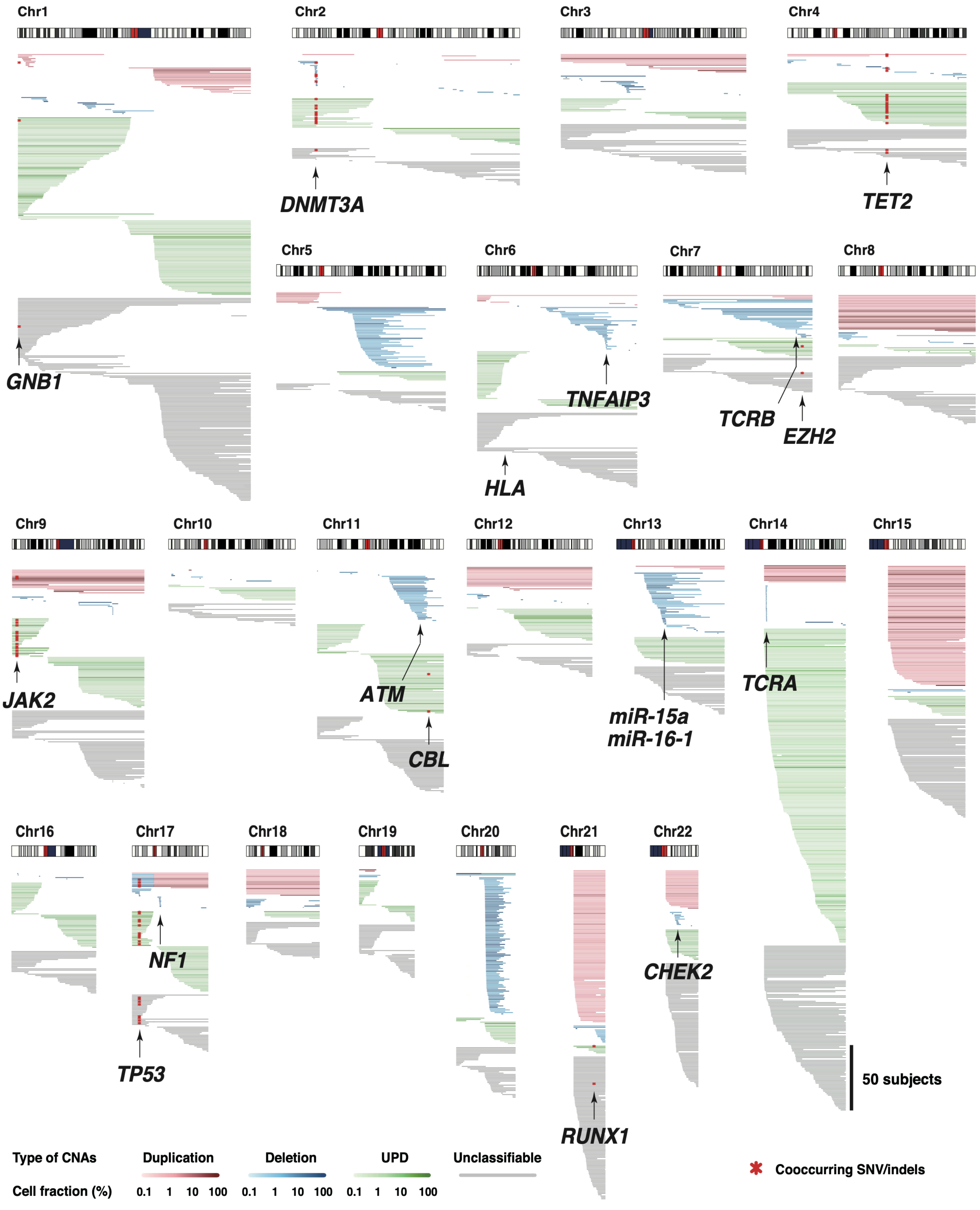


	Gene	SNV/indels	VAF (%)	Cell fractions (%)	CNA	Cell fraction (%)
Subject 1	*TET2*	p.V927fs	39.0	78.0	12qUPD	21.4
	TET2	p.I1873T	39.9	79.8	+21	1.4
Subject 2	*DNMT3A*	p.W313X	57.1	100.0	del(20q)	1.9
					14UPD	7.3
Subject 3	*TP53*	p.G205S	51.3	100.0	del(6q)	9.2
Subject 4	*DNMT3A*	p.R882H	34.8	69.6	+8	67.4
	ETV6	p.G375R	5.5	11.0		
	RUNX1	p.R166Q	19.9	39.8		
	U2AF1	p.S34F	35.1	70.2		
Subject 5	*SF3B1*	p.K700E	27.5	55.0	del(13q)	57.0
	TET2	p.R1216X	2.2	4.4	del(20q)	2.3
Subject 6	*U2AF1*	p.S34Y	34.0	68.0	9p+	60.4
					8+	60.7
					3+	61.2
Subject 7	*RUNX1*	p.T188A	31.1	62.2	del(20q)	54.6
	SRSF2	p.P95R	30.2	60.4		
	TET2	p.V1900A	32.4	64.8		
Subject 8	*JAK2*	p.V617F	64.8	100.0	9pUPD	28.5
	TET2	p.L748X	48.0	96.0		
	TET2	p.E184X	41.5	83.0		

Extended Data Fig. 2 I Landscape of genetic alterations in CH.

a-b, The number of subjects with individual SNVs/indels (a) and CNAs (b). The vertical axis represents the number of subjects with indicated alterations. Unclassifiable CNAs are not included in (b). c, Landscape of SNVs/indels and CNAs in 11,234 subjects. Those without CH-related alterations are omitted. d, The correlations between individual genetic alterations. Combinations seen in 5 or more cases are indicated by asterisks. e-i, VAF of cooccurring SNV/indels in diagonal plot. Dots above the dashed line fulfill “pigeonhole principle” . j, Venn diagram illustrating the overlap between subjects with SNV/indels and those with CNAs. Frequencies within all subjects in whom SNVs/indels and CNAs were examined (n=11,234) are indicated. k, Subjects in whom coocurring SNVs/indels and CNAs were suspected to coexist in the same cells on the basis of “pigeonhole principle.” l, A magnified illustration of microdeletions around *TCRA* locus (14q11.2). A gray bar represents gene body of *TCRA*. Blue horizontal bars represent microdeletions. Cooccurring *TET2* SNVs are indicated by red dots. Genomic coordinates in hg19 are indicated above. m, Proportions of subjects with different number of cooccurring alterations within those who harbor SNVs/indels in the indicated genes. The proportions of subjects with 1, 2, 3, and ≥4 CNAs are depicted by different colors.

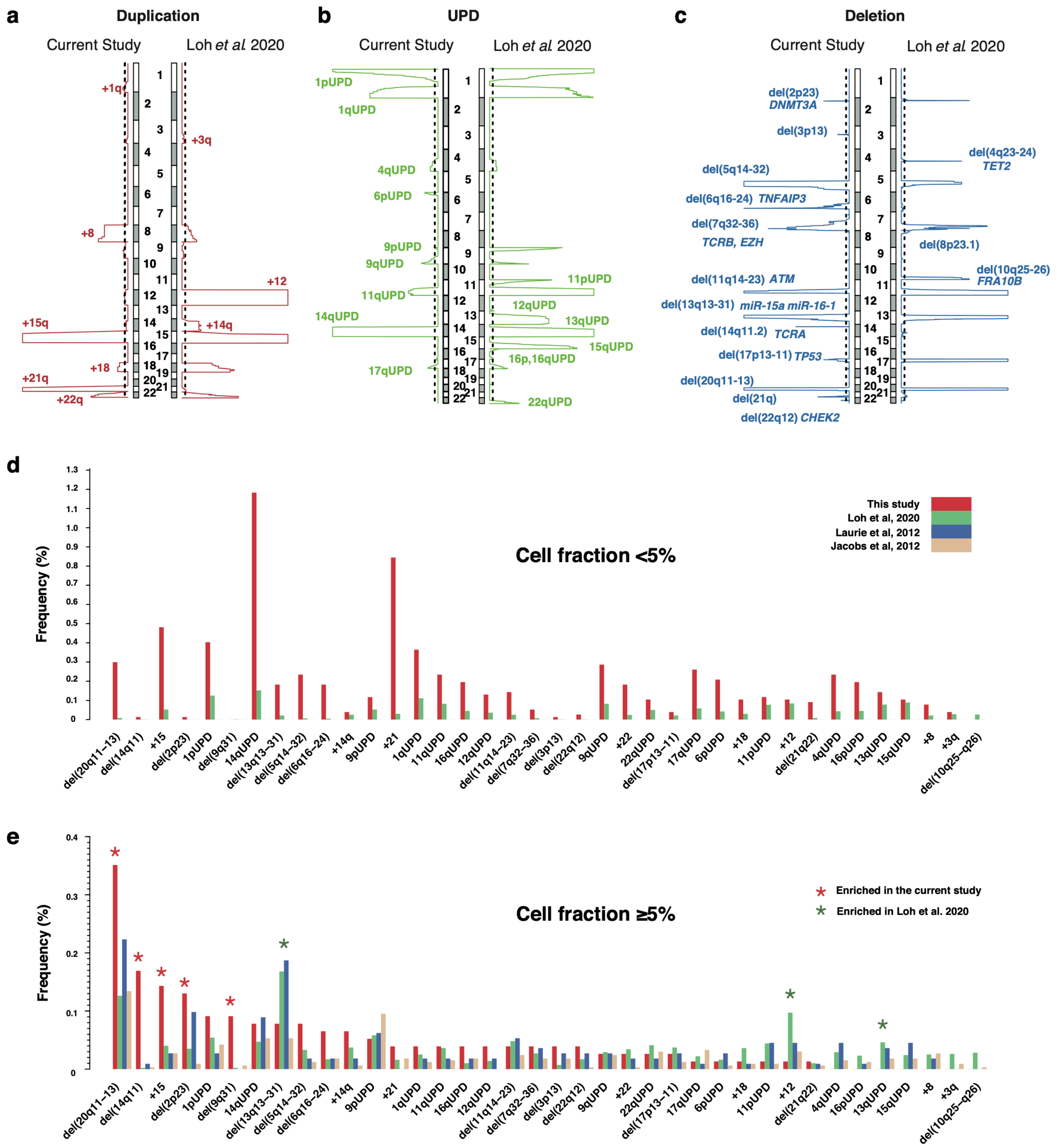
Extended Data Fig. 3



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 SNV/indels are indicated by red dots. Types of CNAs are depicted by different colors as indicated in the annotations.

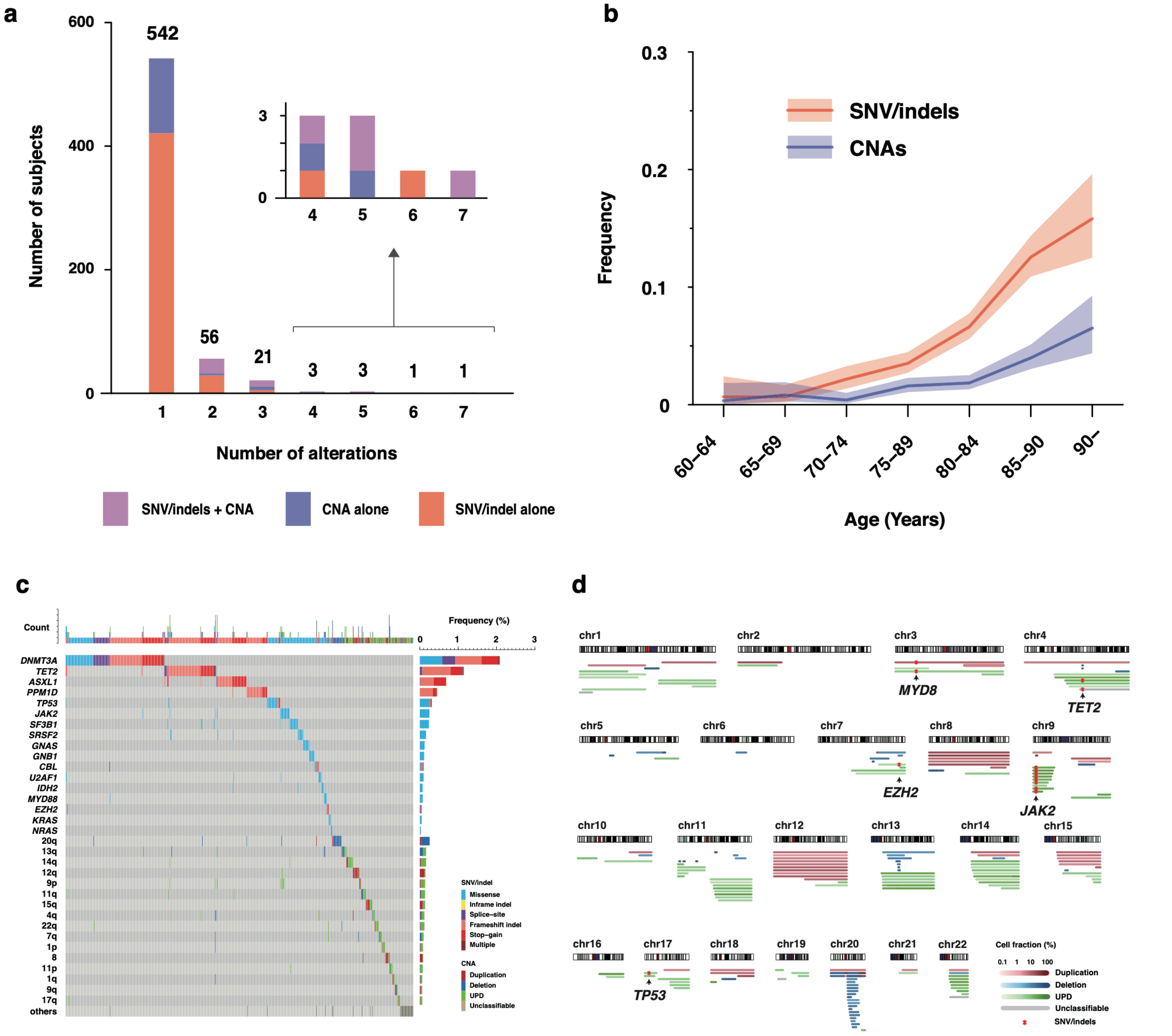
Extended Data Fig. 4



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 Japanese cohort (current study) and in British cohort[11]. Statistical significance for recurrence of CNAs were evaluated by PART[49]. 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).

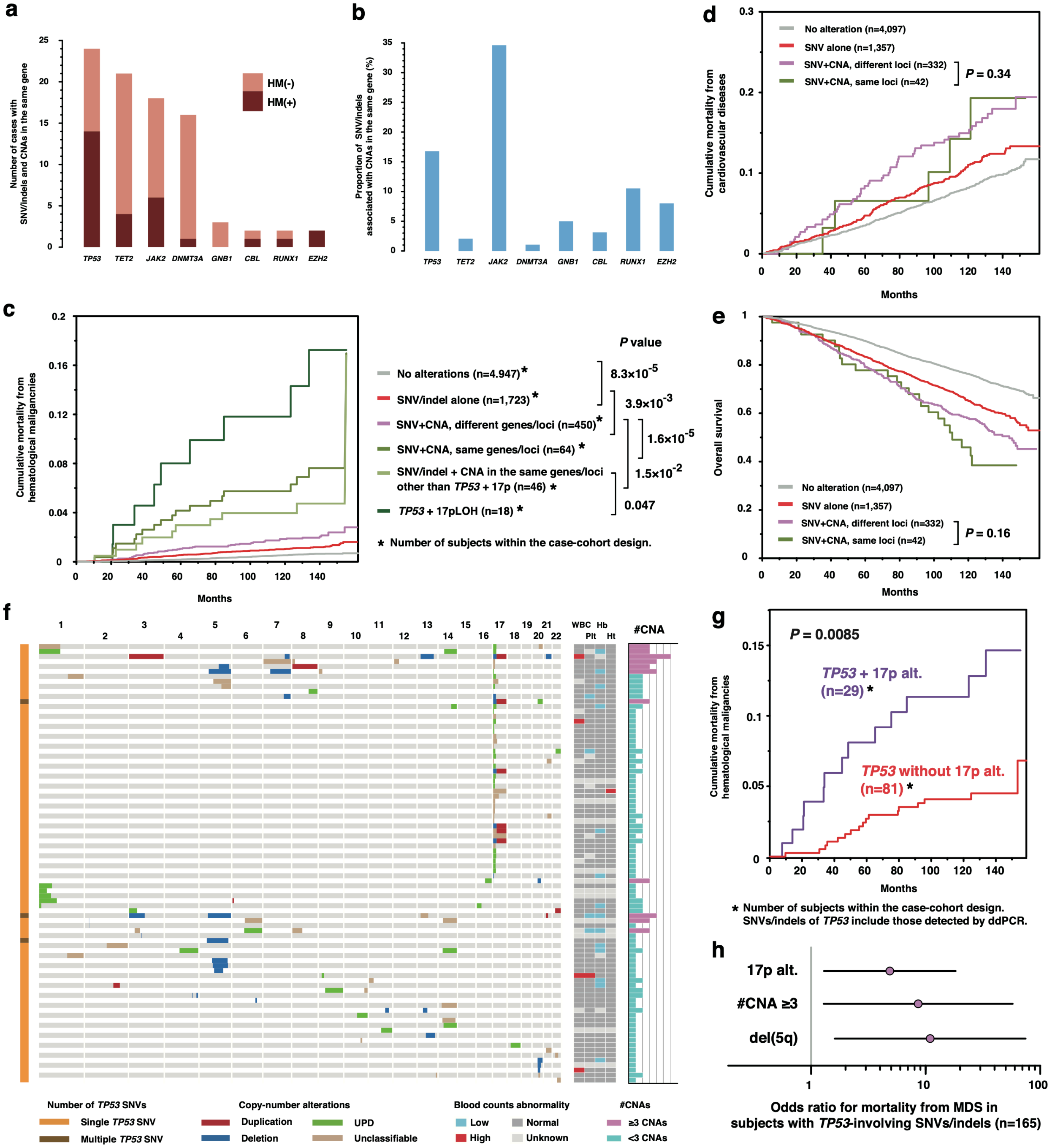
Extended Data Fig. 5



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.

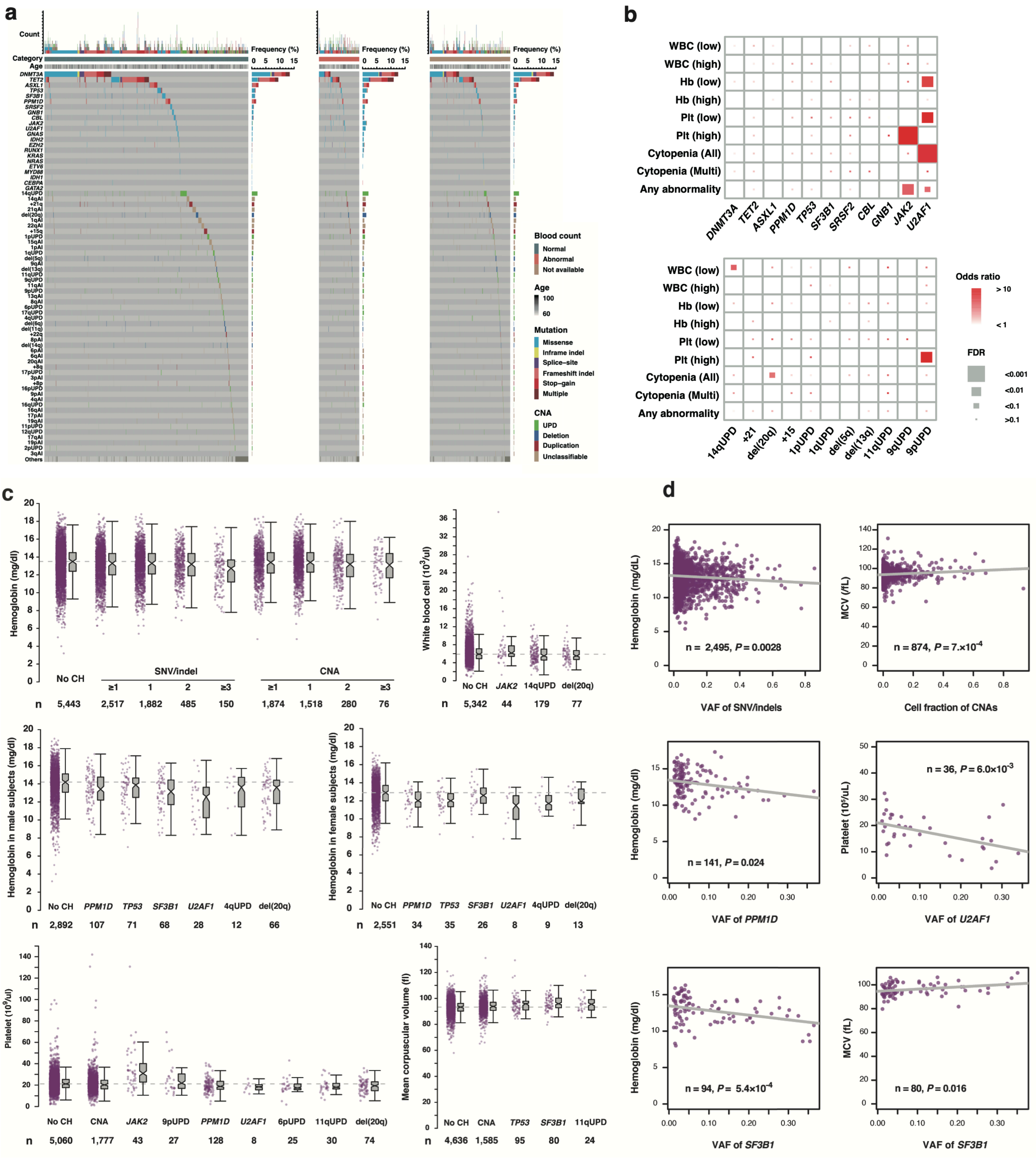
Extended Data Fig. 6



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 SNV/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).

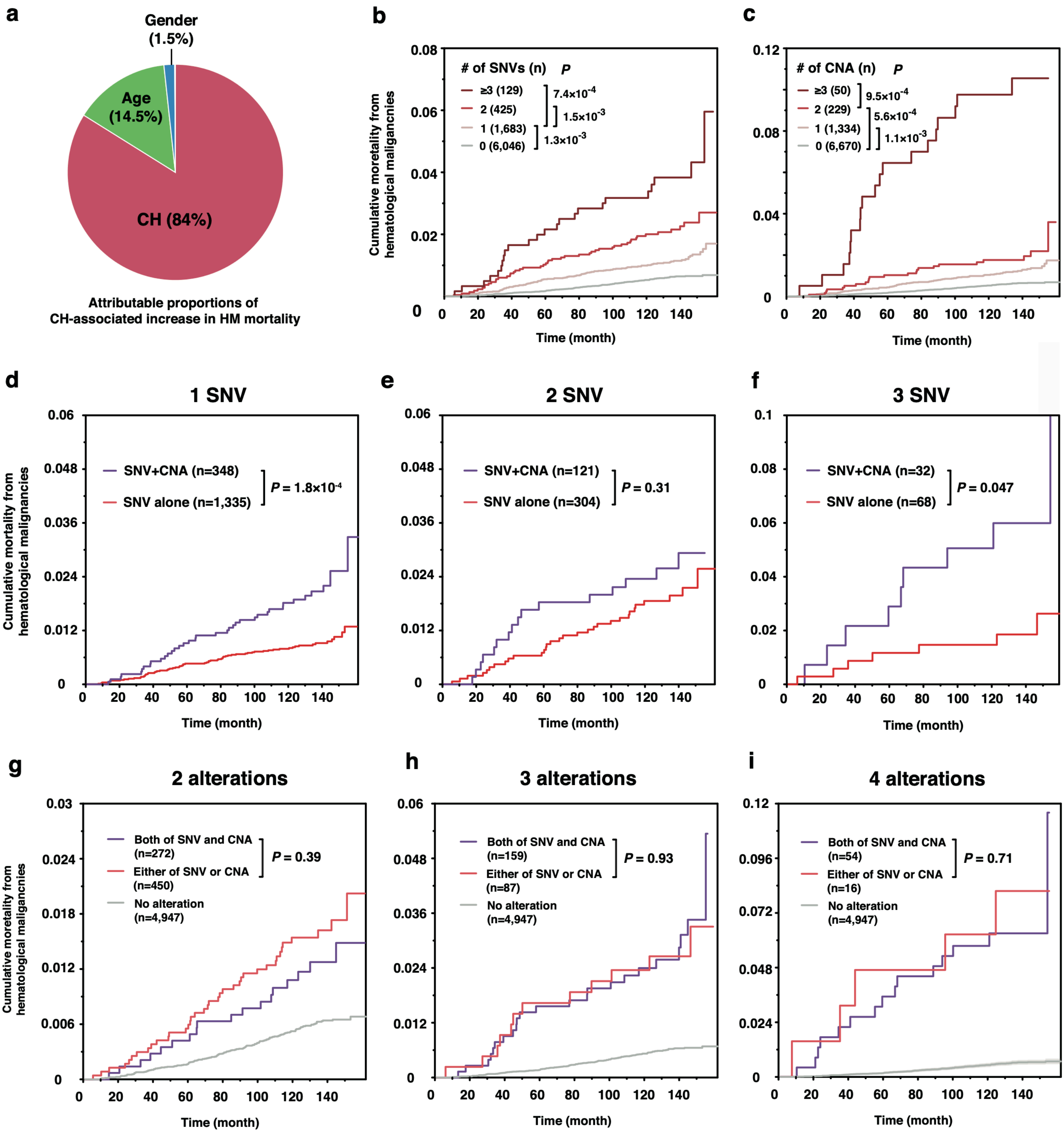
Extended Data Fig. 7



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×the interquartile range (IQR) and Q3 + 1.5×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.

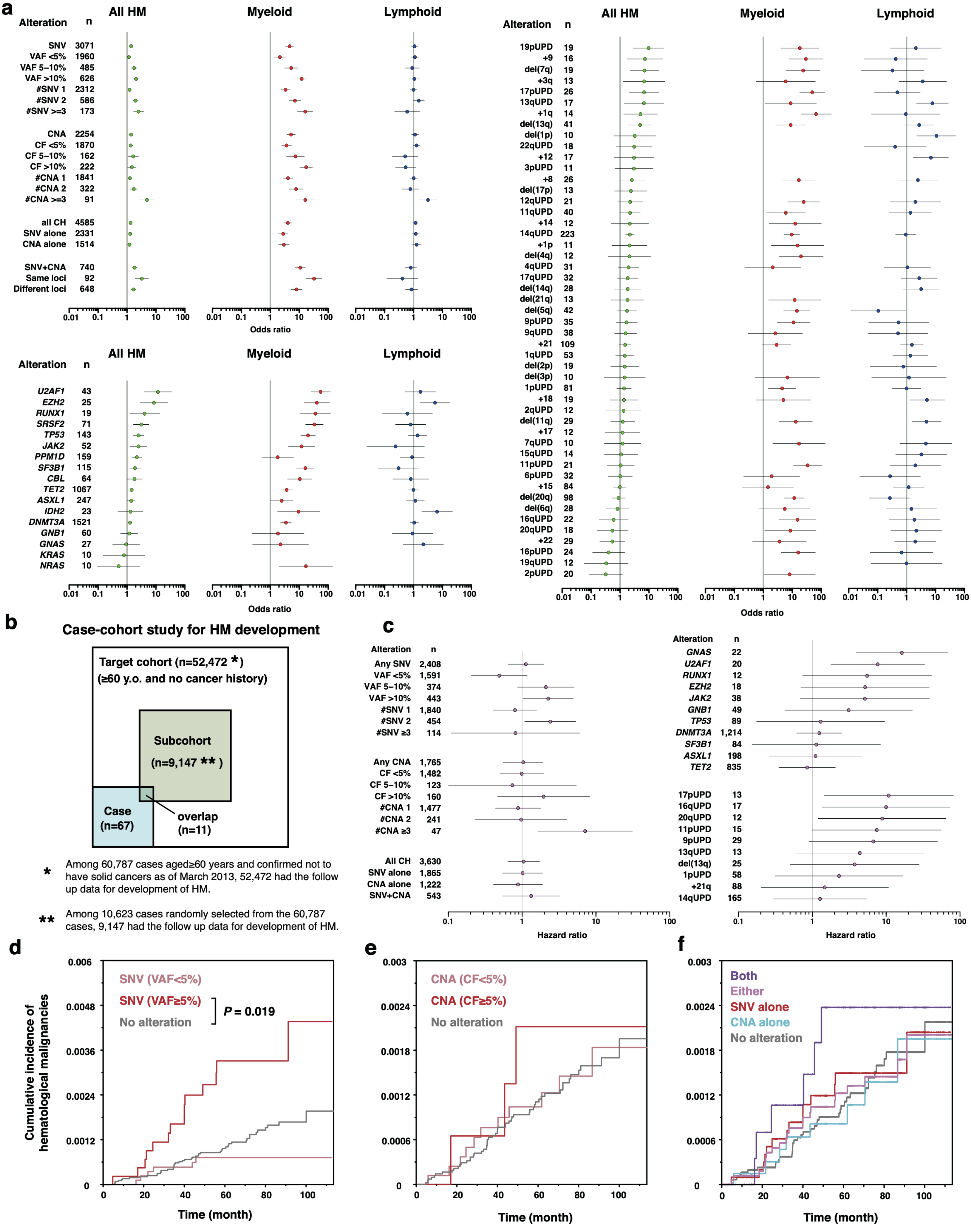
Extended Data Fig. 8



Extended Data Fig. 8 I 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 SNV/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.

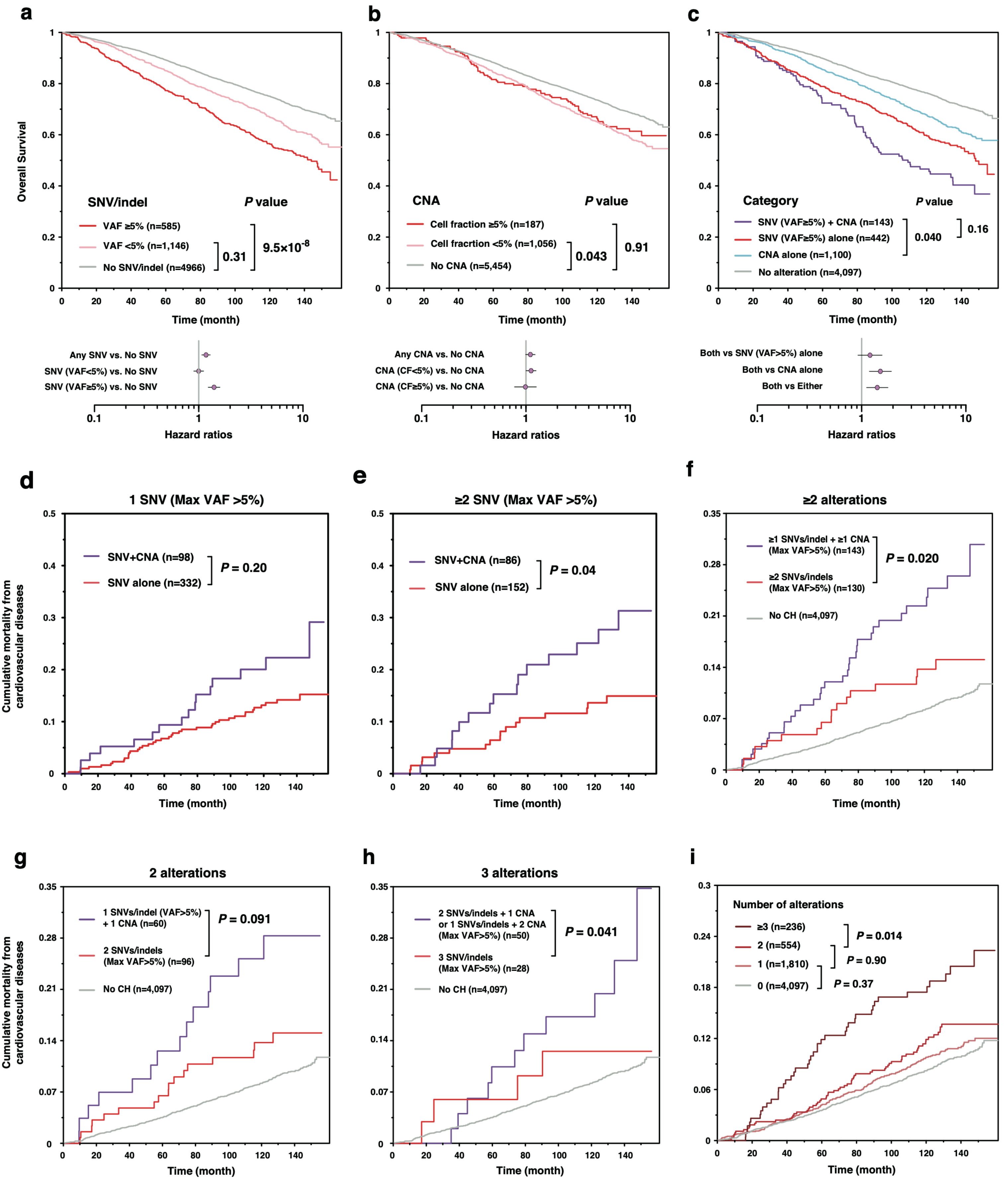
Extended Data Fig. 9



Extended Data Fig. 9 | Association of CH-related SNV/indel and CNA 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.

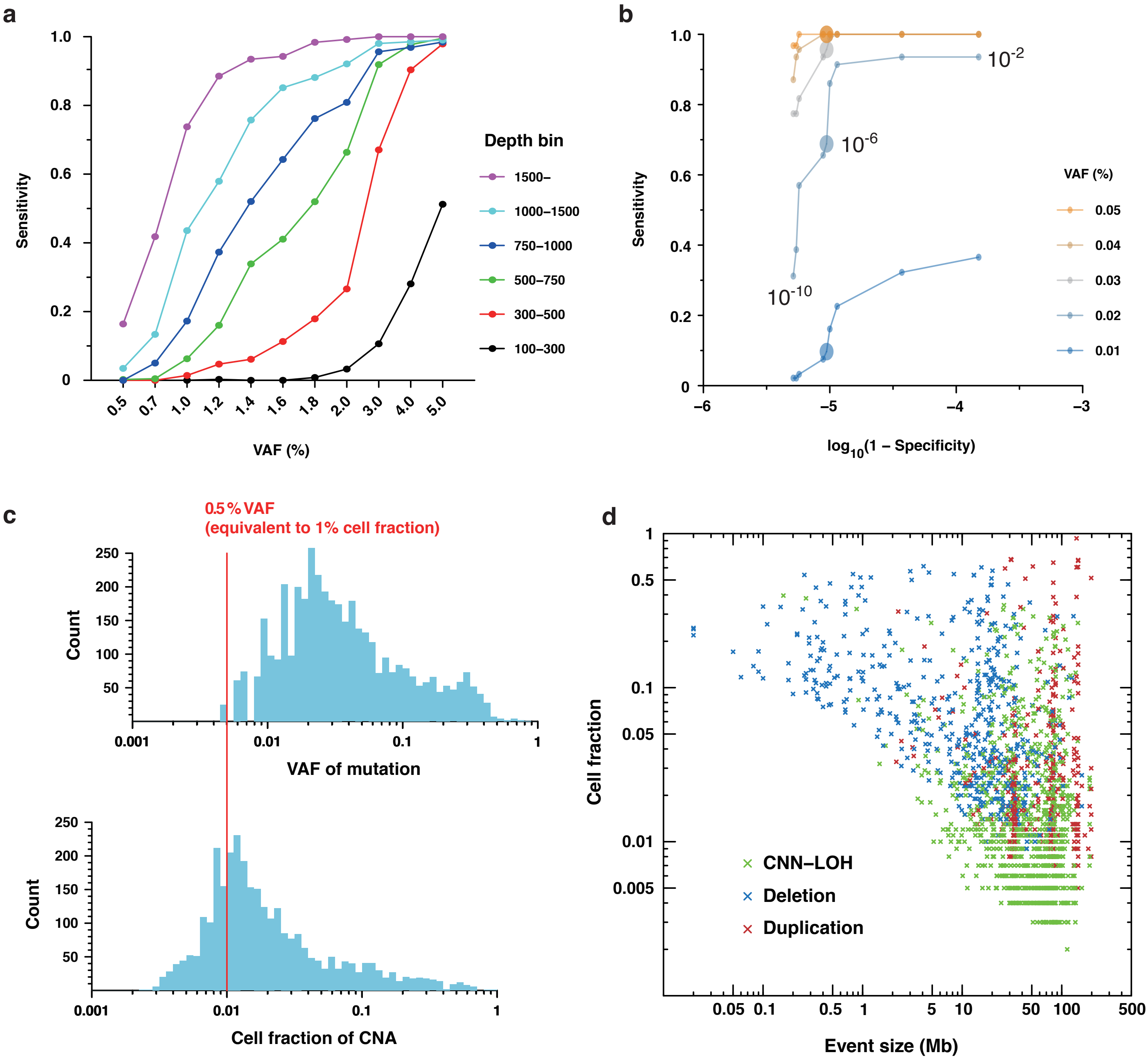
Extended Data Fig. 10



Extended Data Fig. 10 | Combined effect of SNV/indel and CNA on overall survival and cardiovascular mortality.

a-c, Effect of SNV/indels(a), CNAs(b), or combined SNV/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 SNV/indels (Max VAF>5%) and CNAs. Subject 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.

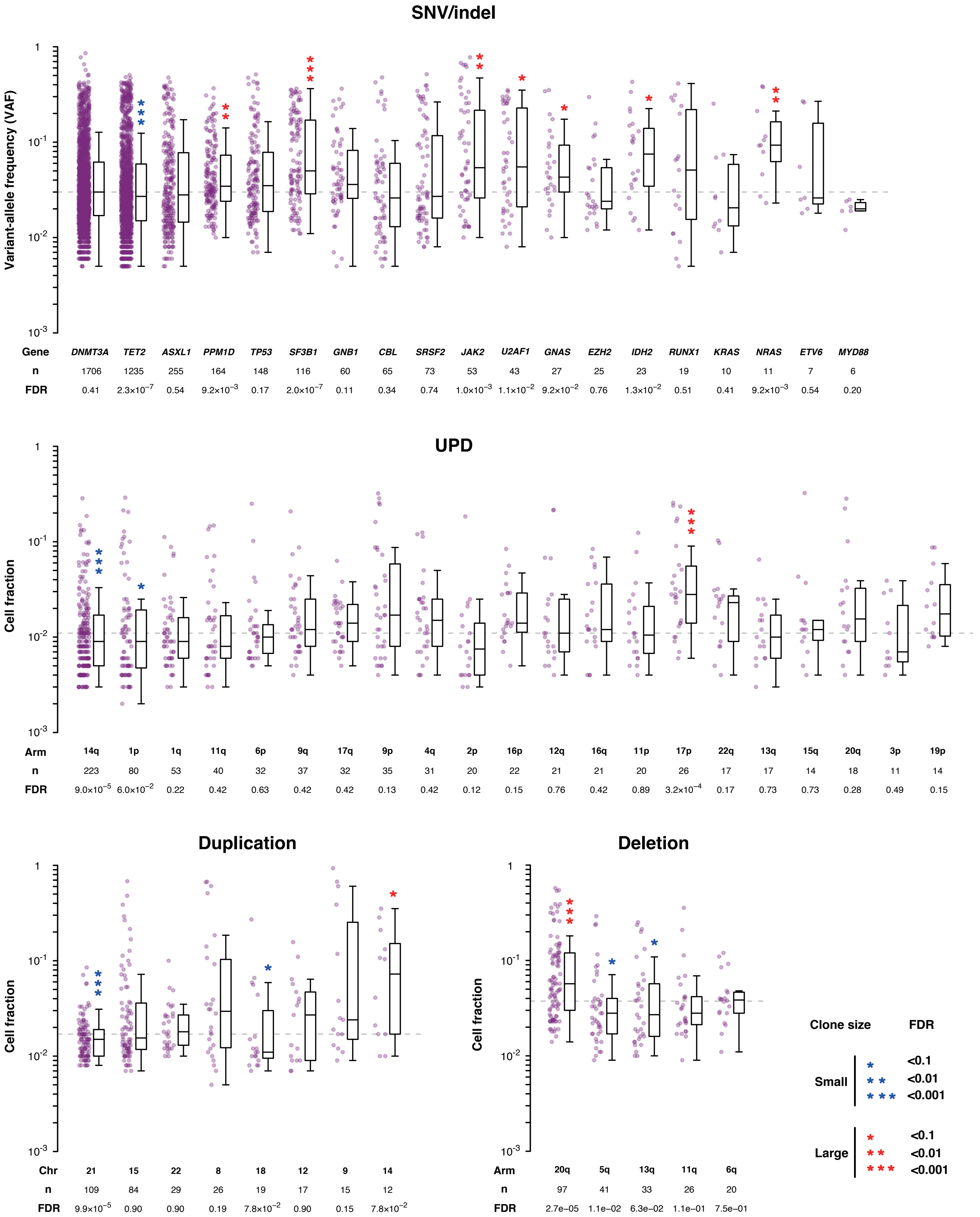
Supplementary Fig. 1



Supplementary Fig. 1 I Distributions of clone sizes and performance evaluation.

a, Sensitivities to detect SNVs simulated for different VAFs and sequencing depths. The horizontal axis represents target VAF of simulated SNVs. The vertical axis represents sensitivity, which was calculated as fractions of detected SNVs out of all simulated ones. b, Receiver operating characteristic (ROC) curves for detection of SNVs/indels, illustrating sensitivity on the vertical axis and $\log_{10}$(1 - specificity) on the horizontal axis. In this panel, we show sensitivity assuming sequencing depth is within x700-x900, which largely represents for the mean coverage in this study (x800). Dots represent variable cutoffs on beta-binominal P values (10^{-2} to 10^{-10}) (Online Method). Large dots represent a cutoff of 10^{-6}, which we adopted in the actual SNV call. c, Histograms of VAFs of SNVs/indels (top), and cell fractions of CNAs (bottom). The red vertical line indicates 0.5 % in VAFs, which is equivalent to 1% in cell fractions. It was impossible to precisely calculate cell fractions for unclassifiable CNAs. Instead, we calculated upper limits of cell fractions by assuming they were duplication. d, Distribution of detected CNAs with cell fractions on the vertical axis and event sizes on the horizontal axis. Unclassifiable CNAs are abbreviated from panel (d).

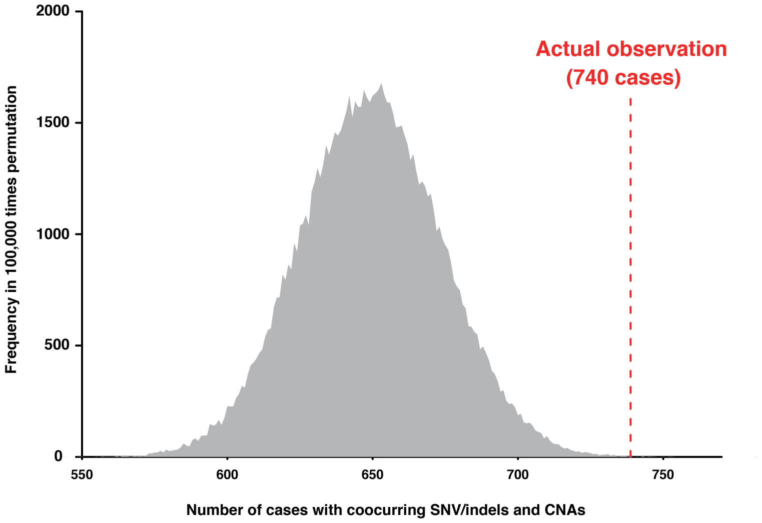
Supplmentary Fig. 2



Supplementary Fig. 2 I Clone size of individual CH-related alterations

VAF/clone size of individual CH-related alterations are compared within each category (SNVs/indels, UPDs, duplications, and deletions). FDR are calculated in comparison with VAF/clone size of all other alterations by two-sided wilcoxon rank sum test. Dashed horizontal lines indicate median VAF/clone size within each category. 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×the interquartile range (IQR) and Q3 + 1.5×IQR. n, number of subjects with the indicated alterations; Arm, names of affected chromosomal arms; Chr, names of affect chromosomes.

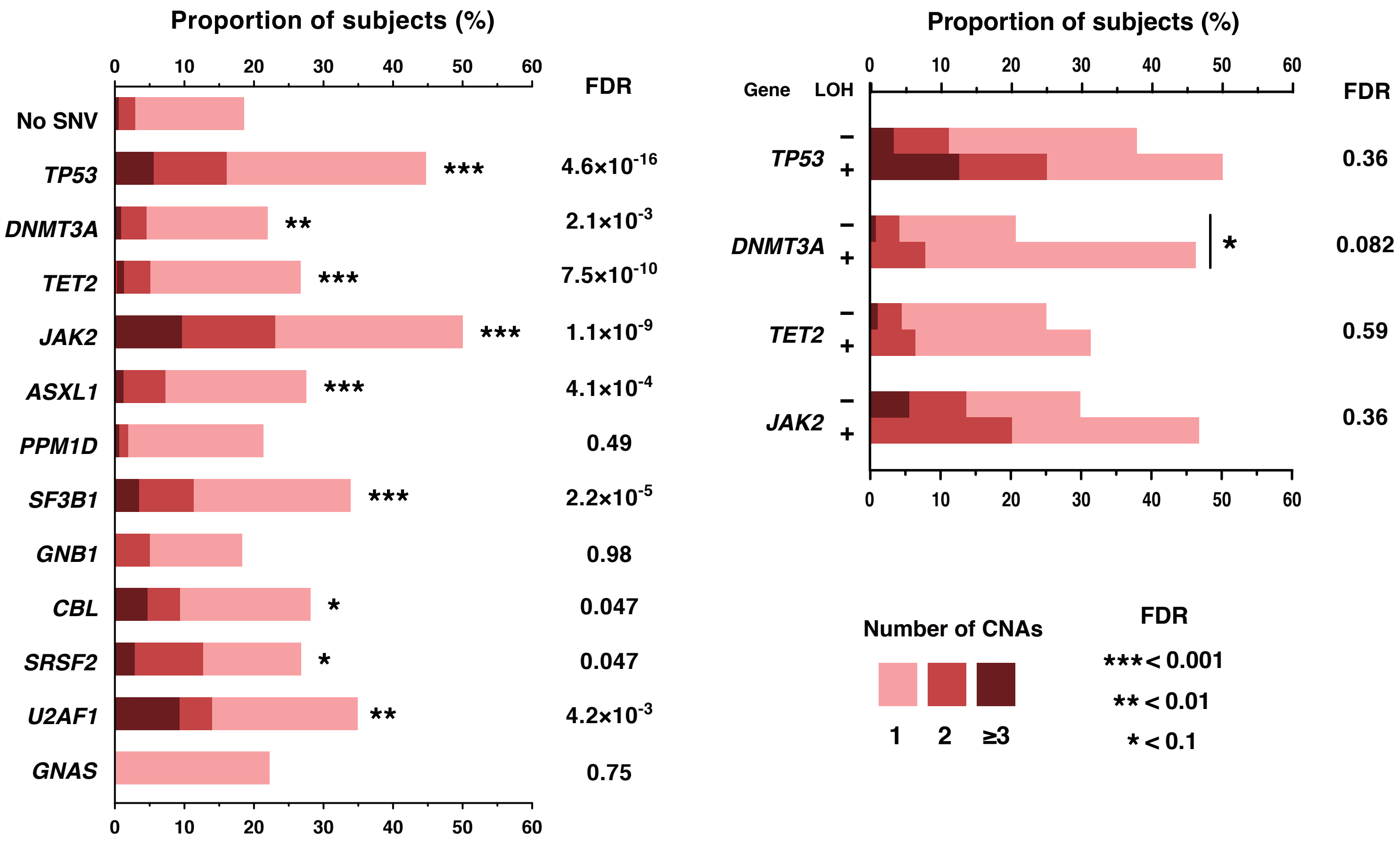
Supplmentary Fig. 3



Supplementary Fig 3. Significant cooccurrences of SNVs/indels and CNAs in CH.

The result of age-stratified permutation test (Online methods). The gray area indicates null distribution of the number of cases with cooccurring SNVs/indels and CNAs which was generated in 100,000 times age-stratified permutation.

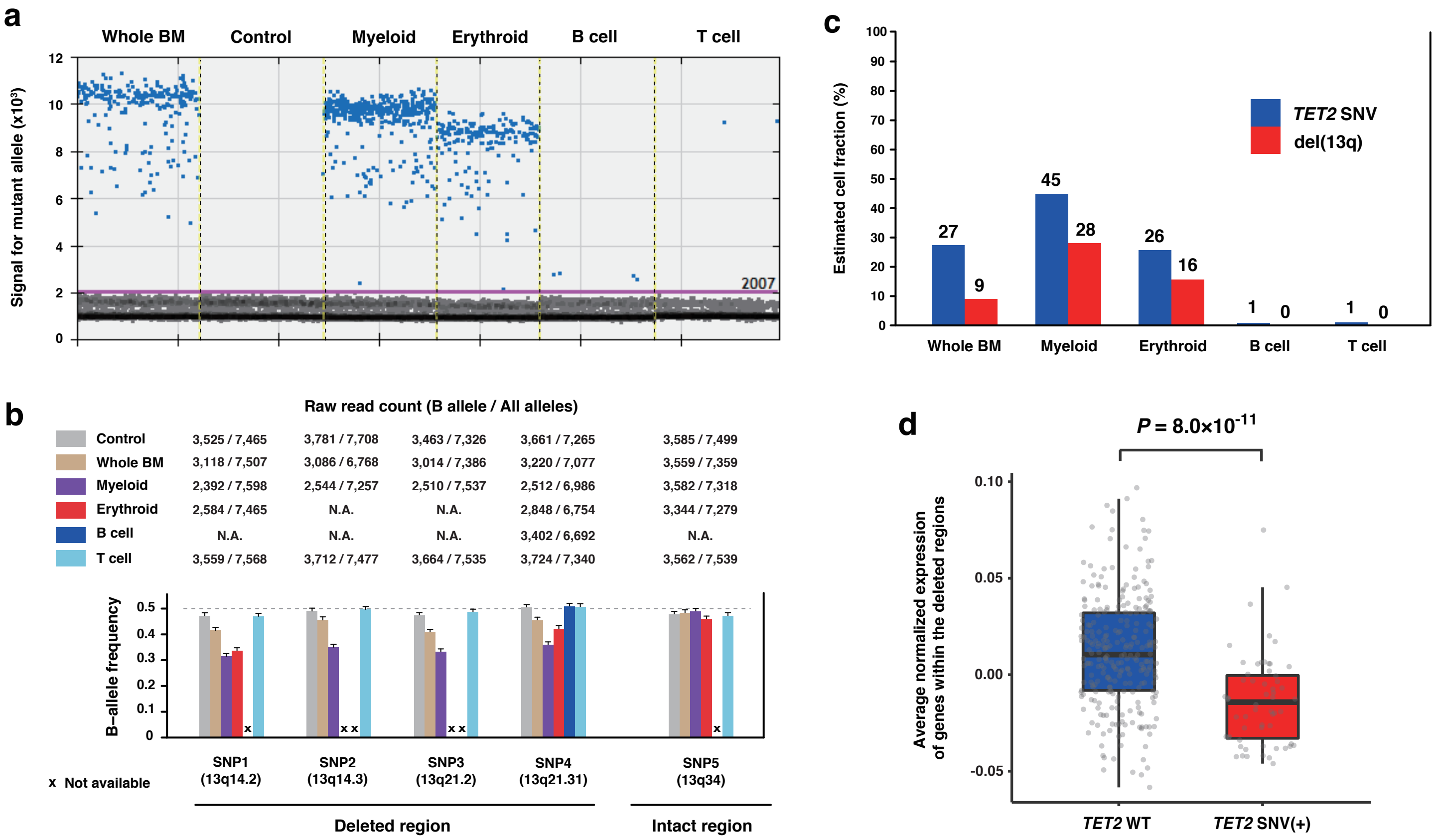
Supplementary Fig. 4



Supplementary Fig. 4 l Relationships of SNVs/indels and number of cooccurring CNAs.

Proportions of subjects with CNAs within those who harbor SNVs/indels in the indivated genes (left) and comparison of the number of cooccurring CNAs between subjects with SNVs/indels with or without LOH in *TP53*, *DNMT3A*, *TET2*, and *JAK2* (right). The proportions of subjects with 1, 2, and ≥3 CNAs are depicted by different colors. In the left, the numbers of CNAs are compared with subjects without SNVs/indels (labeled as “No SNV”) by twp-sided Wilcoxon test. In the right panel, we did not counted CNAs responsible for the LOH in the number of cooccurring CNAs and FDR was calculated by comparing those with and without LOH. Significantly larger numbers of CNAs are indicated by asterisks.

Supplementary Fig. 5



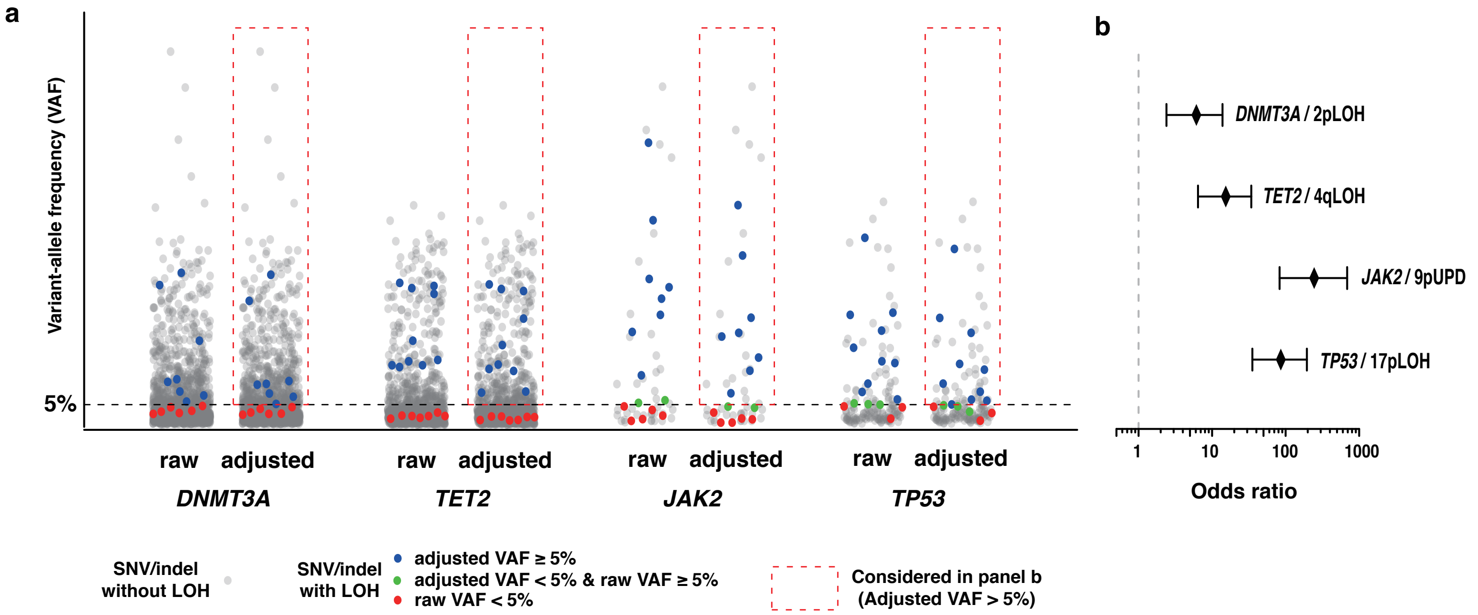
Raw read count (B allele / All alleles)

Control	3,525 / 7,465	3,781 / 7,708	3,463 / 7,326	3,661 / 7,265	3,585 / 7,499
Whole BM	3,118 / 7,507	3,086 / 6,768	3,014 / 7,386	3,220 / 7,077	3,559 / 7,359
Myeloid	2,392 / 7,598	2,544 / 7,257	2,510 / 7,537	2,512 / 6,986	3,582 / 7,318
Erythroid	2,584 / 7,465	N.A.	N.A.	2,848 / 6,754	3,344 / 7,279
B cell	N.A.	N.A.	N.A.	3,402 / 6,692	N.A.
T cell	3,559 / 7,568	3,712 / 7,477	3,664 / 7,535	3,724 / 7,340	3,562 / 7,539

Supplementary Fig. 5 | Analysis of a representative case with a SNV in *TET2* and del(13q).

a, Results of ddPCR for A1153V substitution in *TET2* performed on DNA samples extracted from whole bone marrow cells, myeloid cells (CD13/33+), erythroid cell (CD235a+), B cells (CD19+), and T cells (CD3+). Sample for negative control is taken from a CH-negative subject. b, Raw read counts and B-allele frequencies (BAF) for 4 heterozygous SNPs (Supplementary Table 6) within del(13q) and one in an intact region. Beucase of small amounts of DNA, data for SNP1, 2, 3, and 5 were not available for B cell, and that for SNP 2 and 3 were not available for erythroid. In the barplot, error bars indicate upper limits of 95% confidence intervals for the estimated fraction of B allele. N.A., not available. c, Cell fractions of the SNV in *TET2* and del(13q) in each fraction. Cell fraction for the *TET2* SNV was calculated as 2 × VAF. That for del(13q) were calculated on the basis of allelic imbalance observed at SNP4 in (b). d, Results of single-cell gene expression analysis and SNV detection in Fluidigm C1 platform. Average normalized expression of genes within del(13q), which can be a surrogation for DNA copy-number of the deleted region, are plotted for each cell with or without A1153V substition in *TET2*. *P* value is calculated by two-sided Wilcoxon rank sum test. The box plot indicates the median, first and third quartiles (Q1 and Q3) and whiskers extend to the furthest value between Q1 – 1.5×the interquartile range (IQR) and Q3 + 1.5×IQR.

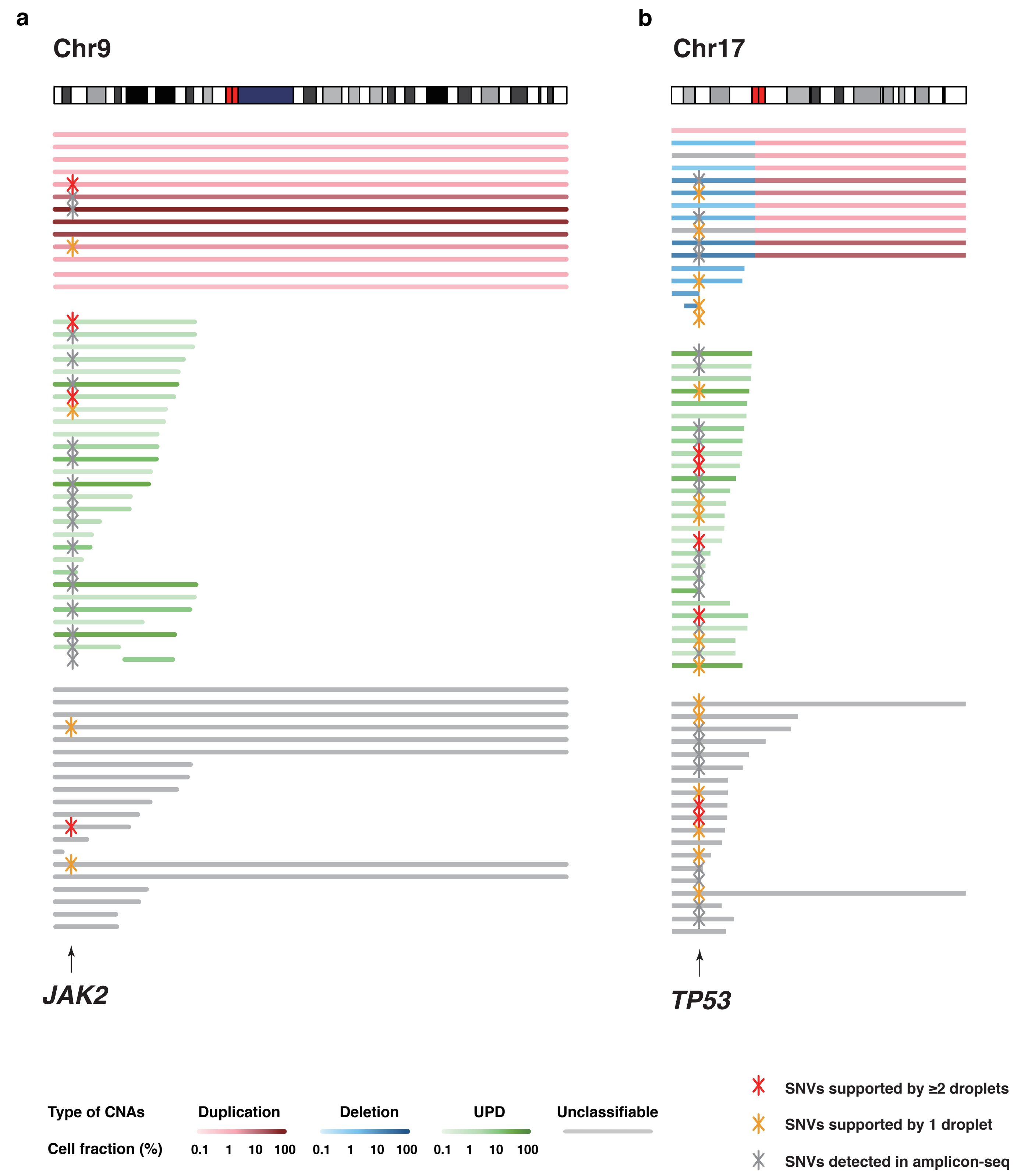
Supplmentary Fig. 6



Supplementary Fig. 6. Analysis of the cooccurrences of SNVs/indels and CNAs adjusted for the effect of VAF inflation.

a, Distributions of raw and adjusted VAF (Online methods) for SNVs/indels in *DNMT3A*, *TET2*, *JAK2*, and *TP53*. b, Odds ratios for cooccurrences of corresponding SNVs/indels and CNAs. Only SNVs/indels enclosed by red rectangles in (a) were taken into consideration. Error bars indicate 95% confidence intervals.

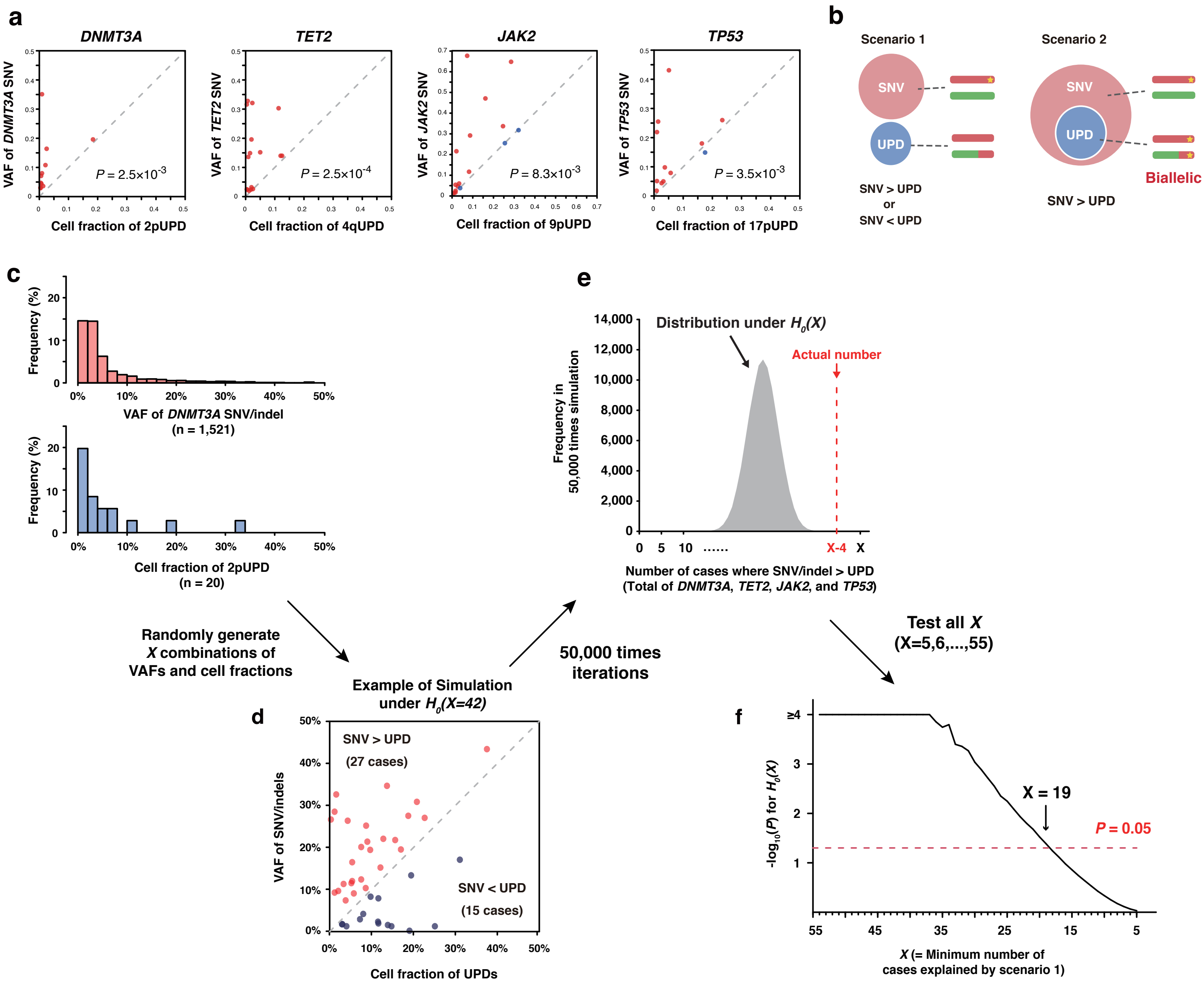
Supplementary Fig. 7



Supplementary Fig. 7 | ddPCR for mutational hotspots in *JAK2* and *TP53*.

a-b, Hotspot SNVs newly detected in ddPCR are illustrated by red or orange asterisks. We tested V617F in *JAK2* and R175H, Y220C, R248Q/W, R273C/H in *TP53*. SNVs supported by multiple droplets are shown in red, those suppported by single in orange, and those already detected in targeted sequencing in gray.

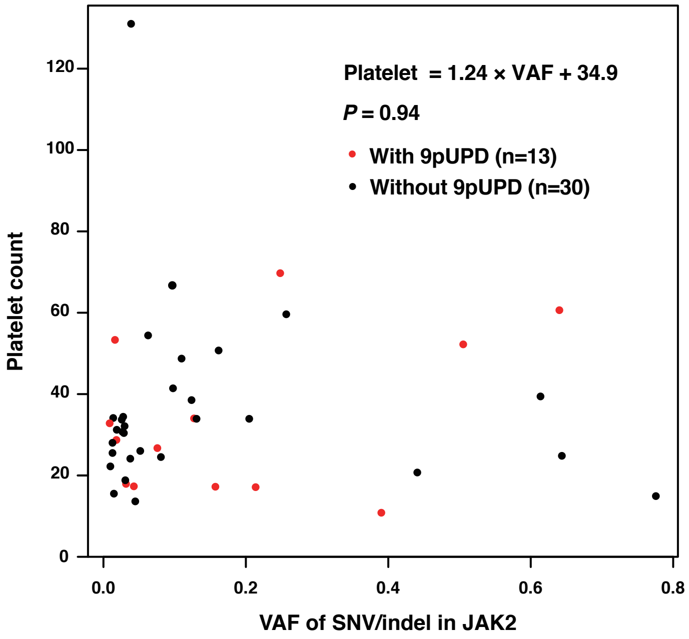
Supplementary Fig. 8



Supplementary Fig. 8 | Simulation for VAF/cell fractions of coocurring SNVs/indels and UPDs.

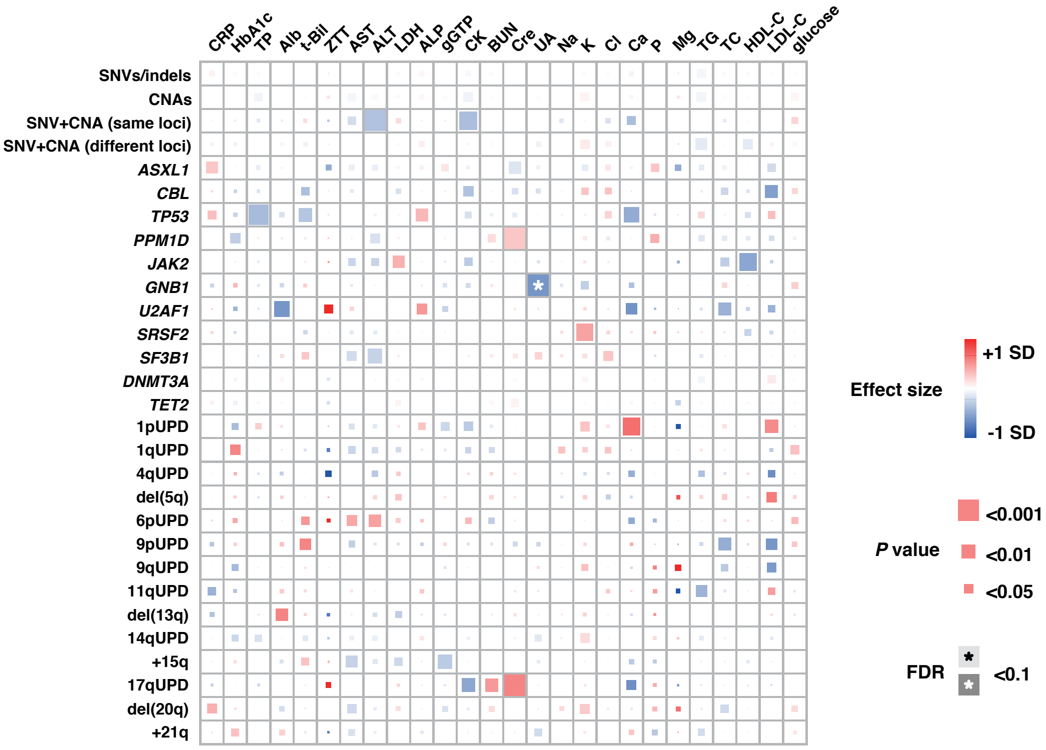
a, Comparison of VAFs and cell fractions of SNV/indels and UPDs involving the same genes: *DNMT3A* (n=11, a), *TET2* (n=16, b), *JAK2* (n=16, c), and *TP53* (n=12, d). Compared with cell fractions of UPDs, VAF of SNVs/indels were larger in 51 of the 55 (red) and smaller in the remaining 4 cases (blue). *P* values were calculated by random simulations as described below. b, Two possible scenarios underlying the observations in (a). In scenario 1, SNVs/indels and UPDs exist in discrete cells. In that case, VAF of SNVs/indels can be larger or smaller than cell factions of UPDs. In scenario 2, UPD subclonally exists within the clone carrying SNVs/indels, causing biallelic alterations. The 54 observations in (a) can be regarded as a mixture of the two scenarios. In the following simulation, we put a null hypothesis $H_0(X)$, that at least X of the 54 cases in (a) are explained by scenario 1. c, Histgrams of VAFs of *DNMT3A* SNVs/indels and cell fractions of 2pUPDs in the entire cohort. In the simulation, we randomly sample *X* combinations of the VAF of SNVs/indels and cell fractions of UPDs from these distributions. Only distributions of *DNMT3A*/2pUPD are shown, but we also sample VAFs and cell fractions from the distributions of *TET2*/4qUPD, *JAK2*/9pUPD and *TP53*/17pUPD as well (not shown). d, An example of simulated combinations of VAF and cell fractions. Here, we supposed *X*=42 and simulated VAFs of SNVs/indels were bigger than cell fractions of UPDs in 27 cases. e, Iterating the procedures illustrated in (c) and (d), we obtained a null distribution of the number of cases in which VAFs of SNVs/indels were larger than cell fractions of UPDs (shown in gray). Comparing the null distribution with the actually observed number, *X*-2 (shown in red), we calculated P value for $H_0(x)$ (x=5,...,55) and looked for the minimum *X* with P<0.05. f, *P* values for $H_0(x)$ (x=5,6,...,55) are shown. The minimum *X* with $P < 0.05$ was 19, which suggested scenario 1 can explain less than 19 cases out of the 51 cases in which VAFs of SNVs/indels were larger than cell fractions of UPDs. Thus, the remaining 32 cases should be explained by scenario 2.

Supplmentary Fig. 9



Supplementary Fig. 9 | The relationship between VAF of SNVs in *JAK2* and platelet counts.

Supplementary Fig. 10



Supplementary Fig. 10 | Association of CH wtih blood test values.

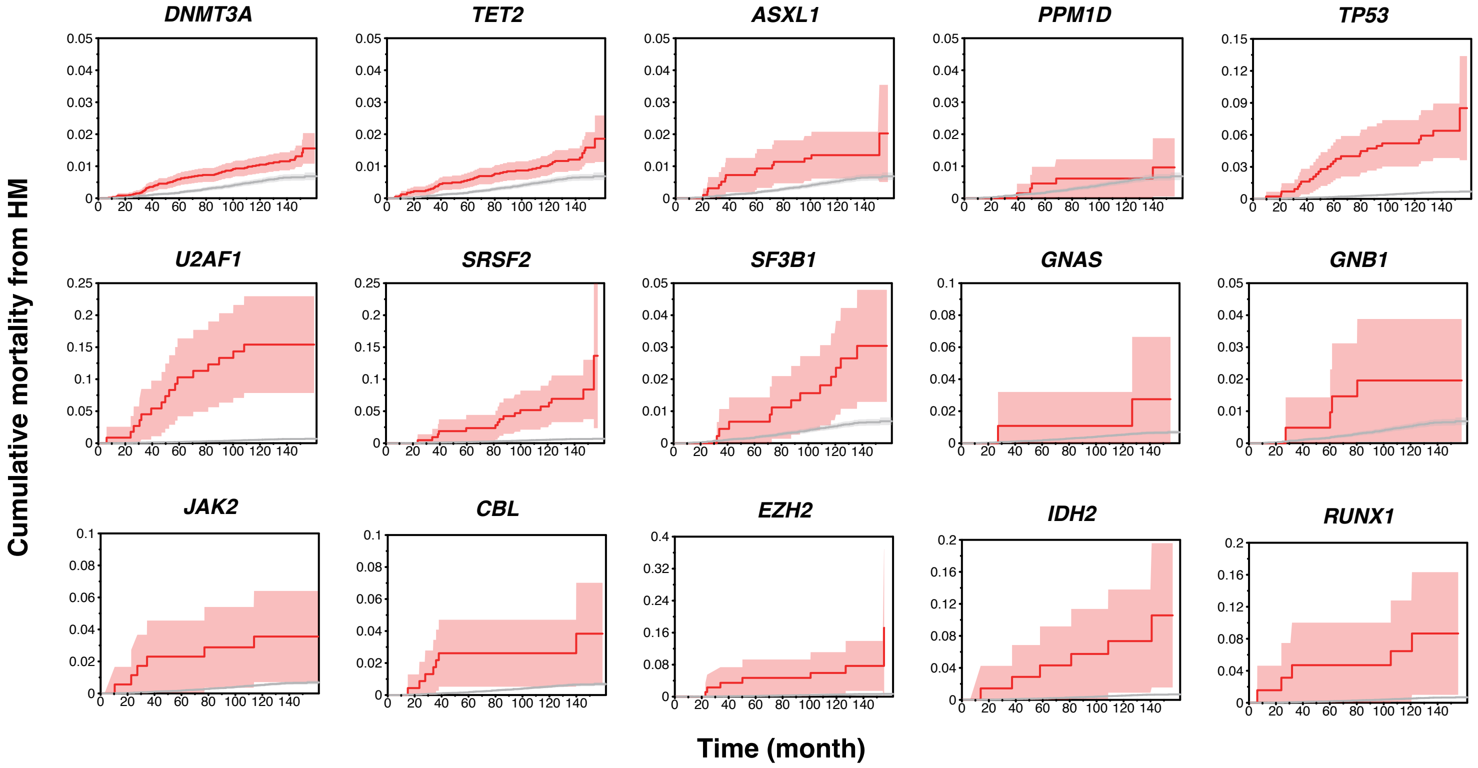
Positive or negative correlation between CH-related alterations and blood test values are illustrated in red or blue rectangles, respectively. Only the correlation between *GNB1*-involving SNVs and uric acid achieved statistical significance (P=1.3×10^{-4}, FDR=0.094). Exact effect size, P value, and FDR are shown in Supplementary Data. CRP, C-reactive protein; HbA1c, Hemoglobin A1c; TP, total protain; Alb, albumine; t-Bil, total bilirubin; ZTT, zinc sulfate turbidity test; AST, aspartate aminotransferase; ALT, alanine aminotransferase; LDH, lactate dehydrogenase; ALP, alkaline phosphatase; gGTP, gamma-glutamyltransferase; CK, creatinine kinase; BUN, blood urea nitrogen; Cre, creatinine; UA, uric acid; Na, sodium ion; K, potassium ion; Cl, chloride ion; Ca, calcium ion; P, phosphate ion; Mg, magnesium ion; TG, triglycerol; TC, total choresterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol.

Supplmentary Fig. 11

	Current study			Jaiswal *et al.*[13]			Genovese *et al.*[12]			Laurie *et al.*[9]		
	CH(+)	CH(−)	Total	CH(+)	CH(−)	Total	CH(+)	CH(−)	Total	CH(+)	CH(−)	Total
All HM	376	296	672	5	11	16	15	12	27	14	90	104
Myeloid	160	55	215	3	1	3	7	0	7	5	0	5
AML	63	27	90	1	1	1	2	0	2	2	0	2
MDS	75	25	100	1	0	1	3	0	3	1	0	1
MPN	4	1	5	1	0	1	2	0	2	1	0	1
CML	7	2	9	0	0	0	0	0	0	0	0	0
Others	11	0	11	0	0	0	0	0	0	1	0	1
Lymphoid	191	229	420	2	6	9	6	0	6	9	0	9
B-NHL	123	143	266	1	0	3	1	0	1	1	0	1
T-NHL	15	17	32	0	0	0	0	0	0	0	0	0
CLL	7	0	7	0	0	0	3	0	3	5	0	5
ALL	7	12	19	0	0	0	0	0	0	0	0	0
MM	36	53	89	0	2	2	2	0	2	1	0	1
Others	3	4	7	1	4	4	0	0	0	2	0	2
Others/Unknown	25	12	37	0	4	4	2	12	14	0	90	90
No HM	4,209	6,353	10,562	741	16,425	17,166	N.A.	N.A.	12,353	381	49,818	50199
Total	4,585	6,649	11,234	746	16,436	17,182	N.A.	N.A.	12,380	404	49,818	50,222

Supplementary Fig. 11 I Number and diagnosis of hematological malignancies in the current and previous studies.
CH, clonal hematopoiesis; HM, Hematological malignancies; 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; N.A., not available.

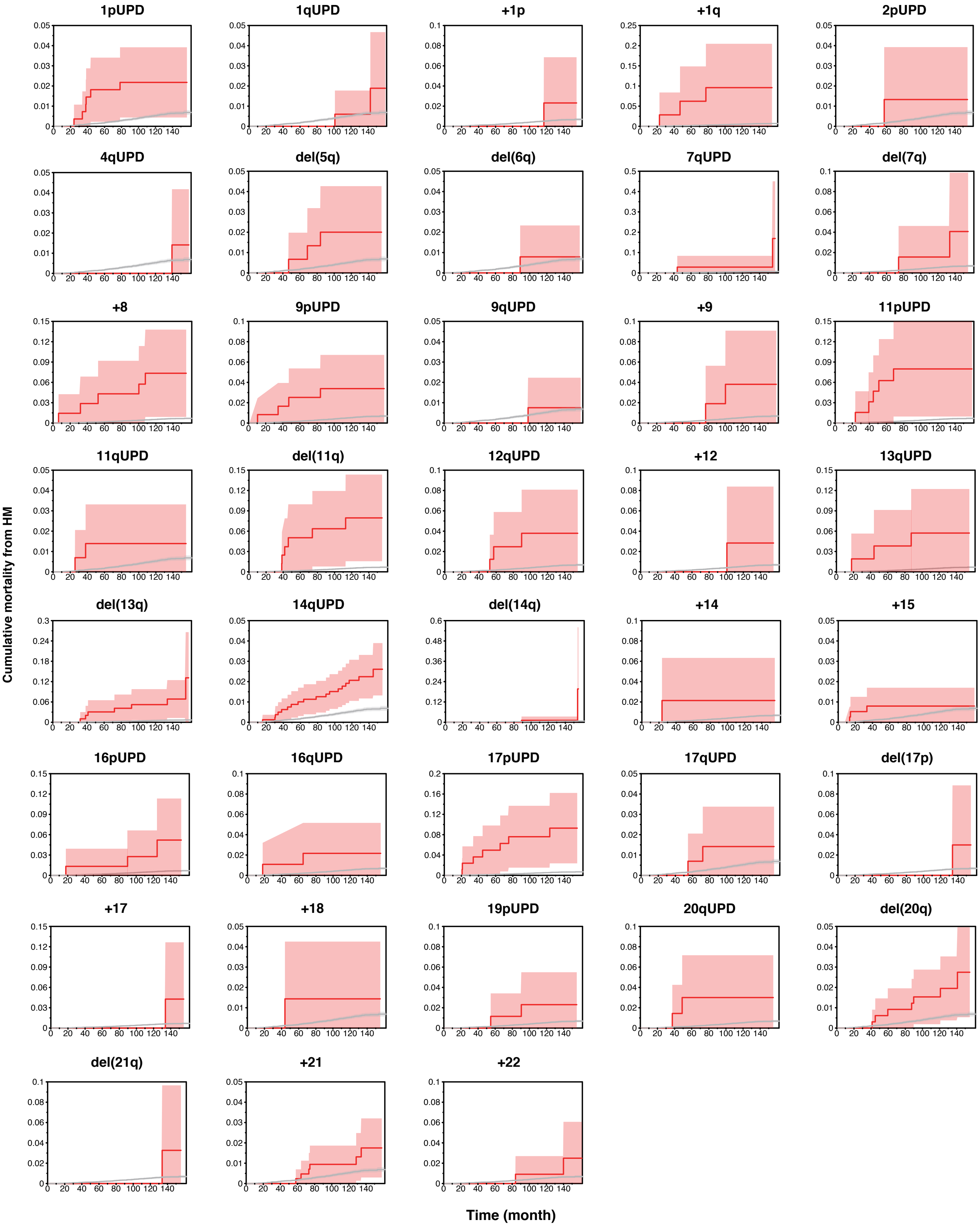
Supplementary Fig. 12



Supplementary Fig. 12 | Cumulative mortality from HM in subjects with individual SNVs/indels.

Cumulative mortality from HM in subjects with SNVs/indels in the indicated genes are shown by red lines. For comparison, cumulative mortality from HM in subjects without any alteration is also shown by gray lines. Colored bands indicate 95% confidence intervals.

Supplementary Fig. 13



Supplementary Fig. 13 l Cumulative mortality from HM in subjects with individual CNAs.

Cumulative mortality from HM in subjects with the indicated CNAs are shown by red lines. For comparison, cumulative mortality from HM in subjects without any alteration is also shown by gray lines. Colored bands indicate 95% confidence intervals.

Supplementary Table 1. Demographic summary of subjects.

Category	HM (+) (n = 672)		HM (-) (n = 10,562)	
Characteristics	No. of subjects (%)	Median (range)	No. of subjects (%)	Median (range)
Age at sampling (n = 11,234)		71 (25-94)		70 (60-101)
<50	18 (2.7)		0 (0.0)	
50-59	63 (9.4)		0 (0.0)	
60-69	204 (30.4)		4,826 (45.7)	
70-79	299 (44.5)		4,287 (40.6)	
80-89	84 (12.5)		1,350 (12.8)	
90-101	4 (0.6)		99 (0.9)	
Gender (n = 11,234)				
Female	234 (34.8)		4,933 (46.7)	
Male	438 (65.2)		5,629 (53.3)	
BMI (n = 10,532)		22.9 (14-39.7)		23.1 (12-52.2)
<25	101 (15.9)		1,433 (14.5)	
≥25, <30	514 (80.8)		8,115 (82.0)	
≥30	21 (3.3)		348 (3.5)	
History of smoking (n = 11,192)				
Current or past smoker	387 (58.4)		4,997 (47.5)	
Non-smoker	276 (41.6)		5,532 (52.5)	
History of drinking (n = 11,158)				
Current or past drinker	351 (53.4)		4,828 (46.0)	
Non-drinker	306 (46.6)		5,673 (54.0)	
Hypertension (n = 9,671)				
-	387 (66.6)		5,440 (59.8)	
+	162 (27.9)		2,817 (31.0)	
++	29 (5.0)		678 (7.5)	
+++	3 (0.5)		155 (1.7)	
Hyperlipidemia (n = 11,234)				
-	179 (26.6)		3,895 (36.9)	
+	493 (73.4)		6,667 (63.1)	
Diabetes mellitus (n = 11,234)				
-	171 (25.4)		3,177 (30.1)	
+	501 (74.6)		7,385 (69.9)	

HM (+/-): Subjects with/without event of hematological malignancies during follow-up periods.

Hypertension -: systolic blood pressures(sBP) < 140 and diastolic blood pressures(dBP) < 90, +: sBP≥140 or dBP≥90, ++: sBP≥160 or dBP≥100, +++: sBP≥180 or dBP≥110.

Supplementary Table 2. Number of subjects with individual target diseases.

Disease group	Disease name	Number of subjects (%)	
		HM (+) (n = 672)	HM (-) (n = 10,562)
Malignant tumors	Lung cancer	27 (4.0)	0 (0.0)
	Esophageal cancer	13 (1.9)	0 (0.0)
	Gastric cancer	42 (6.3)	0 (0.0)
	Colorectal cancer	29 (4.3)	0 (0.0)
	Liver cancer	5 (0.7)	0 (0.0)
	Pancreas cancer	2 (0.3)	0 (0.0)
	Gallbladder/Cholangiocarcinoma	3 (0.4)	0 (0.0)
	Prostate cancer	45 (6.7)	0 (0.0)
	Breast cancer	20 (3.0)	0 (0.0)
	Cervical cancer	2 (0.3)	0 (0.0)
	Uterine cancer	4 (0.6)	0 (0.0)
	Ovarian cancer	1 (0.1)	0 (0.0)
Cerebral diseases	Cerebral infarction	90 (13.4)	1740 (16.5)
	Cerebral aneurysm	6 (0.9)	208 (2.0)
	Epilepsy	7 (1.0)	89 (0.8)
Respiratory diseases	Bronchial asthma	24 (3.6)	494 (4.7)
	Pulmonary tuberculosis	**6** (0.9)	29 (0.3)
	Chronic obstructive pulmonary disease	17 (2.5)	263 (2.5)
	Interstitial lung disease/Pulmonary fibrosis	7 (1.0)	68 (0.6)
Cardiovascular diseases	Myocardial infarction	73 (10.9)	1173 (11.1)
	Unstable angina	31 (4.6)	521 (4.9)
	Stable angina	97 (14.4)	1630 (15.4)
	Arrhythmia	96 (14.3)	1522 (14.4)
	Heart failure	41 (6.1)	909 (8.6)
	Peripheral arterial diseases	19 (2.8)	373 (3.5)
Liver diseases	Chronic hepatitis B	4 (0.6)	36 (0.3)
	Chronic hepatitis C	29 (4.3)	278 (2.6)
	Liver cirrhosis	11 (1.6)	83 (0.8)
Urologic diseases	Nephrotic syndrome	4 (0.6)	42 (0.4)
	Urolithiasis	2 (0.3)	263 (2.5)
Metabolic diseases	Osteoporosis	44 (6.5)	770 (7.3)
	Diabetes mellitus	171 (25.4)	3177 (30.1)
	Dyslipidemia	179 (26.6)	3895 (36.9)
Endocrine diseases	Graves' disease	3 (0.4)	74 (0.7)
Connective tissue diseases	Rheumatoid arthritis	26 (3.9)	292 (2.8)
Allergic diseases	Hay fever	7 (1.0)	176 (1.7)
Dermatologic diseases	Drug eruption	1 (0.1)	31 (0.3)
	Atopic dermatitis	1 (0.1)	13 (0.1)
	Keloid	3 (0.4)	19 (0.2)
Gynecologic diseases	Uterine fibroid	2 (0.3)	45 (0.4)
	Endometriosis	0 (0.0)	1 (0.0)
Pediatric diseases	Febrile seizure	0 (0.0)	0 (0.0)
Ophthalmologic diseases	Glaucoma	21 (3.1)	446 (4.2)
	Cataract	87 (12.9)	1991 (18.9)
Dental diseases	Periodontitis	1 (0.1)	131 (1.2)
Other	Amyotrophic lateral sclerosis	0 (0.0)	0 (0.0)

HM (+/-): Subjects with/without event of hematological malignancies during follow-up periods.

Supplementary Table 3. Summary of blood cell counts.

Blood cell count	HM(+) (n=672)		HM(-) (n=10,562)	
	Median (range)	No. of subjects (%)	Median (range)	No. of subjects (%)
White blood cell (/μL)	5450 (840-37500)	551	5,900 (1,050-27,600)	8,519
Normal		507 (92.0)		8,092 (95.0)
≥10000		18 (3.2)		341 (4.0)
<3000		26 (4.7)		86 (1.0)
Hemoglobin (g/dL)	13.2 (4.8-17.7)	573	13.5 (3.2-19.0)	8,651
Normal		517 (90.2)		7,814 (90.3)
≥16.5 (male), 16 (female)		22 (3.8)		427 (4.9)
<10		34 (5.9)		410 (4.7)
Hematocrit (%)	39.7 (21.8-52.6)	571	40.4 (15.5-69.4)	8,645
Normal		561 (98.2)		8,458 (97.8)
≥50		10 (1.8)		187 (2.2)
Platelet (10^4/μL)	20.0 (1.1-131)	511	21.0 (1.1-387)	8,117
Normal		477 (93.3)		7,920 (97.6)
≥45		7 (1.4)		54 (0.7)
<10		27 (5.3)		143 (1.8)

HM (+/-): Subjects with/without event of hematological malignancies during follow-up periods.

Supplementary Table 4. Antibodies for cell sorting.

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

Supplementary Table 5. Primer sequences for detection of allele imbalances in the regions of del(13q).

SNP ID	Status	SNP position	Forward primer sequence	Reverse primer sequence
rs731779	Deleted	chr13:47452038	AAGCGGCCGCAAAGCAGGGCAAGTACCTCA	AAGCGGCCGCTGAGTGTCTCTCTTGCCCCA
rs1350457	Deleted	chr13:54355150	AAGCGGCCGCGGTAAGAATACAAACCTGGAAAAAGTG	AAGCGGCCGCCCTTGGACCCGCTTCACTC
rs341506	Deleted	chr13:60420314	AAGCGGCCGCACACAGGCTTTCCTCCAAGT	AAGCGGCCGCTGTGTAAGAGTGAGTGTGGCA
rs359362	Deleted	chr13:65239972	AAGCGGCCGCTTGGTCAAATGGCACCCCTT	AAGCGGCCGCCAATTAGATTTGGAATTTGCTTGTGA
rs4773419	Intact	chr13:112311079	AAGCGGCCGCAAGAAAGGCAGGTCCAAGGG	AAGCGGCCGCGTGTTGACAAAGCCGGTTGG

Supplementary Table 6. Probes used for ddPCR.

Gene	Amino acid substitution	BioRad Assay ID
TET2	p.A1153V	dHsaMDS869039740
JAK2	p.V617F	dHsaMDS488977115
TP53	p.R175H	dHsaMDV2010105
TP53	p.Y220C	dHsaMDV2510536
TP53	p.R248Q	dHsaMDV2010127
TP53	p.R248W	dHsaMDV2010107
TP53	p.R273H	dHsaMDV2010109
TP53	p.R273C	dHsaMDV2510538