Genetic drivers and clinical consequences of mosaic chromosomal alterations in 1 million individuals

Authors: Kun Zhao[1,2]*, Yash Pershad[1,2]*, Hannah M Poisner[1,2], Xiaolong Ma[3], Kali Quade[3], Caitlyn Vlasschaert[4], Taralynn Mack[1,2], Nikhil K Khankari[1,2], Kelly von Beck[1,2], James Brogan[1], Ashwin Kishtagari[1], Robert W Corty[1], Yajing Li[5,6], Yaomin Xu[5,6], Alexander P Reiner[7,8], Paul Scheet[9], Paul L Auer[3], Alexander G Bick[1,2]

Affiliations: [1]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [2]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA; [3]Division of Biostatistics, Data Science Institute and Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA; [4]Department of Medicine, Queen's University, Kingston, Ontario, Canada; [5]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA; [6]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [7]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center. Seattle, WA, USA; [8]Department of Epidemiology, University of Washington, Seattle, WA, USA; [9]Department of Epidemiology, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

*These authors contributed equally to this work

Correspondence to: Dr. Bick, Alexander.Bick@VUMC.org

Address: 550 Robinson Research Building, Nashville, TN 37232

Word Count: 4,624

Figures: 4

Table: 1

Abstract

Mosaic chromosomal alterations of the autosomes (aut-mCAs) are large structural somatic mutations which cause clonal hematopoiesis and increase cancer risk. Here, we detected aut-mCAs in 1,011,269 participants across four biobanks. Through integrative analysis of the minimum critical region and inherited genetic variation, we found that proto-oncogenes exclusively drive chromosomal gains, tumor suppressors drive losses, and copy-neutral events can be driven by either. We identified three novel inherited risk loci in *CHI3L2*, *HLA* class II, and *TERT* that modulate aut-mCA risk and ten novel aut-mCA-specific loci. We found specific aut-mCAs are associated with cardiovascular, cerebrovascular, or kidney disease incidence. High-risk aut-mCAs were associated with elevated plasma protein levels of therapeutically actionable targets: NPM1, PARP1, and TACI. Participants with multiple high-risk features such as high clonal fraction, more than one aut-mCA, and abnormal red cell morphology had a 50% cumulative incidence of blood count abnormalities over 2 years. Leveraging inherited variation, we causally established aut-mCAs as premalignant lesions for chronic lymphocytic leukemia. Together, our findings provide a framework integrating somatic mosaicism, germline genetics, and clinical phenotypes to identify individuals who could benefit from preventative interventions.

Introduction

As individuals age, hematopoietic stem cells (HSCs) accumulate somatic mutations, some of which enhance HSC fitness and result in peripheral blood cells sharing the same mutation. This phenomenon, termed clonal hematopoiesis, affects 10% of people over the age of 60 and increases risk of all-cause mortality, hematologic malignancy, cardiovascular disease, and infection.[1–6] One way to classify clonal hematopoiesis is by the type of driver mutation detectable from blood-derived DNA. Single-nucleotide variants and small insertions or deletions in certain preleukemic genes are termed clonal hematopoiesis of indeterminate potential (CHIP).[1,7–9] Larger >1 megabase gains (+), losses (-), or copy-neutral losses of heterozygosity (=) are termed mosaic chromosomal alterations (mCAs).[10–15] mCAs of sex chromosomes such as mosaic loss of Y and loss of X are the most common type of mCA and have been well-characterized.[16,17] In contrast, mCAs of the autosomes (aut-mCAs) are more heterogeneous in type and phenotypic consequence.[4,18–20]

Prior work has shown that aut-mCAs increase in prevalence with age, the presence of aut-mCAs is influenced by inherited genetic variants, and aut-mCAs often co-occur with blood count abnormalities and increase risk of chronic lymphocytic leukemia (CLL) and severe infections.[3,13,18,21] However, several important questions about aut-mCAs remain unanswered. First, for most aut-mCAs, the driver gene conferring their selective advantage remains unknown. Second, research on aut-mCAs has been limited to single-ancestry populations (e.g., UK Biobank[13,14] and Biobank Japan[15]), which along with sample size, has limited the study of how germline genetics influence aut-mCA prevalence. Third, unlike in CHIP where it is well-known that different driver genes have distinct risk factors and phenotypic consequences, researchers have not had sufficient sample size to profile the heterogeneity of specific aut-mCAs. Fourth, while certain aut-mCAs are associated cross-sectionally with abnormal blood counts,[10,13,14] for a person with an aut-mCA but normal blood counts, it is not well-understood what their risk of developing incident blood count abnormalities, the earliest sign of progression to hematologic malignancy. Fifth, there are no candidate preventative therapies which may prevent progression.

Here, using 1,011,269 individuals across four large-scale biobanks, NIH All of Us (AoU), NHLBI TOPMed, Vanderbilt's BioVU, and UK Biobank (UKB) with genetic and longitudinal clinical and laboratory data, we have constructed the largest and most diverse, deeply phenotyped cohort of aut-mCAs (**Extended Data Fig.1**). Using this data, we identified putative causal drivers for all aut-mCAs, discovered novel aut-mCA-specific and aut-mCA-wide germline genetic variants which alter risk of aut-mCAs, and reported the proteomic signature of aut-mCAs to nominate therapeutic targets. We are the first to delineate the health consequences of specific aut-mCA types beyond hematologic malignancy. Then, we identified a subgroup of people with aut-mCAs at high risk of incident blood count abnormalities. Finally, we demonstrated that aut-mCAs cause CLL and that polygenic germline risk and somatic aut-mCAs interact to alter CLL risk.

Results

The profile of autosomal mosaic chromosomal alterations

We profiled aut-mCAs in 1,011,269 individuals using genotyping arrays in UKB[13,14], BioVU[22], and AoU and whole genome sequencing in TOPMed[10] using MoChA (**Extended Data Fig.2**). Due to the expected low prevalence of aut-mCAs in those under 40 years old, we then focused on the 850,810 participants between 40-90 years old. The baseline characteristics of each cohort are shown in **Supplementary Table 1.** The participants in the study had a mean age at DNA collection of 58 years; the mean age of those with an aut-mCA was 63 years. 55% of participants were male. 80% clustered with the 1000 Genomes European ancestry group (1KG-EUR-like), 10% with the 1000 Genomes African ancestry group (1KG-AFR-like), and 2.5% with the 1000 Genomes East Asian ancestry group (1KG-EAS-like). 23,766 (2.70%) individuals had ≥1 detectable aut-mCA. Among those aut-mCAs, 86% had one, 8% had two, 6% had multiple (≥3) (**Extended Data Fig.3**). The prevalence of aut-mCA increased with age, with 1KG-EUR-like and 1KG-AFR-like populations having higher prevalence of aut-mCAs than 1KG-EAS-like populations in each age group (**Extended Data Fig.4**).

Certain mCA subtypes were markedly overrepresented in different age and sex groups. A majority of people with chr15+ (72%) and chr20q– (67%) aut-mCAs were males (**Figure 1A**). The male bias of chr15+ and chr20q– persists after exclusion of those with mosaic loss of chromosome X and Y (**Extended Data Fig.5)**. People with = aut-mCAs were on average younger than + or – aut-mCAs (**Figure 1A**). Prevalence of certain aut-mCAs also varied by genetic ancestry: chr15+ was more prevalent in those of 1KG-AFR-like ancestry (1.4% of 1KG-AFR-like and 1.1% of 1KG-EUR-like, $P = 0.008$), while chr13q– (0.8% of 1KG-AFR-like and 2.5% in 1KG-EUR-like, $P = 1\times10^{-15}$) and chr14q= (0.6% in 1KG-AFR-like and 2.0% in 1KG-EUR-like, $P = 6\times10^{-13}$) were more prevalent in 1KG-EUR-like individuals (**Figure 1B**).

CHIP and autosomal mosaic chromosomal alterations co-occurrence

Given that co-occurring somatic mutations may influence the prevalence of aut-mCAs, we characterized the patterns of co-occurrence between CHIP mutations and aut-mCAs. Among the 23,766 people with ≥1 aut-mCA, 2,014 (8.5%) also had detectable CHIP. CHIP mutations tended to co-occur with = aut-mCAs of the same chromosome (**Figure 1C**), such as *TET2*-chr4= (OR: 7.5, 95% CI 5.5-10.2, $P = 1\times10^{-35}$), *MPL*-chr1= (OR: 31.7, 95% CI 12.0-84.9, $P = 2\times10^{-8}$), *JAK2*-chr9= (OR: 36.3, 95% CI 27.6-48.0, $P = 3\times10^{-162}$), *DNMT3A*-chr2= (OR: 9.1, 95% CI 4.7-18.6, $P = 3\times10^{-11}$), *CREBBP*-chr16= (OR: 20.8, 95% CI 6.6-56.1, $P = 1\times10^{-3}$), *CBL*-chr11= (OR: 10.8, 95% CI 4.8-22.6, $P = 6\times10^{-5}$), and *TP53*-chr17= (OR: 13.2, 95% CI 6.4-28.5 $P = 1\times10^{-10}$) (**Supplementary Table 2**). Only three CHIP genes tended to co-occur with aut-mCAs with copy number changes. Chr9+ and chr12+ co-occurred with CHIP mutations in proto-oncogenes *JAK2* (OR: 9.34, 95% CI 3.96-22.06, $P = 1.75\times10^{-4}$) and *KRAS* (OR: 24.73, 95% CI 8.86-63.59, $P = 1.47\times10^{-5}$). Chr17– co-occurred with CHIP mutations in *TP53* (OR: 13.8, 95% CI 8.2-23.8, $P = 3\times10^{-22}$). *TP53* CHIP tended to co-occur with a wide range of loss aut-mCAs in addition to chr17q.

Among CHIP-aut-mCA pairs, the CHIP mutations had greater cell fractions than the aut-mCAs on the same chromosome (i.e., $\log_{10}$ of the ratio of cell fraction of CHIP to aut-mCA < 1), implying that the CHIP mutation preceded the aut-mCA. Notably, *MPL* and *TET2* CHIP cell fraction was much higher than chr1p= and chr4q= in people with both mutations, whereas the difference in cell fraction between *TP53* and chr17p– tended to be smaller (**Figure 1D**).

Identification of the putative driver gene of autosomal mosaic chromosomal alterations

For most aut-mCAs, the driver gene conferring their selective advantage remains unknown. To identify the putative driver genes of aut-mCAs, we identified the smallest overlapping region for each aut-mCA chromosome and type in all participants and within this region nominated genes implicated in hematologic malignancy as a putative driver gene (**Methods**; **Supplementary Table 3**). This approach identified putative drivers for all aut-mCAs (**Table 1)**. Annotation of driver genes as tumor suppressors or proto-oncogenes revealed that every + aut-mCA contained proto-oncogenes as putative drivers, notably every – aut-mCA contained tumor-suppressors as putative drivers, and = aut-mCAs contained either proto-oncogenes and tumor-suppressor drivers. The smallest overlapping region of chr11+, –, and = contained *WT1* which may act as both a tumor suppressor and proto-oncogene in a context-specific manner.[23] Many aut-mCA putative drivers were genes implicated in CHIP such as *DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *SF3B1*, *TP53*, and *RUNX1*. This approach also clarifies prior observations. For example, others have identified germline genetic mutations in *FRA10B* as contributing to chr10q– risk, which is not known to drive hematologic malignancy.[14] *FRA10B* – a known site of chromosomal instability – is located near putative driver and tumor suppressor *SMC3*, which likely explains the proliferative advantage after loss of chr10q.[24]

Furthermore, in the UKB, burden tests of rare germline missense, frameshift, deletion, and stop-gain variants within 10 putative drivers altered risk of their aut-mCAs, thereby providing further evidence that these are in

fact the driver mutations (**Supplementary Table 4**). Missense variants in proto-oncogenes increased risk of gains, such as those in *JAK2* and *KRAS* for chr9p+ and chr12q+. Similarly, frameshift, deletion, and stop-gain variants increased risk of losses, such as *TET2* for chr4q–, *TP53* for chr17p–, and *CHEK2* for chr22q–. Interestingly, frameshift mutations in proto-oncogene *BCR* increased risk of chr22p+. Moreover, missense variants in *DLK1* and *TCL1A* – both of which are not considered proto-oncogenes or tumor suppressors – increased risk of chr14q aut-mCAs; these findings support that *DLK1* and *TCL1A* are putative drivers.

Germline 'cis' genetic risk of autosomal mosaic chromosomal alteration

Along with rare variants in driver genes, we hypothesized that more common germline variants would alter risk of specific aut-mCAs and may propel clonal selection. In a meta-analysis of the 4 cohorts, we identified 33 significant *cis* associations – that is, genetic variants which increase risk of an aut-mCA of the chromosome and arm of the variant – after multiple hypothesis correction for the number of variants tested on each respective chromosome arm. 23 significant associations were for – aut-mCAs, 10 for = aut-mCAs, and 0 for + aut-mCAs (**Figure 2A**; **Supplementary Table 5**). For = aut-mCAs, we confirmed associations with variants in *MPL* (chr1p=, OR: 3.1, 95% CI 2.1-4.7, $P = 3\times10^{-8}$), *JAK2* (chr9p=, OR: 1.8, 95% CI 1.6-2.0, $P = 2\times10^{-27}$), *ATM* (chr11q=, OR: 0.04, 95% CI 0.01-0.1, $P = 1\times10^{-10}$), and *TARS3* (chr15q=, OR: 120, 95% CI 71-202, $P = 2\times10^{-71}$) (**Extended Data Fig.6**). Novel associations with = aut-mCAs included variants in *LRPAP1* (chr4p=, OR: 1.1×10^{-4}, 95% CI 5.0×10^{-6}-2.0×10^{-3}, $P = 6\times10^{-9}$), *AP3M2* (chr8p=, OR: 3.9×10^{-4}, 95% CI 3.2×10^{-5}-5.0×10^{-3}, $P = 5\times10^{-10}$), and *PIK3C3* (chr18q=, OR: 0.009, 95% CI 0.002-0.04, $P = 2\times10^{-9}$). We also identified new associations with – aut-mCAs, including variants in *COL15A1* (chr9q–, OR: 2.6×10^{-5}, 95% CI 7.7×10^{-7}-1.1×10^{-3}, $P = 5\times10^{-9}$) and *SHOC2* (chr10q–, OR: 12.0, 95% CI 8.5-17, $P = 4\times10^{-45}$) (**Extended Data Fig.7**). *PI3KC3* is the only human class III PI-3-kinase and is involved in stem cell proliferation, *COL15A1* is a known tumor suppressor, and *SHOC2* has been observed as mutated in leukemias and is part of the RAS-MAPK signaling

pathway. The *PIK3C3* variant rs16974244-T was ancestry-specific, as it is common in 1KG-AFR-like populations (minor allele frequency = 0.068), but not present in 1KG-EUR-like or 1KG-EAS-like populations.

Germline 'trans' genetic risk of autosomal mosaic chromosomal alteration

In contrast to germline variants on the same chromosome altering risk of specific aut-mCAs, certain germline genetic variants altered the risk of any detectable aut-mCA. We identified 5 such genome-wide significant loci, three of which were novel (**Figure 2B**). Two loci which had been previously shown to be associated with aut-mCA risk in the UK Biobank by Loh et al[14] are (1) rs7705526-C (OR: 0.90, 95% CI 0.88-0.92, $P = 1\times10^{-20}$, 1KG-EUR-like allele frequency (AF) = 0.33, 1KG-AFR-like AF = 0.19) on chr5 in *TERT* and (2) rs1356532206-TA (OR: 1.1, 95% CI 1.05-1.10, $P = 8\times10^{-9}$, 1KG-EUR-like AF = 0.42, 1KG-AFR-like AF = 0.67) on chr2 in *SP140*. Using iterative conditional analysis, we found that rs2736100-C (D' = 0.99, $R^2 = 0.49$ with rs7705526) in *TERT* was an independent locus (**Supplementary Table 6**). The other two novel loci associated with increased aut-mCA prevalence were: (1) rs12072791-G (OR: 1.1, 95% CI 1.05-1.12, $P = 4\times10^{-8}$, 1KG-EUR-like AF = 0.11, 1KG-AFR-like AF = 0.25) in *CHI3L2* on chr1 and (2) rs3134968-G (OR: 1.1, 95% CI 1.05-1.10, $P = 2\times10^{-9}$, 1KG-EUR-like AF = 0.79, 1KG-AFR-like AF = 0.77) on chr6 in *HLA*. These five significant loci did not exhibit a strong association with any particular aut-mCA, but rather were moderately associated with increased risk of multiple aut-mCAs (**Figure 2C**).

Genetically predicted leukocyte telomere length and autosomal mCAs

Given that variants in *TERT* alter aut-mCA and CLL risk, we hypothesized that genetically predicted leukocyte telomere length (gLTL) may alter risk of aut-mCAs (**Supplementary Table 7**). The prevalence of aut-mCAs varied by gLTL stratification across age groups. In all age groups, individuals with the highest gLTL (top 20%) had the highest aut-mCA prevalence (**Extended Data Fig.8**). The difference in aut-mCA prevalence between gLTL groups was more pronounced among aut-mCAs considered high-risk for CLL[19,22] (**Supplementary**

Table 8). In a meta-analysis of the four cohorts, eight specific types of aut-mCAs were positively associated with longer gLTL, such as chr9q+ ($\beta = 19.18$, 95% CI 12.71-25.66, $P = 6.40 \times 10^{-9}$), chr12= ($\beta = 18.55$, 95% CI 9.73-27.38, $P = 3.77 \times 10^{-5}$), and chr15+ ($\beta = 16.90$, 95% CI 9.03-24.76, $P = 2.55 \times 10^{-5}$) after correction for multiple hypothesis testing **(Supplementary Table 9)**. No aut-mCAs were significantly associated with shorter gLTL.

Proteomic associations with autosomal mCAs

In order to identify actionable therapeutic targets, we sought to identify the proteomic signature of aut-mCAs. Using the UK Biobank plasma proteomics data for 52,705 participants, we examined the associations between ~1500 proteins and the presence of an aut-mCA and found a number of putative targets with FDA-approved drugs, such as PARP1, NPM1, and TNFRSF13B (TACI) (**Supplementary Table 10**). Aut-mCAs ($\beta = 0.25$, SE = 0.01, $P = 2\times10^{-73}$) and high-risk CLL-associated aut-mCAs ($\beta = 1.37$, SE = 0.04, $P = 9\times10^{-307}$) were associated with increased Fc receptor-like 2 (FCRL2) levels (**Figure 3A; Figure 3B**). Several proteins from the tumor necrosis factor receptor superfamily (TNFRSF) also showed strong associations with aut-mCAs and high-risk CLL-associated aut-mCAs, including TNFRSF9 ($\beta = 0.20$-0.98, $P < 1\times10^{-54}$), TNFRSF13B ($\beta = 0.11$-0.47, $P < 6\times10^{-29}$), and TNFRSF13C ($\beta = 0.13$-0.83, $P < 1\times10^{-26}$). Higher levels of PARP1, TCL1A, and NPM1 were also associated with aut-mCAs, lymphoid aut-mCAs, and high-risk CLL-associated aut-mCAs, with a stronger association for high-risk CLL-associated aut-mCAs. We then tested the association between proteins and cell fraction of individuals with mCAs, indicating the clonal expansion (**Supplementary Table 11**). FCER2, which is considered a positive prognostic marker for CLL[25], exhibited the strongest association with the cell fraction of aut-mCAs ($\beta = 2.34$, SE = 0.17, $P = 3\times10^{-40}$) and lymphoid mCAs ($\beta = 4.62$, SE = 0.36, $P = 8\times10^{-32}$). Similarly, PARP1, TCL1A, and NPM1 were also associated with clonal fraction of aut-mCAs, lymphoid aut-mCAs, and high-risk CLL-associated aut-mCAs. Proteins associated with presence of high-risk

CLL-associated aut-mCAs were enriched for GO biological processes involving activation of T cells (N = 8 proteins, $P = 6.4\times10^{-7}$) and B cells (N = 7 proteins, $P = 8.5\times10^{-7}$).

Risk of incident blood cell abnormalities of autosomal mosaic chromosomal alterations

Prior work has found that individuals with mCAs are more likely to have abnormal blood counts; however, the future risk of developing blood count abnormalities conferred by an mCA that is necessary for clinical counseling of patients with mCAs and for the design of clinical trials, is presently unknown. To quantify the risk of incident blood count abnormalities, we constructed a matched case-control cohort of 3,223 people with unique longitudinal blood counts data in UKB and AoU (**Supplementary Table 12**). The median time at risk was 1.6 years in those with aut-mCAs and 2.8 years in those without mCA. Incident persistent cytoses were two-fold more likely in those with aut-mCAs than those without. Cytoses occurred in 99/837 (11.2%) participants with aut-mCAs and 143/2,386 (5.9%) people without aut-mCAs. The major cause of cytosis was lymphocytosis, with 85 (10.2%) in mCA cases and 113 (4.7%) in controls. Those with aut-mCAs had significantly increased risk of incident cytosis versus matched controls (Hazard Ratio [HR]: 2.3, 95% CI 1.8-2.9, $P = 1\times10^{-11}$) (**Figure 4A**) (**Supplementary Table 13**). In contrast, the risk of incident cytopenia was not significantly different between those with and without aut-mCAs (HR: 1.1, 95% CI 0.96-1.30, $P = 0.16$) (**Supplementary Table 14**).

Risk factors for incident persistent cytoses included multiple aut-mCAs, CF ≥ 0.10, high-risk CLL-associated aut-mCAs, and red cell distribution width ≥ 15 femtoliters (**Figure 4B)**. The presence of three or more of the aforementioned risk factors greatly increased risk of cytosis. In those in UKB with aut-mCAs and three or more risk factors, the cumulative incidence of cytosis over 2 years was 50% (37% - 66%) compared to 3% (1% - 8%) for those without any high risk features. Similarly, those with aut-mCAs and three or more risk factors in AoU, had a cumulative incidence of cytosis of 47% (44% - 51%) over 2 years (**Supplementary Table 15**).

Furthermore, specific aut-mCAs predisposed individuals to incident cytosis and/or cytopenias (**Figure 4C**). Persistent anemia and leukocytosis were the most common blood abnormalities. Certain high-risk CLL-associated aut-mCAs such as chr11q–, chr12+, chr13q-, and chr13q= significantly increased risk of incident leukocytosis. Moreover, chr17q– increased risk of leukocytosis and anemia and chr6q– suggestively increased risk of cytopenia and decreased risk of cytosis. Myeloid-associated mCAs tended to increase risk of isolated persistent anemia without cytosis, and lymphoid associated mCAs tended to increase risk of isolated leukocytosis.

Phenotypic consequences of mosaic chromosomal alterations

We sought to systematically profile risk of disease conferred by aut-mCAs. Time-to-event phenome-wide association studies (PheWAS) across BioVU, UKB, and AoU identified significant associations between mCAs and 132 distinct Phecodes spanning multiple disease categories. The majority of significant associations clustered in neoplasm systems with the most pronounced associations being observed with leukemias, particularly CLL and acute myeloid leukemia (**Figure 4D**). Among 20,640 individuals with mCAs and no CLL diagnosis at baseline, we observed 445 incident CLL events during follow-up over 15 years. The presence of any mCA was significantly associated with incident CLL (HR: 22.4, 95% CI 12-41, $P = 7\times10^{-23}$) after adjustment **(Extended Data Fig.9)**. The risk of CLL development varied by aut-mCA characteristics, with high clonal fraction (≥10%) demonstrating substantially higher CLL risk (HR: 51, 95% CI 27-97, $P = 1\times10^{-33}$), compared to those with lower clonal fraction (HR: 7, 95% CI 3-16, $P = 6\times10^{-6}$). Moreover, individuals carrying high-risk, CLL-associated aut-mCAs showed a 120-fold increased CLL risk compared to those without an aut-mCA (95% CI 49-295, $P = 1\times10^{-25}$).

In PheWASes of specific aut-mCA types, we not only confirmed previously defined high CLL-risk mCAs, such as chr13q=, chr13q-, and chr12+, but also identified specific mCA types that were strongly associated with

cardiovascular, cerebrovascular, and kidney disease, along with infections. For example, chr13q= increased the risk of arterial stenosis by 109-fold (95% CI: 55–126, $P = 6\times10^{-42}$) and was significantly associated with sepsis (HR: 8.2, 95% CI: 4.6-14.4, $P = 4\times10^{-12}$) (**Supplementary Table 16**). Additionally, chr9p=, which co-occurs with *JAK2* CHIP, was associated with a six-fold increased risk of stroke (95% CI: 3.2–11.3, $P = 2\times10^{-8}$). Chr13q= was highly associated with a four-fold increased risk of chronic kidney disease (95% CI: 2.7–6.2, $P = 4\times10^{-11}$). Based on these PheWAS results, we defined chr4q=, chr9p=, chr11q=, chr13q=, chr13q-, chr20q-, and chr22q= as high cardiovascular-risk mCAs; chr9p= as a high stroke-risk mCA; chr13q=, chr11q=, chr1q=, chr4q=, chr6p=, chr9p=, and chr9q= as high kidney disease-risk mCAs; and chr4p=, chr9p=, chr11q=, chr13q=, chr13q-, chr15+ and chr15= as high infection-risk mCAs (**Supplementary Table 16**).

Autosomal mCAs partially mediate and enhance germline polygenic CLL risk

Since aut-mCAs are strongly associated with CLL and CLL has a high single-nucleotide-polymorphism-based heritability, we hypothesized that (1) aut-mCAs mediate the risk of CLL in those with high CLL polygenic risk and (2) risk of CLL is higher in those with aut-mCAs and a high CLL polygenic germline genetic background than in those with aut-mCAs without high germline genetic risk. We calculated polygenic risk of chronic lymphocytic leukemia (CLL-PRS) from Law et al[26] in UKB, AoU, and BioVU (**Supplementary Table 17**). Individuals with the highest 20% of CLL-PRS risk had the highest prevalence in all age groups for aut-mCAs and high-risk CLL-associated aut-mCAs (**Extended Data Fig.10**). Six specific types of aut-mCAs were associated with CLL-PRS (**Supplementary Table 8**). Aut-mCAs mediated 23.8% ($P < 2 \times 10^{-16}$) of the total CLL-PRS effect between CLL-PRS and CLL (indirect effect $\beta = 1.4\times10^{-4}$, $P < 2\times10^{-16}$; direct effect $\beta = 4.4\times10^{-4}$, $P < 2\times10^{-16}$; total effect $\beta = 5.8\times10^{-4}$, $P < 2\times10^{-16}$) **(Figure 4E)**. Notably, high-risk CLL-associated aut-mCAs mediated 43.7% ($P = 0.04$) of the total CLL-PRS effect (indirect effect $\beta = 1.9\times10^{-4}$, $P < 2\times10^{-16}$; direct effect $\beta = 1.9\times10^{-4}$, $P = 0.22$; total effect $\beta = 3.8\times10^{-4}$, $P = 0.04$) **(Figure 4F)**. These findings suggest that aut-mCAs,

particularly those highly associated with CLL, substantially mediate the relationship between germline genetic risk and subsequent CLL development.

We found that those with aut-mCAs and high (top 20%) germline polygenic risk of CLL had substantially higher risk of CLL than those with either aut-mCAs or high polygenic risk. The 133,893 participants without detectable aut-mCAs but with high polygenic risk were at modest increased risk of CLL (HR: 1.81, 95% CI = 1.03-3.18, $P = 0.04$). The 21,903 people with detectable aut-mCAs but not at high polygenic risk were at higher risk of incident CLL than those with only high polygenic risk (HR: 10.3, 95% CI = 3.0-35, $P = 2\times10^{-4}$). The 6,189 people with both detectable aut-mCAs and high polygenic risk had the highest risk of CLL (HR: 23.5, 95% CI = 5.7-97, $P = 1\times10^{-5}$). The interaction analysis indicates synergism between high polygenic risk and detectable aut-mCAs on the hazard of developing CLL (**Supplementary Table 18**).

Autosomal mCAs are in the causal pathway of CLL

We next hypothesized from our mediation analysis that mCAs may be causally related to CLL. We performed Mendelian randomization (MR) using the 5 independent loci which our GWAS of aut-mCAs described above from UKB, TOPMed, AoU, and BioVU as instrumental variables and summary statistics of chronic lymphocytic leukemia from FinnGen as outcome data.[27] Genetic predisposition to aut-mCAs was associated with increased risk of CLL in inverse-variance-weighted (IVW) MR (IVW MR= 6.37, 95% CI = 2.3-17.7; $P = 3.7\times10^{-4}$). The results remained robust after leave-one-out MR for each instrumental variable and suggested that rs7705526 in *TERT* had the most significant effect (**Extended Data Fig.11**).

Discussion

This large-scale analysis of aut-mCAs across one million diverse individuals from four large biobanks enabled us to profile the remarkable heterogeneity of aut-mCAs compared to other forms of clonal hematopoiesis such

as loss of sex chromosomes or CHIP. Certain aut-mCAs exhibit demographic specificity by ancestry and sex. Chr13q– and chr20q– predominate in 1KG-EUR-like populations, while chr15+ is more prevalent in 1KG-AFR-like populations. The male predominance of chr15+ and chr20q– is supported by mathematical modeling of clonal fitness showing higher clonal expansion rate of these aut-mCAs in males after age 40.[10] Moreover, germline variants increase risk of specific aut-mCAs. Expectedly, both common and rare variants on the chromosome affected by the aut-mCA modulate risk of mCA acquisition, particularly for copy-neutral loss of heterozygosities and losses. Our diverse study enabled discovery of additional such variants, such as a variant exclusively found in 1KG-AFR-like populations in *PIK3C3* which reduces risk of chr18p=. Moreover, longer genetically predicted telomere length increased the risk of certain aut-mCAs, which suggests that this subset of aut-mCAs may specifically impose a selection pressure for hematopoietic stem cells as they replicate. While high-risk aut-mCA types for hematologic malignancy have been reported, we are the first to identify specific aut-mCAs associated with cardiovascular, cerebrovascular, and kidney disease.

Despite the heterogeneity of aut-mCAs, our findings provide shared insights into the biology and clinical implications of clonal hematopoiesis. Most fundamentally, by exhaustively identifying putative driver genes for aut-mCAs, we showed that proto-oncogenes explained each aut-mCA gain event and tumor suppressors explained each aut-mCA loss, with support from rare variant collapsing analysis. Copy-neutral loss of heterozygosity events were explained by both proto-oncogenes and tumor suppressors, likely based on the zygosity of the driver mutation. This pattern may apply to copy-number variants in non-cancerous tissues throughout the body. Among those with co-occurring CHIP and aut-mCAs, CHIP mutation cell fractions consistently equal or exceed those of aut-mCAs on the same chromosome, suggesting that CHIP mutations typically arise first and aut-mCAs serve as secondary events enhancing clonal fitness when the two lesions are found to co-occur. Moreover, the largest genome-wide association study of aut-mCAs in aggregate demonstrates that genetic variants associated with aut-mCA prevalence broadly modulate either clonal

expansion or immunosurveillance across all aut-mCA types. Loh et al had previously identified the risk loci in *TERT* and *SP140* for detectable mCAs on any autosome.[13,14] The leading variant in *TERT* rs7705526-C, which reduces risk of aut-mCAs, was also shown to be associated with reduced risk of CHIP and CLL.[9,26,28–30] We discovered novel associations in *CHI3L2* and HLA class II genes. *CHI3L2* is accessible in lymphoid progenitor cells and may promote clonal proliferation based on its role in immune cell proliferation.[31–33] While HLA class I variants near *HLA-C* have been shown to be associated with CHIP and mosaic loss of chromosome X, no HLA class II variants have been previously associated with clonal hematopoiesis. The HLA associations cluster in the DRB1*04:01 haplotype, which is carried by ~15% of 1KG-EUR-like individuals and is known to increase risk of multiple autoimmune conditions, suggesting impaired immune surveillance is a permissive environment for clonal expansion.

Along with the aforementioned biologic insights, our study demonstrates how personalized interventions may enable prevention of progression from aut-mCAs to CLL. Mediation analysis and Mendelian randomization illustrate that aut-mCAs cause CLL, thereby creating a critical need to identify which subset of people with aut-mCAs are at highest risk of transformation. The critical step in this transformation is development of abnormal blood counts. Therefore, we show that (1) those with aut-mCAs develop abnormally high blood counts (i.e., lymphocytosis) at a rate of about 12% every 2 years and (2) among those with aut-mCAs and multiple clinical high-risk features (i.e., ≥ 1 mutation, abnormal red cell morphology, expanded clonal fraction, and CLL-associated aut-mCA), half develop cytoses within 2 years of sequencing. Another high-risk feature is germline polygenic risk, which synergistically interacts with somatic aut-mCAs, to increase CLL risk. The proteomic signature of expanded, high-risk CLL-associated aut-mCAs nominated druggable targets which may be exploited for therapies to prevent progression to hematologic malignancy in those with aut-mCAs at high risk. For example, those with high-risk CLL-associated mCAs had increased levels of NPM1 and PARP1, which are known druggable targets for leukemia (i.e., menin and PARP1 inhibitors).[34–36] High-risk CLL-associated aut-

mCAs were also associated with higher levels of TNFRSF13B (TACI), which interacts with B-cell activating factor (BAFF) and a proliferation-inducing ligand (APRIL) to increase B cell proliferation and class switching. Given the role of TACI in promoting the expansion of clonal B cells, increased circulating TACI levels may reflect the pathogenesis from aut-mCAs to monoclonal B cell lymphocytosis and then to CLL.[21] Telitacicept and atacicept both block the interaction of TACI with BAFF and APRIL and are investigative therapies targeting B cell proliferation for systemic lupus erythematosus.[37]

Limitations of this study include the single time point assessment of aut-mCAs, the heterogenous methods for mCA detection across cohorts, and use of telomere length and CLL polygenic scores derived in individuals of European ancestry. Moreover, the proteomic analysis was performed in only those with measured protein levels – which was approximately 50,000 UKB participants. Finally, incident blood count abnormality analyses could only be performed in those with longitudinal blood counts, which may overestimate the rate of incident blood count abnormalities, as these blood counts were collected in the course of routine clinical care.

In conclusion, this study provides fundamental insights into somatic structural variants which occur in the blood and likely in tissues throughout the body. Moreover, we provide evidence for an approach to incorporate germline genetics, somatic mosaicism, and clinical phenotypes to prioritize high-risk individuals who may benefit from preventative interventions. This paradigm likely applies to premalignant lesions in tissues throughout the body.

References

1. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N Engl J Med* **371**, 2488–2498 (2014).
2. Jakubek, Y. A., Reiner, A. P. & Honigberg, M. C. Risk factors for clonal hematopoiesis of indeterminate potential and mosaic chromosomal alterations. *Translational Research* **255**, 171–180 (2023).
3. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* **27**, 1012–1024 (2021).
4. Weeks, L. D. *et al.* Prediction of Risk for Myeloid Malignancy in Clonal Hematopoiesis. *NEJM Evidence* **2**, (2023).
5. Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
6. Zhao, K. *et al.* Somatic and Germline Variants and Coronary Heart Disease in a Chinese Population. *JAMA Cardiol* **9**, 233 (2024).
7. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
8. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
9. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
10. Jakubek, Y. A. *et al.* Mosaic chromosomal alterations in blood across ancestries using whole-genome sequencing. *Nat Genet* **55**, 1912–1919 (2023).
11. Vattathil, S. & Scheet, P. Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. *The American Journal of Human Genetics* **98**, 571–578 (2016).
12. Vattathil, S. & Scheet, P. Haplotype-based profiling of subtle allelic imbalance with SNP arrays.

Genome Res. **23**, 152–158 (2013).

13. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).

14. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).

15. Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130–135 (2020).

16. Liu, A. *et al.* Genetic drivers and cellular selection of female mosaic X chromosome loss. *Nature* **631**, 134–141 (2024).

17. Jakubek, Y. A. *et al.* Genomic and phenotypic correlates of mosaic loss of chromosome Y in blood. *Am J Hum Genet* S0002-9297(24)00456–7 (2025) doi:10.1016/j.ajhg.2024.12.014.

18. Niroula, A. *et al.* Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med* **27**, 1921–1927 (2021).

19. Lin, S.-H. *et al.* Incident disease associations with mosaic chromosomal alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association study in the UK Biobank. *Cell Biosci* **11**, 143 (2021).

20. Pershad, Y. *et al.* Determinants of mosaic chromosomal alteration fitness. *Nat Commun* **15**, 3800 (2024).

21. Sekar, A. *et al.* Mosaic chromosomal alterations (mCAs) in individuals with monoclonal B-cell lymphocytosis (MBL). *Blood Cancer J.* **14**, 193 (2024).

22. Kishtagari, A. *et al.* Driver mutation zygosity is a critical factor in predicting clonal hematopoiesis transformation risk. *Blood Cancer J.* **14**, 6 (2024).

23. Yang, L., Han, Y., Suarez Saiz, F. & Minden, M. D. A tumor suppressor and oncogene: the WT1 story. *Leukemia* **21**, 868–876 (2007).

24. Huijsdens–van Amsterdam, K. *et al.* Mosaic maternal 10qter deletions are associated with FRA10B

413 expansions and may cause false-positive noninvasive prenatal screening results. *Genetics in Medicine* **20**,
414 1472–1476 (2018).
415 25. Geisler, C. H. *et al.* Prognostic importance of flow cytometric immunophenotyping of 540 consecutive
416 patients with B-cell chronic lymphocytic leukemia. *Blood* **78**, 1795–1802 (1991).
417 26. Law, P. J. *et al.* Genome-wide association analysis implicates dysregulation of immunity genes in
418 chronic lymphocytic leukaemia. *Nat Commun* **8**, 14175 (2017).
419 27. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population.
420 *Nature* **613**, 508–518 (2023).
421 28. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*
422 **590**, 290–299 (2021).
423 29. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and
424 consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).
425 30. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis phenotypes.
426 *Nature* **612**, 301–309 (2022).
427 31. Comabella, M. *et al.* CSF Chitinase 3–Like 2 Is Associated With Long-term Disability Progression in
428 Patients With Progressive Multiple Sclerosis. *Neurol Neuroimmunol Neuroinflamm* **8**, e1082 (2021).
429 32. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis
430 and leukemia evolution. *Nat Genet* **48**, 1193–1203 (2016).
431 33. Liu, L. *et al.* CHI3L2 Is a Novel Prognostic Biomarker and Correlated With Immune Infiltrates in
432 Gliomas. *Front. Oncol.* **11**, 611038 (2021).
433 34. Padella, A. *et al.* Targeting PARP proteins in acute leukemia: DNA damage response inhibition and
434 therapeutic strategies. *J Hematol Oncol* **15**, 10 (2022).
435 35. Candoni, A. & Coppola, G. A 2024 Update on Menin Inhibitors. A New Class of Target Agents against
436 KMT2A-Rearranged and NPM1-Mutated Acute Myeloid Leukemia. *Hematology Reports* **16**, 244–254

(2024).

36. Uckelmann, H. J. *et al.* Therapeutic targeting of preleukemia cells in a mouse model of *NPM1* mutant acute myeloid leukemia. *Science* **367**, 586–590 (2020).

37. Fan, Y., Gao, D. & Zhang, Z. Telitacicept, a novel humanized, recombinant TACI-Fc fusion protein, for the treatment of systemic lupus erythematosus. *Drugs Today (Barc)* **58**, 23–32 (2022).

Tables

Table 1. Putative driver genes associated with autosomal mosaic chromosomal alterations.

Genes are annotated as proto-oncogenes (+), tumor suppressors (-), or both (+/-). Genes were identified from the minimum shared altered region in each chromosomal event type and filtered for those implicated in hematologic malignancy. Annotations in parentheses (p) and (q) indicate chromosome arm specificity when relevant. CN-LOH = copy-neutral loss of heterozygosity. Genes are italicized. Bold and underlined genes have rare variant support as putative drivers.

Chromosome	Gain	CN-LOH	Loss
chr1	*NOTCH2*$^{+-}$, *NRAS*$^{+}$	*NOTCH2*$^{+-}$, *NRAS*$^{+}$, *MPL*$^{+}$	*NOTCH2*$^{+-}$
chr2	*IDH1*$^{+}$, *SF3B1*$^{+}$	***XPO1***$^{+}$	*DNMT3A*$^{-}$, *ASXL2*$^{-}$
chr3	*BCL6*$^{+-}$, *FOXP1*$^{+-}$	*FOXP1*$^{+-}$, *GATA2*$^{+-}$	*SETD2*$^{-}$
chr4	*KIT*$^{+}$, *FGFR3*$^{+}$ (p)	*KIT*$^{+}$	***TET2***$^{-}$
chr5	*TERT*$^{+}$, *PDGFRB*$^{+}$, *IL7R*$^{+}$	*APC*$^{-}$, *SDHA*$^{-}$ (p)	*APC*$^{-}$
chr6	*SGK1*$^{+}$	*SGK1*$^{+}$, *DUSP22*$^{-}$(p)	*BACH2*$^{-}$, *PRDM1*$^{-}$
chr7	*EGFR*$^{+}$	*EGFR*$^{+}$, *IKZF1*, *GNA12*$^{+}$, *PSM2*$^{-}$ (p)	*CUX1*$^{-}$, *KMT2C*$^{-}$
chr8	*MYC*$^{+}$	*ATP6V1B2*$^{-}$	*NBN*$^{-}$
chr9	***JAK2***$^{+}$, *NOTCH1*$^{+-}$ (q)	***JAK2***$^{+}$, ***NOTCH1***$^{+-}$ (q)	*CDKN2A*$^{-}$, *CDKN2B*$^{-}$, *MOB3B*$^{-}$
chr10	*RET*$^{+}$	*RET*$^{+}$	*SMC3*$^{-}$
chr11	*WT1*$^{+-}$, *HRAS*$^{+}$ (p)	*WT1*$^{+-}$, *HRAS*$^{+-}$ (p), *IGF2*$^{+-}$ (p)	*WT1*$^{+-}$
chr12	***KRAS***$^{+}$, *KDM5A*$^{+}$ (p)	*KRAS*$^{+}$	*SH2B3*$^{-}$
chr13	*FLT3*$^{+}$	*FLT3*$^{+}$, *RB1*$^{-}$	*DLEU1* (q), *DLEU2* (q)
chr14	*AKT1*$^{+}$, ***DLK1***, *TCL1A*	*AKT1*$^{+}$, ***DLK1***, *TCL1A*	*BCL11B*$^{-}$, *DLK1*, ***TCL1A***
chr15	*IDH2*$^{+}$, *IGF1R*$^{+}$	*B2M*$^{-}$, *MGA*$^{-}$, *RAD51*$^{-}$	*RAD51*$^{-}$, *MGA*$^{-}$
chr16	*MAPK3*$^{+}$	*SETD6*, *CREBBP*$^{-}$ (p), *CTCF*$^{-}$ (q)	*CREBBP*$^{-}$
chr17	*SRSF2* (q), *STAT5B*$^{+}$	*NF1*$^{-}$, ***TP53***$^{-}$ (p)	***TP53***$^{-}$
chr18	*SETBP1*$^{+}$	*SETBP1*$^{+}$	*PTPN2*$^{-}$, *PI3KC3*$^{-}$
chr19	*NOTCH3*$^{+-}$, *GNA11*$^{+}$ (p), *TYK2*$^{+}$	*NOTCH3*$^{+-}$, *SMARCA4*$^{-}$, *TYK2*$^{+}$, *STK11*$^{-}$ (p), *U2AF2* (q)	*NOTCH3*$^{+-}$, *SMARCA4*$^{-}$, *STK11*$^{-}$
chr20	*SRC*$^{+}$	*ASXL1*$^{-}$, *PAK5*$^{+}$ (p)	*ASXL1*$^{-}$
chr21	*ERG*$^{+}$, *U2AF1*$^{+}$	*RUNX1*$^{-}$, *U2AF1*$^{+}$ (q)	*RUNX1*$^{-}$
chr22	***BCR***$^{+}$, *MAPK1*$^{+}$	***CHEK2***$^{-}$, *EP300*$^{-}$	***CHEK2***$^{-}$, *EP300*$^{-}$

Figures

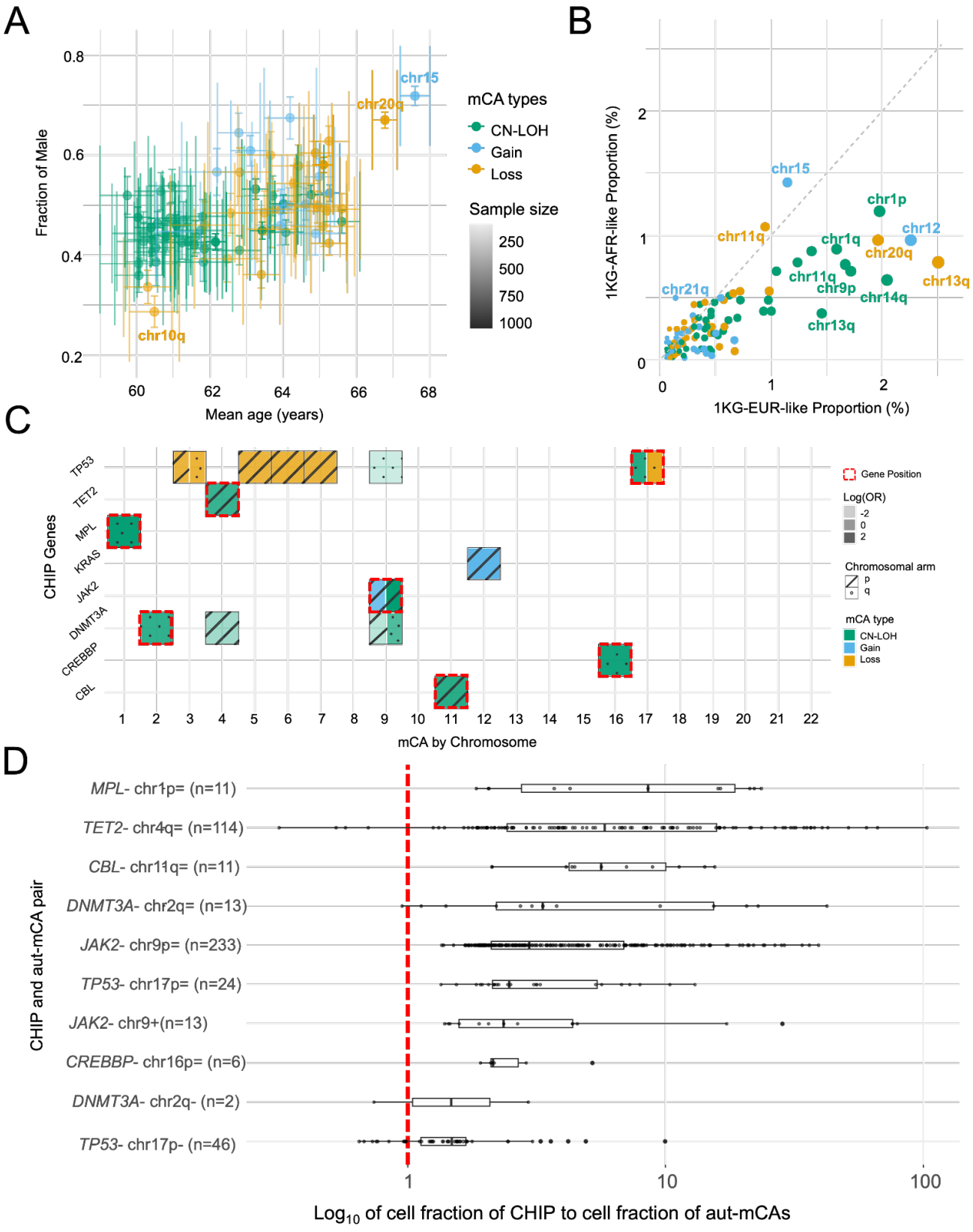


Figure 1. Demographic and genetic patterns of autosomal mosaic chromosomal alterations.

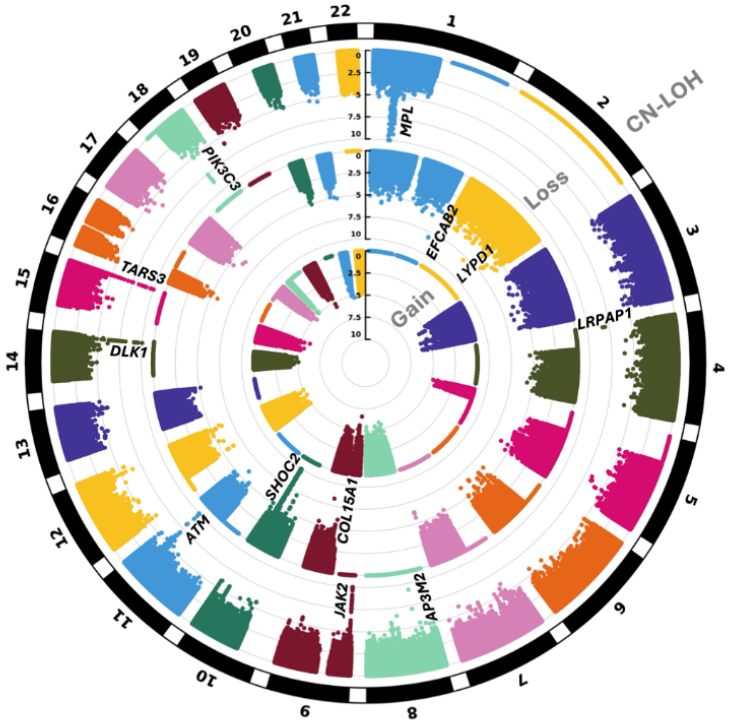
A) Age and sex distribution of aut-mCA types from 23,766 unrelated, multi-ethnic mCA cases with ages of 40-90. Each point represents a specific aut-mCA type, with x-axis showing mean age of cases and y-axis showing proportion of males. Notable findings include male predominance of chr15+ (72%) and chr20q- (67%).

B) Ancestry-specific aut-mCA frequencies comparing 1KG-EUR-like and 1KG-AFR-like populations.

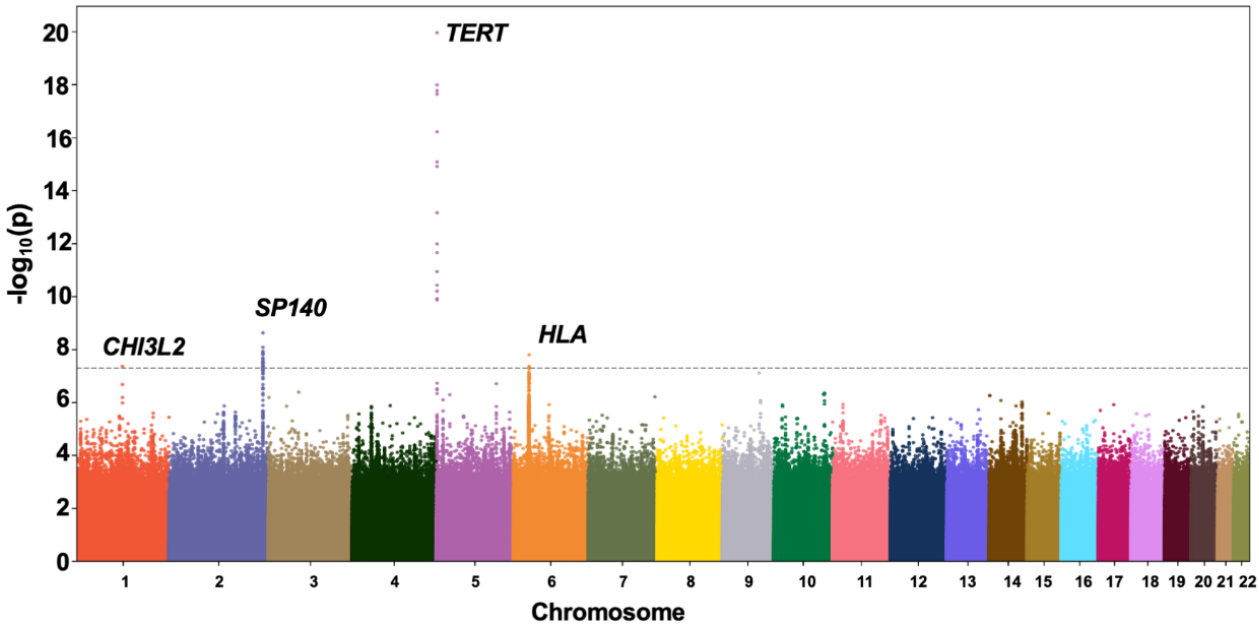
C) Co-occurrence patterns between CHIP mutations and aut-mCAs across chromosomes. Color indicates aut-mCA type (CN-LOH, gain, loss). Significant associations include *TET2*-chr4= (OR: 7.5) and *JAK2*-chr9= (OR: 36.3). The threshold of P values was Bonferroni-adjusted by <0.05/181. Only results with co-occurrence counts ≥5 were presented.

D) Comparison of cell fractions between co-occurring CHIP mutations and aut-mCAs. Box plots show $\log_{10}$ ratios of CHIP to aut-mCA cell fractions, with red line at ratio=1. CHIP mutations generally show higher cell fractions than corresponding aut-mCAs, suggesting that the CHIP mutations preceded the aut-mCAs.

A



B



C

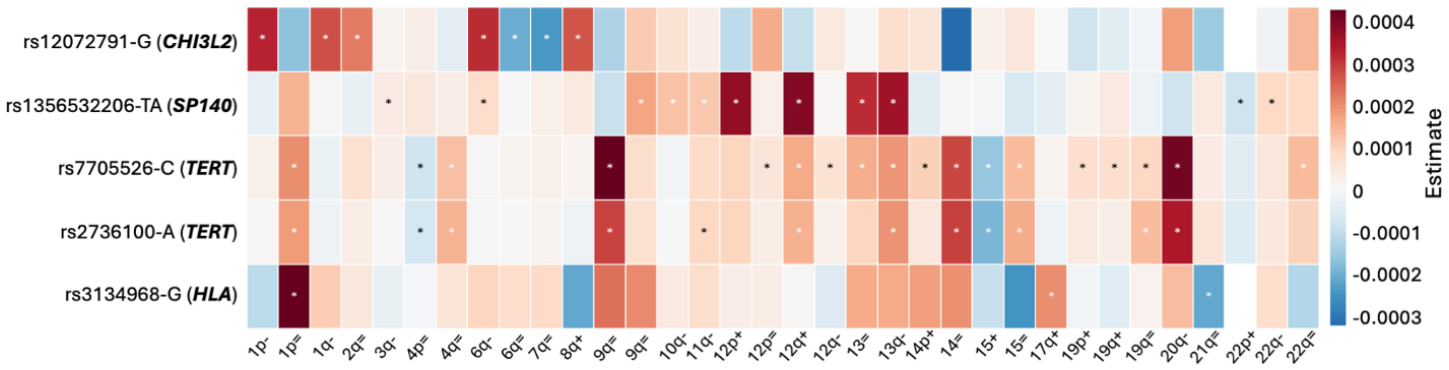


Figure 2. Genome-wide association analysis of autosomal mCA risk.

A) Circular Manhattan plot displaying the results of cis-GWAS for specific mCA types, highlighting genome-wide significant loci ($P<5\times10^{-8}$) associated with the prevalence of aut-mCAs occurring on the same chromosome and arm. The outermost ring represents copy-neutral loss of heterozygosity (CN-LOH), the middle ring represents losses, and the innermost ring represents gains. Each dot corresponds to a genetic variant, with colors indicating different chromosomes. Only aut-mCAs with 25 cases in each cohort were included, and if the chromosome was not tested, the P-values are shown as all being 0.99. P-values $<1\times10^{-10}$ labeled as 1×10^{-10}.

B) Manhattan plot showing genome-wide significant loci ($P<5\times10^{-8}$) associated with aut-mCA prevalence. Novel associations identified at the *TERT*, *CHI3L2* and *HLA* regions.

C) Heatmap presenting the associations between five significant germline genetic loci identified in trans-GWAS by specific types of aut-mCAs. The rows represent the SNPs along with their corresponding genes, while the columns indicate different mCA types. The color scale represents the effect estimates, with red indicating positive associations and blue indicating negative associations. Asterisks (*) denote statistically significant associations at a nominal P value threshold of 0.05. Note that the effect allele shown in this plot for rs7705526 in *TERT* is A rather than C.

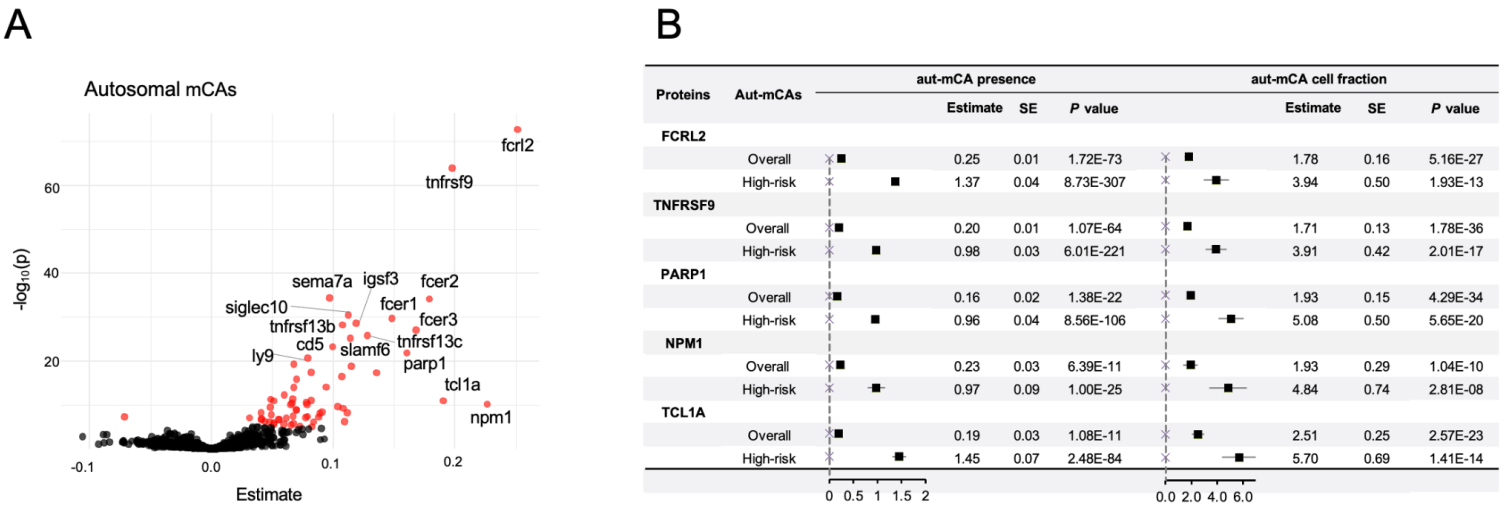
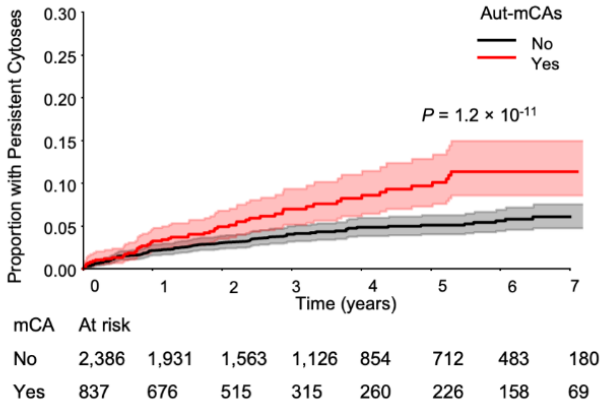


Figure 3. Proteomic signatures of autosomal mCAs

A) Volcano plots showing protein associations with aut-mCA presence. Red points indicate significant associations after Bonferroni correction ($P<3.4\times10^{-5}$).

B) Forest plot showing for the level of each protein, the estimate of the presence or cell fraction of aut-mCAs (overall and high-risk CLL-associated). All analyses adjusted for age, age squared, sex, genetic ancestry, and smoking status.

A

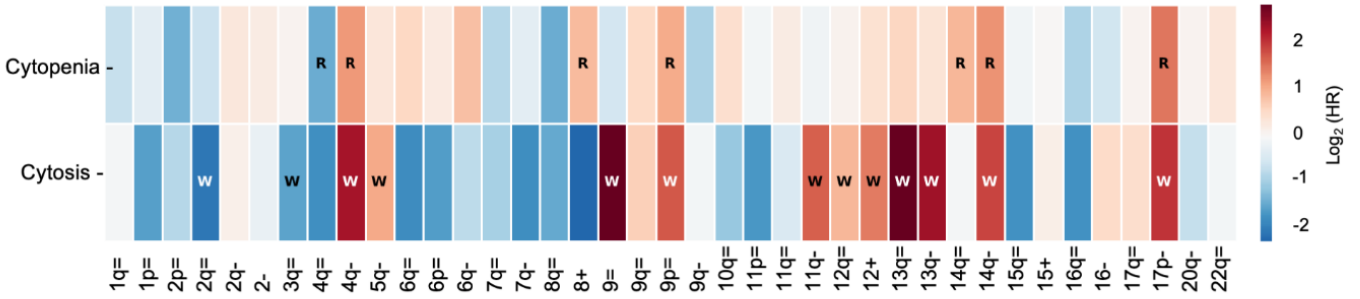


mCA	At risk							
No	2,386	1,931	1,563	1,126	854	712	483	180
Yes	837	676	515	315	260	226	158	69

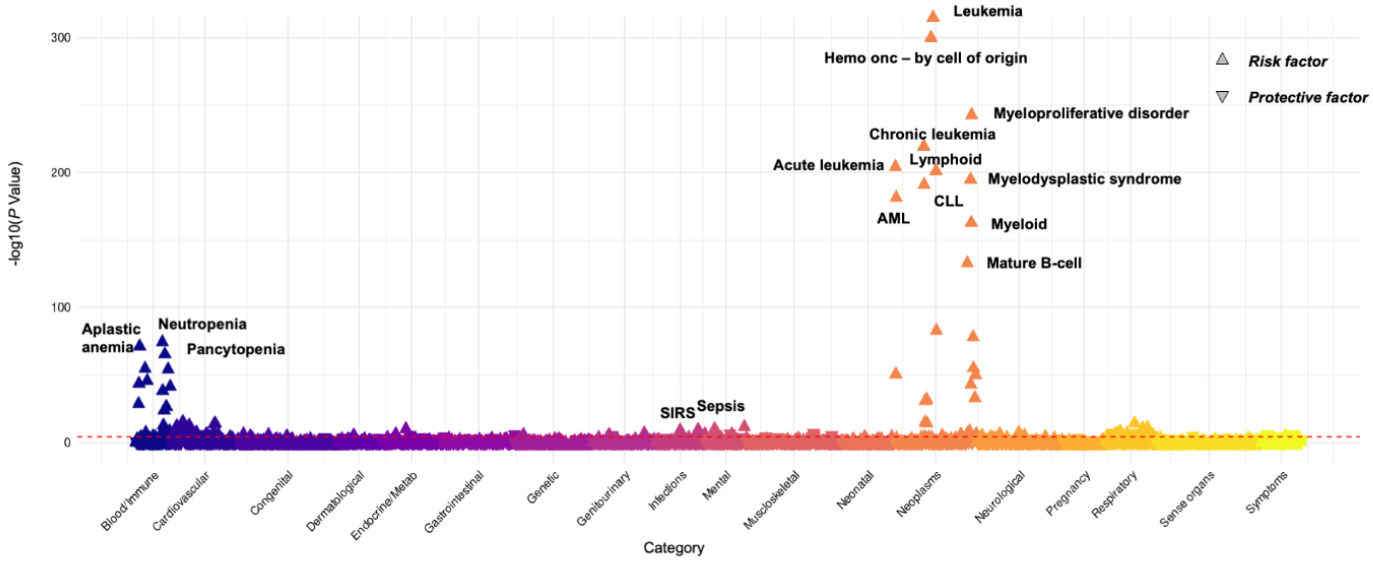
B

Variables	No.	Hazard Ratios (95% CI)	*P* value
Age ≥ 65	427	1.71 (1.18, 2.47)	4.63E-03
Male	353	0.93 (0.64, 1.34)	6.95E-01
Smoker	401	0.97 (0.67, 1.39)	8.60E-01
≥ 2 mutations	544	3.05 (1.92, 4.85)	2.56E-06
VAF ≥ 0.10	254	2.29 (1.60, 3.29)	6.14E-06
high-risk aut-mCAs	90	3.04 (2.00, 4.63)	1.92E-07
MCV ≥ 100	63	1.14 (0.60, 2.18)	6.93E-01
RDW ≥ 15	248	1.98 (1.38, 2.84)	2.30E-04

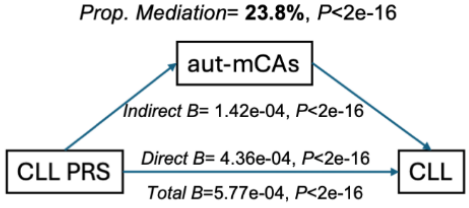
C



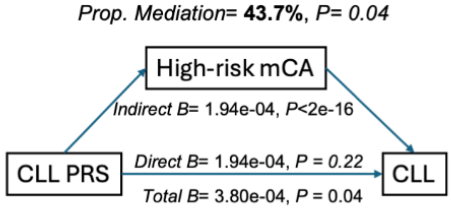
D



E



F



491

Figure 4. Clinical implications of autosomal mCAs

A) Cumulative incidence of cytosis comparing aut-mCA carriers (red) to matched controls (gray), accounting for competing risk of hematologic malignancy. Shaded areas represent 95% confidence intervals. The Fine-Gray model was used to estimate the cumulative incidence of cytosis. For those who did not develop hematologic malignancy, or persistent blood count abnormalities, the last follow-up date was defined by the latest time for CBC measurements.

B) Forest plot showing hazard ratios (HR) for cytosis risk factors in aut-mCA carriers. Key risk factors include multiple mutations (HR=3.05, $P=2.56\times10^{-6}$) and high cell fraction (HR=2.65, $P=5.10\times10^{-7}$). The reference group for each association was defined as the other group of the trait. VAF, variant allele fraction; MCV, mean corpuscular volume; RDW, red cell distribution width.

C) Heatmap showing associations between specific aut-mCA types and blood count abnormalities. Color intensity indicates effect size ($\log_2$HR). "W" indicates white blood cell, "R" red blood cell associations.

D) Manhattan-style plot showing phenome-wide associations with aut-mCAs across disease categories after meta-analysis. Each point represents a phenotype, with $-\log_{10}$(P) values on y-axis. Upward/downward triangles indicate risk/protective associations.

E) Mediation analysis showing aut-mCAs mediate 23.8% of CLL-PRS effect on CLL risk and F) high-risk CLL-associated aut-mCAs mediate 43.7% of CLL-PRS effect on CLL risk. Mediation model was used after adjusting for age, age squared, sex, ancestry, and current smoking status.

Acknowledgments

This work was supported by National Institutes of Health grant R01AG083736 (A.G.B., P.L.A., P.S.), National Institutes of Health grant R01HL117626, National Institutes of Health contract HHSN268201800002I, National Institutes of Health grant R01HL120393, National Institutes of Health grant U01HL120393, National Institutes of Health contract HHSN268201800001I, National Institutes of Health grant DP5OD029586 (A.G.B.) and National Institutes of Health grant T32GM007347 (Y.P.).

Author Contributions

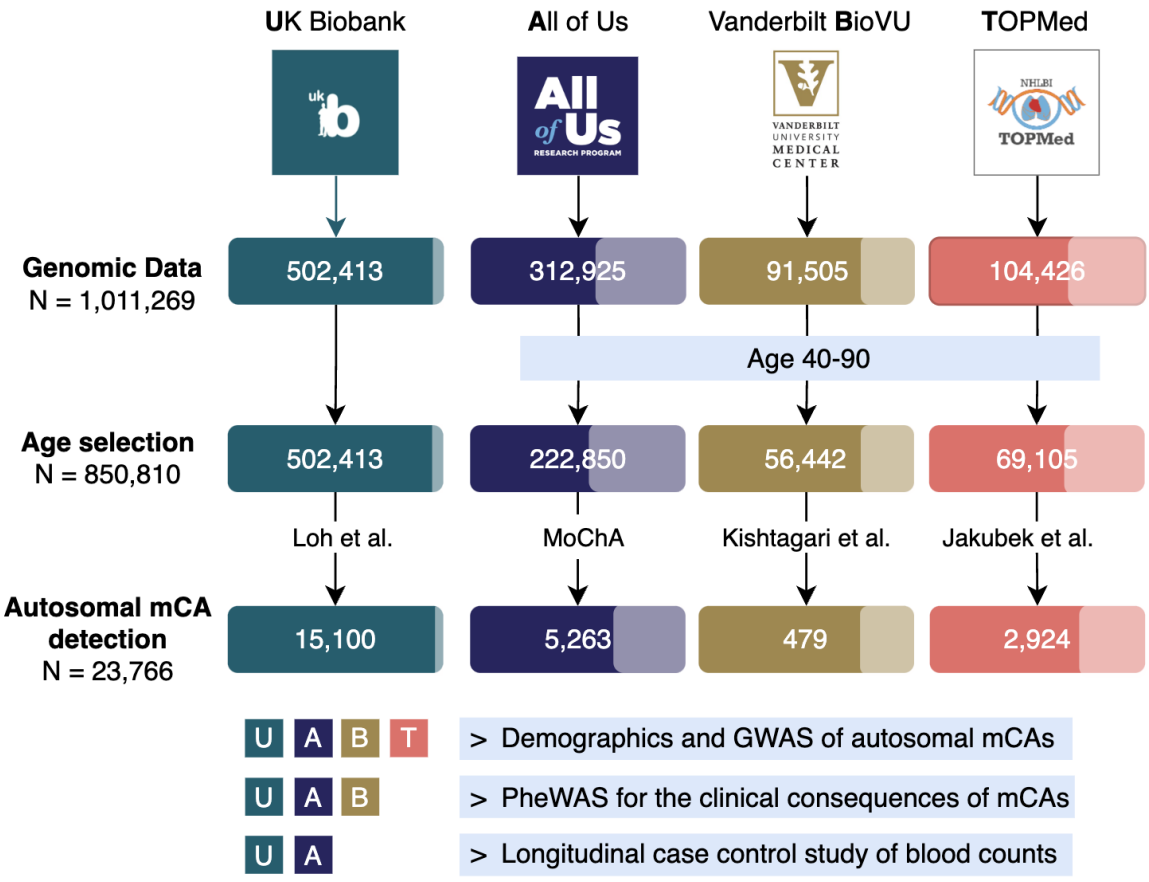
These authors contributed equally: Kun Zhao and Yash Pershad. K.Z., Y.P., A.G.B., A.P.R, P.S and P.L.A conceived of the study. X.M., K.Q, H.M.P., T.M. and R.W.C helped design statistical analysis. Y.P., C.V, K.v.B., and Y.L. performed mCAs and CHIP detection. K.Z., Y.P. and A.G.B. performed statistical analyses and wrote the manuscript. All authors reviewed and revised the manuscript. A.G.B. supervised the work.

Disclosures and Conflicts of Interest

A.G.B. has received honoraria for advisory board membership from, and holds equity in, TenSixteen Bio.

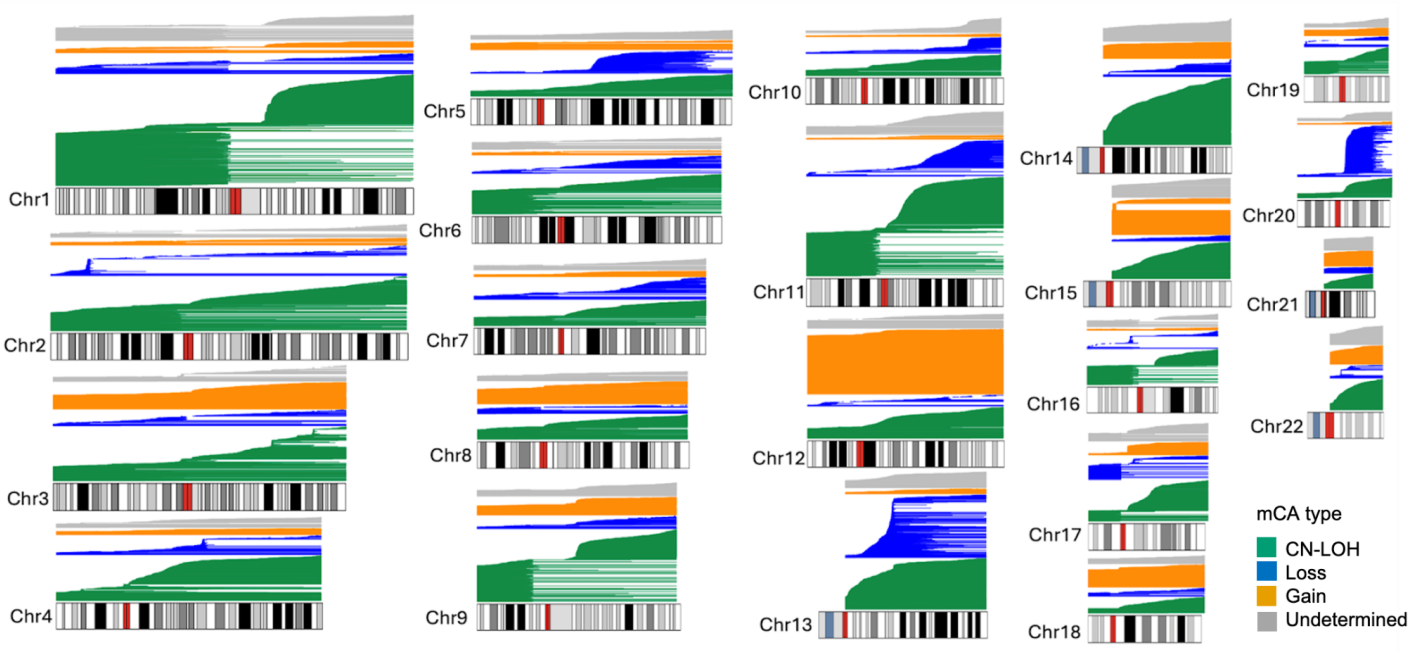
Extended Data Figures

Extended Data Fig.1 Study design and cohort selection.



Flowchart showing the selection of participants from four large biobank cohorts: UK Biobank (N=502,413), NIH All of Us (N=222,850), Vanderbilt BioVU (N=56,442), and TOPMed (N=69,105). After age selection (40-90 years), 23,766 individuals with autosomal mCAs were identified using different detection methods (Loh et al., MoChA, Kishtagari et al., and Jakubek et al.). Lighter sections indicate the proportion of non-European ancestry individuals. Bottom panels show which cohorts contributed to different analyses: demographics/GWAS (U,A,B,T), phenome-wide associations (U,A,B), and longitudinal blood count study (U,A).

538 **Extended Data Fig.2 Genomic distribution of autosomal mosaic chromosomal alterations.**

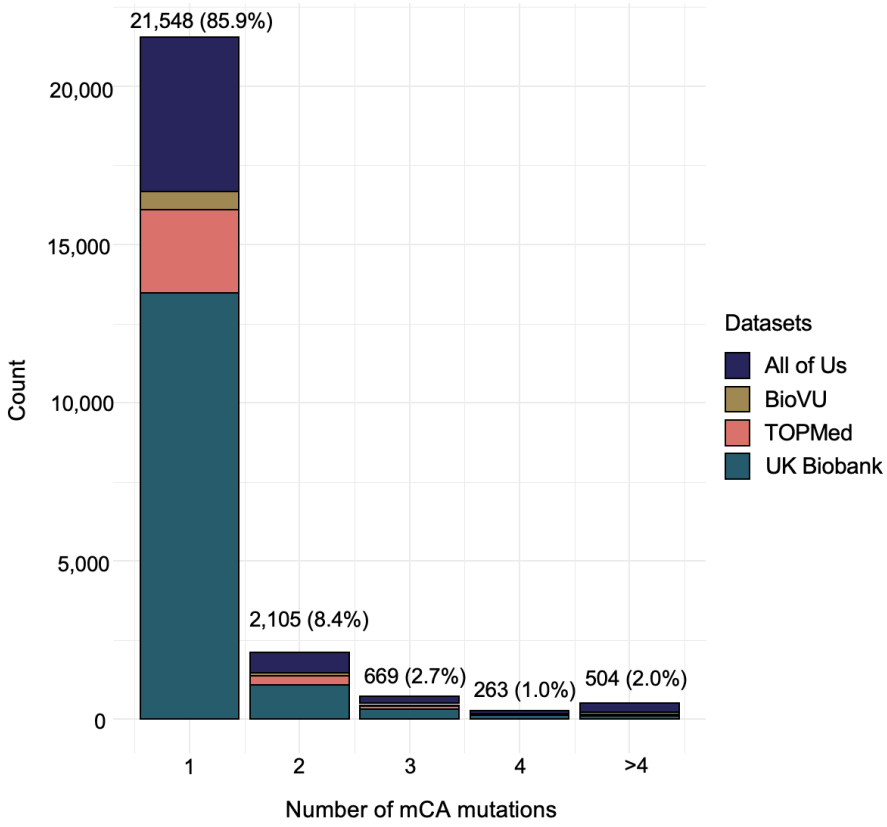


539

540 Comprehensive visualization of aut-mCAs detected across autosomes in ~1 million participants. Events are
541 color-coded by type: gains (orange), copy-neutral loss of heterozygosity/CN-LOH (green), losses (blue), and
542 undetermined copy number events (grey). Horizontal lines represent individual events, demonstrating the
543 chromosomal locations and extent of detected alterations.

544

545 **Extended Data Fig.3 Distribution of the number of mCA mutations per carrier**

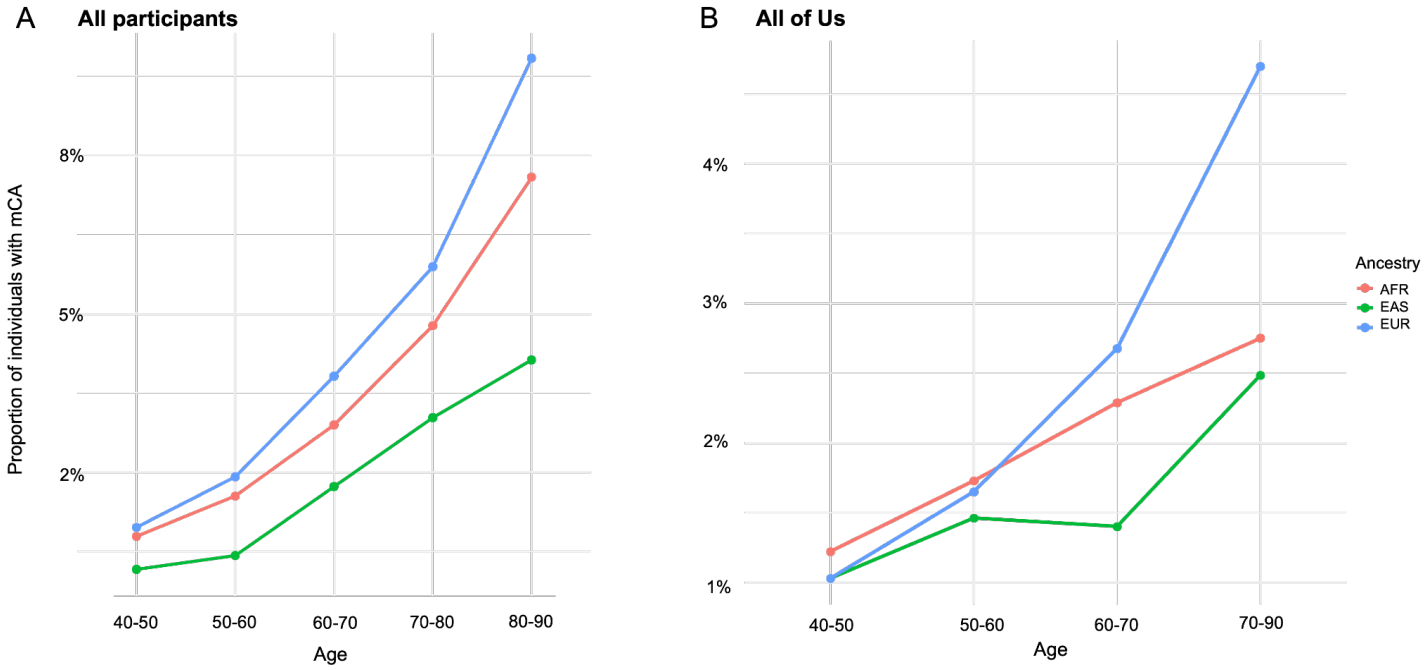


546

547 Stacked bar plot showing the number of aut-mCA mutations per carrier across four cohorts. The majority
548 (85.9%, N=21,548) carried a single mutation, while 8.4% (N=2,105) had two mutations, 2.7% (N=669) had
549 three mutations, and 3% had four or more mutations. Data is stratified by cohort (UK Biobank, TOPMed,
550 BioVU, and All of Us).

551

552 **Extended Data Fig.4 Age and ancestry-specific patterns of mCA prevalence**

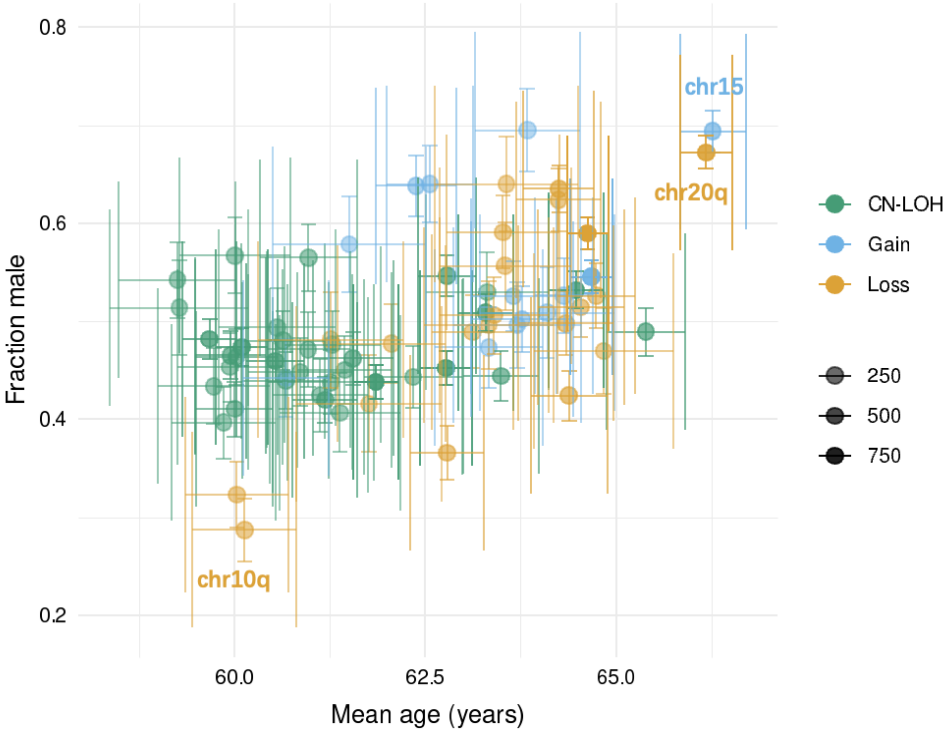


553

554 A) Overall aut-mCA prevalence across age groups (40-90 years) stratified by genetic ancestry. B) Subset
555 analysis in All of Us cohort showing similar patterns. Lines represent different ancestry groups: 1KG-AFR-like
556 (red), 1KG-EAS-like (green), and 1KG-EUR-like (blue). Demonstrates increasing prevalence with age and
557 ancestry-specific differences in accumulation rates.

558

559 **Extended Data Fig.5 Sex and age distribution of mCA types excluding chrX/Y events**

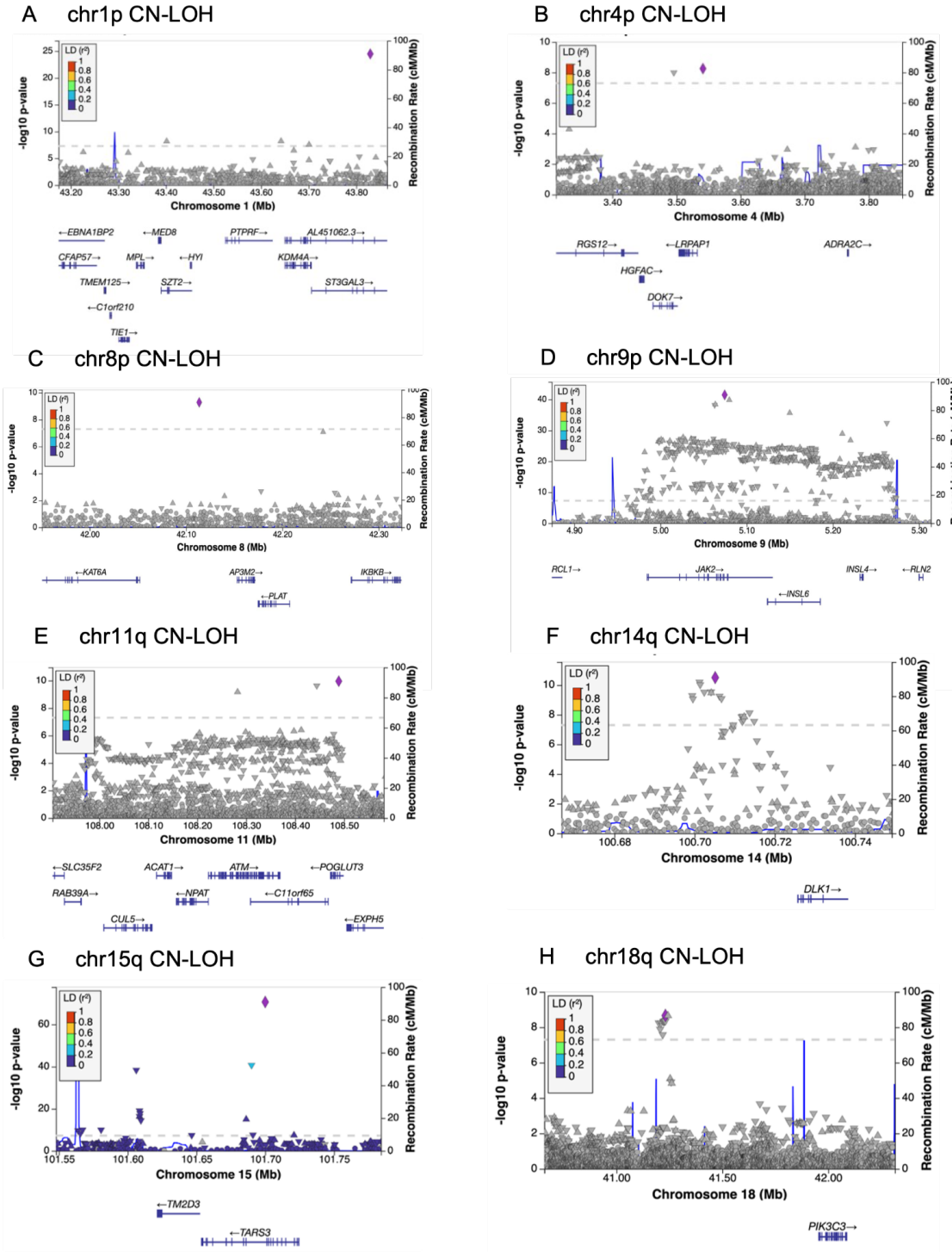


560

561 Scatter plot showing the relationship between mean age (x-axis) and proportion of males (y-axis) for different
562 mCA types after excluding individuals with mosaic loss of chromosomes X or Y. Point size indicates case
563 count, colors denote mCA type (CN-LOH, gain, loss). Demonstrates that male predominance of chr15+ and
564 chr20q- persists independently of sex chromosome mosaicism.

565

566 **Extended Data Fig.6 Regional association plots for CN-LOH events**

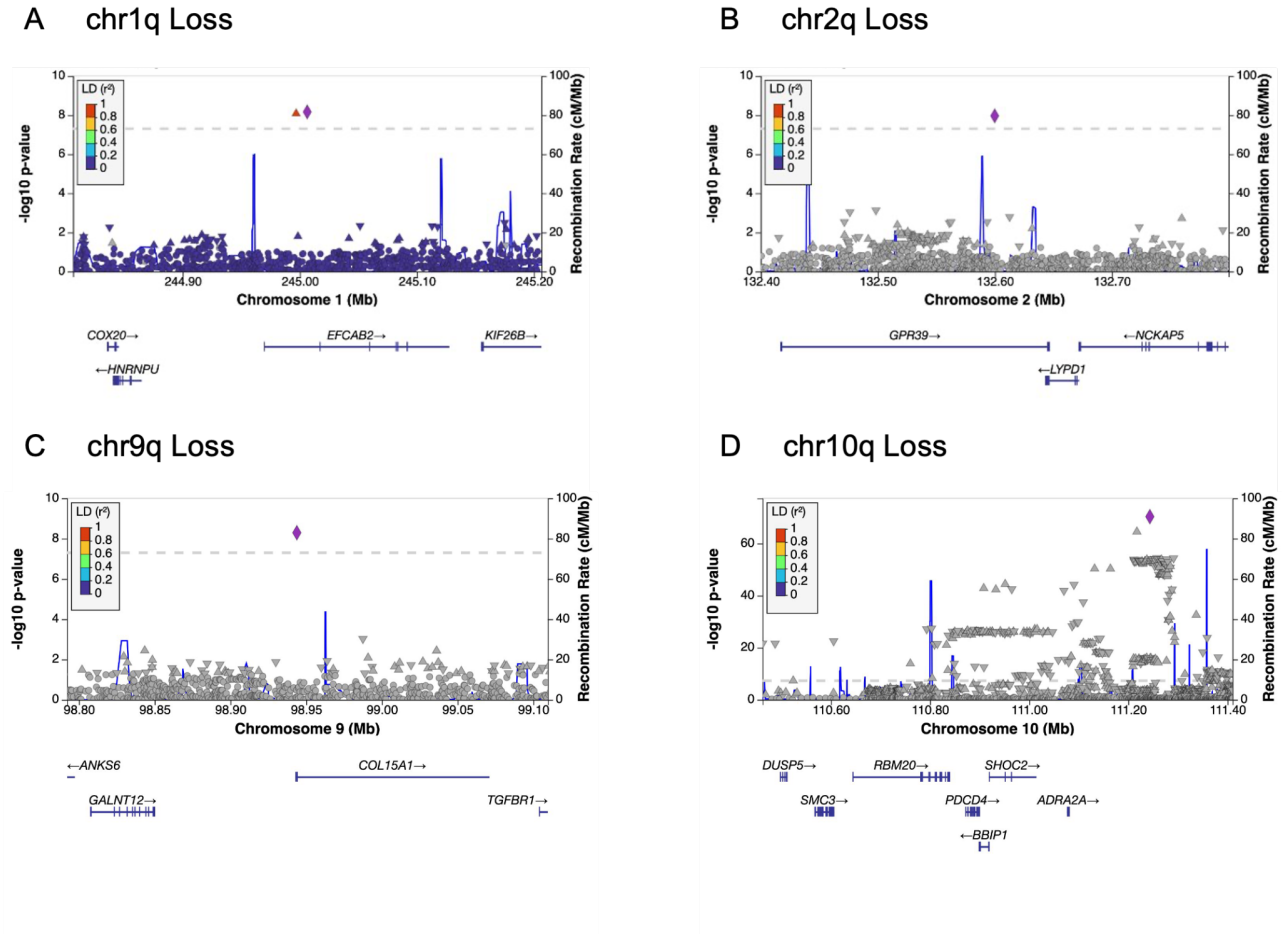


567

568 Locus-zoom plots for eight significant CN-LOH associations: A) chr1p, B) chr4p, C) chr8p, D) chr9p, E)
569 chr11q, F) chr14q, G) chr15q, and H) chr18q. Each plot shows -log10(P) values of variants (y-axis) against
570 genomic position (x-axis), with colors indicating linkage disequilibrium with lead variant. Gene annotations
571 shown below each plot.

572

573 **Extended Data Fig.7 Regional association plots for chromosomal losses**



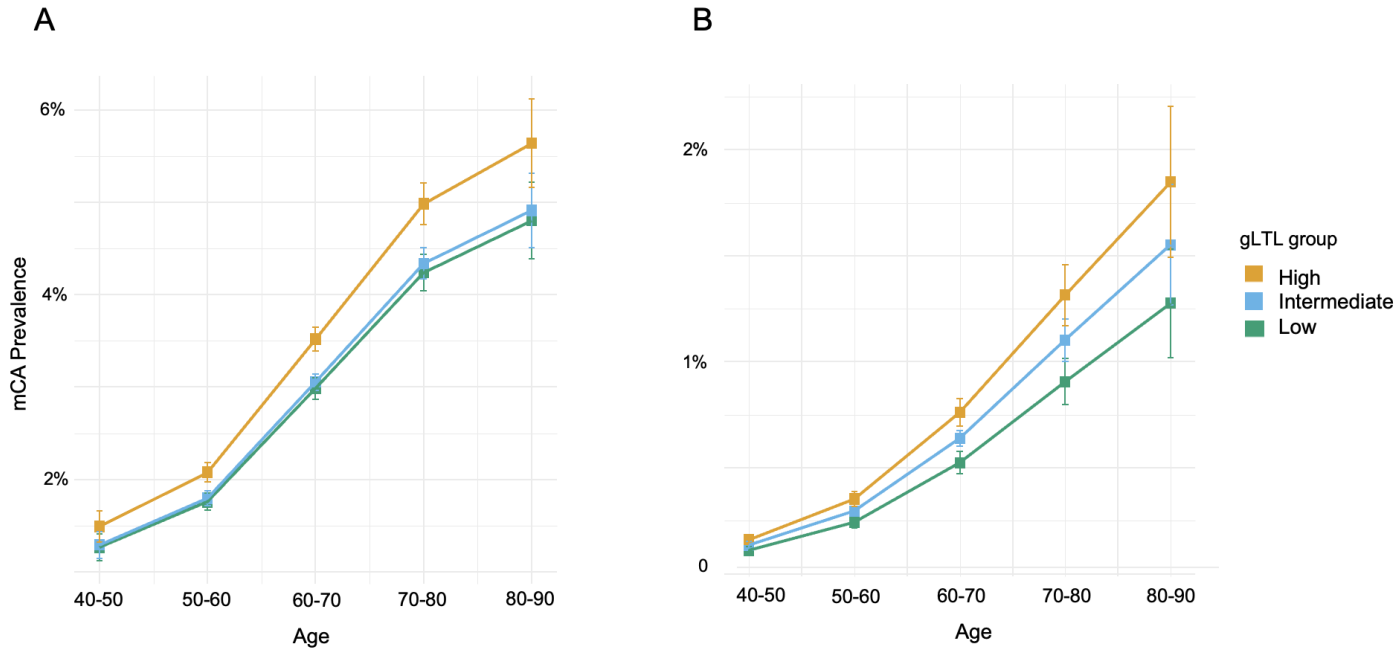
574

575 Locus-zoom plots for four significant loss associations: A) chr1q, B) chr2q, C) chr9q, and D) chr10q. Format

576 matches Extended Data Fig.6, showing regional genetic architecture around significant loss-associated loci.

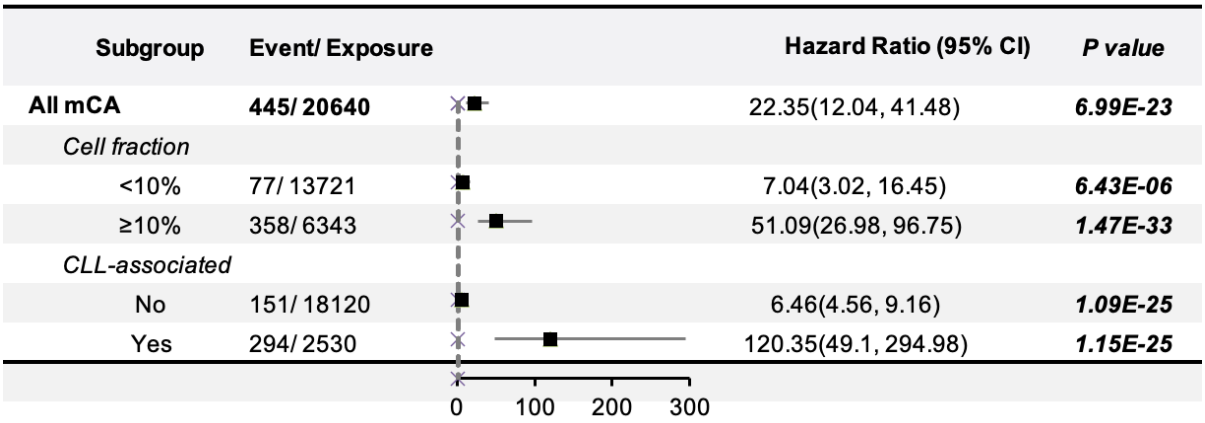
577

Extended Data Fig.8 Relationship between genetically predicted leukocyte telomere length and mCA prevalence



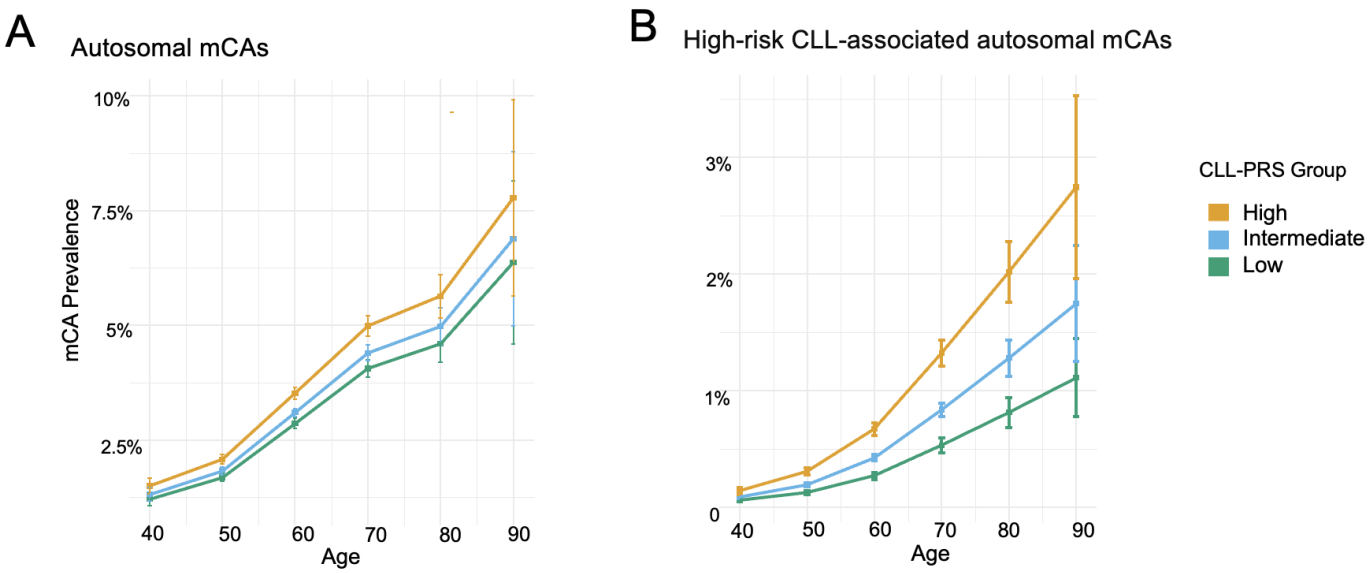
A) Overall aut-mCA prevalence and B) high-risk CLL-associated aut-mCA prevalence stratified by genetically predicted leukocyte telomere length (gLTL) groups and age. Analysis restricted to European ancestry individuals from UKB, AoU, and BioVU. Shows consistently higher mCA prevalence in individuals with high LTL-PRS (top 20%, orange) compared to intermediate (blue) and low (green) groups.

Extended Data Fig.9 Impact of mCA characteristics on CLL risk



Forest plot showing hazard ratios for CLL incidence associated with different mCA features after meta-analysis across cohorts. Demonstrates strong effects of cell fraction (≥10% HR=51.09 vs <10% HR=7.04) and CLL-risk classification (CLL-associated HR=120.35 vs non-CLL-associated HR=6.46). All analyses adjusted for age, sex, ancestry, and smoking status.

Extended Data Fig.10 Relationship between CLL polygenic risk score and mCA prevalence

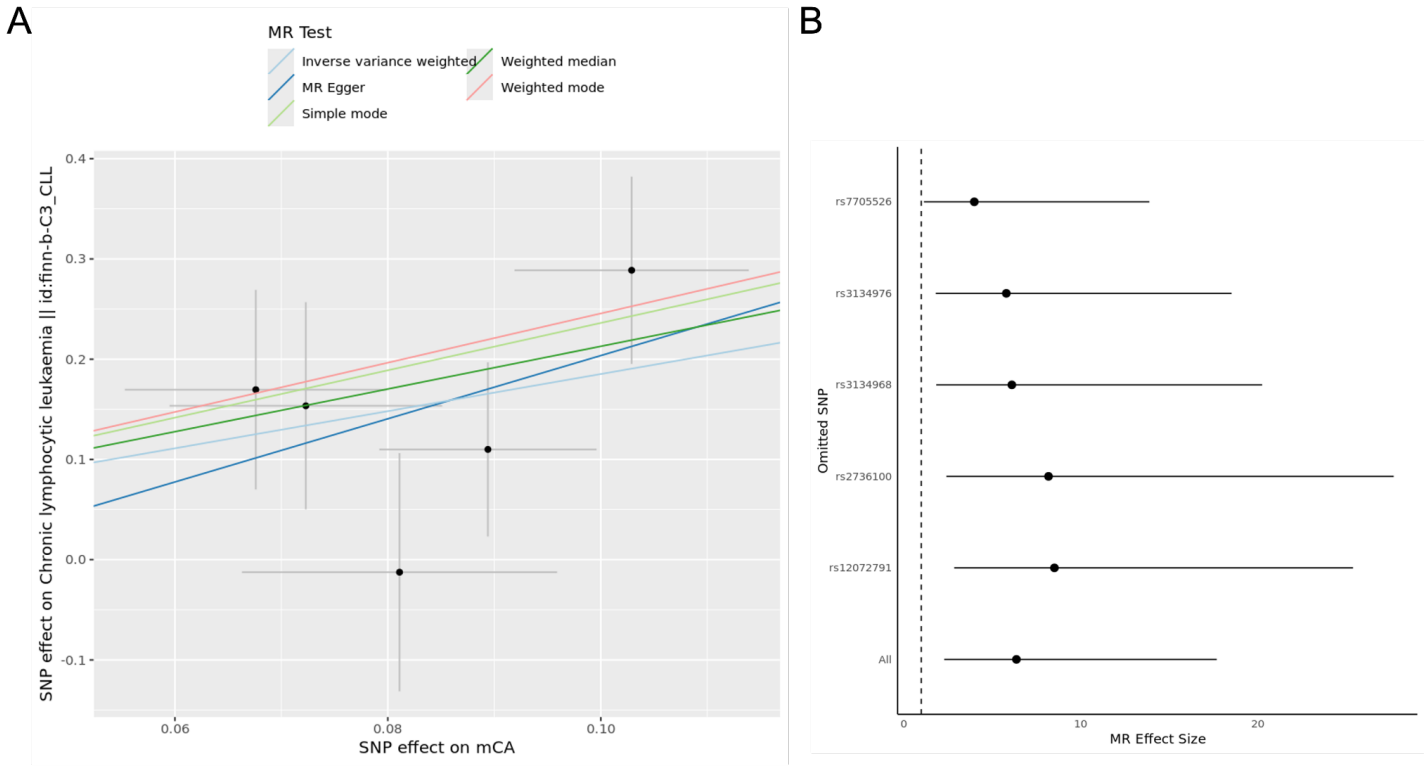


1KG-EUR-like populations from UKB, AoU, and BioVU were included in the analysis.

A) Age-specific aut-mCA prevalence stratified by CLL polygenic risk score (PRS) groups.

B) Age-specific prevalence of high-risk CLL-associated aut-mCAs by CLL-PRS groups.

598 **Extended Data Fig.11 Mendelian randomization analysis of mCA-CLL relationship**



599

600 A) Scatter plot showing different MR methods (inverse variance weighted, MR-Egger, weighted median/mode)
601 relating SNP effects on mCAs (x-axis) to effects on CLL (y-axis). B) Leave-one-out analysis testing robustness
602 of causal inference to removal of individual genetic instruments. Demonstrates consistent positive causal effect
603 estimates across methods and sensitivity analyses.

604

Methods

Study cohort

This study utilized four datasets: the NIH All of Us (AoU) Research Program, UK Biobank (UKB), NHLBI Trans-Omics for Precision Medicine (TOPMed), and Vanderbilt's BioVU. To ensure consistency across these datasets, participants aged 40–90 years were further analyzed (**Extended Data Fig.1**). Participants in AoU, BioVU, UKB, and TOPMed were assigned ancestry labels based on genetic clustering using principal components analysis. Individuals were categorized as 1KG-EUR-like, 1KG-AFR-like, or 1KG-EAS-like by comparing their genetic profiles to the 1000 Genomes (1KG) reference panel superpopulations, following the 2023 National Academies guidelines for population descriptors in genetics and genomics research.[1,2] The four cohorts are described as follows:

1) AoU aims to establish a longitudinal cohort of one million or more U.S. participants, with a particular focus on populations historically underrepresented in biomedical research. Detailed protocols for the AoU cohort and its genomic data have been described previously.[3,4] In brief, eligible participants are U.S. residents or residents of U.S. territories aged 18 and older who can provide informed consent. The genotyping array data analyzed in this study were part of the AoU version 7 release.

2) UKB is a population-based cohort of >500,000 UK adult residents recruited between 2006 and 2010 and followed prospectively via linkage to national health records.[5] UKB included adults aged 40 to 70 years at blood draw with available genotyping array data. Follow-up in the UK Biobank occurred through March 2020 for inpatient diagnosis.

3) The NHLBI TOPMed program comprises data from over 51 studies, with whole-genome sequencing (WGS) conducted on all included samples, as previously described.[6] The whole genome sequencing data analyzed in this study were part of the TOPMed Freeze 8 data release. Each sample was assigned to one of seven super-populations, coded numerically from 1 to 7, based on the primary contribution of global ancestry.

4) BioVU is a Vanderbilt University Medical Center biobank with linked de-identified electronic health records (EHRs), spanning 2006 to 2021.[7] The Vanderbilt University Medical Center's Institutional Review Board oversees BioVU and approved this project.

Autosomal mosaic chromosomal alteration detection

We identified a total of 23,766 individuals with autosomal mosaic chromosomal alterations (aut-mCAs). mCA detection in the UK Biobank was previously described.[8] In TOPMed, MoChA version 1.11, a haplotype-phasing tool for mCA detection was used for detection. MoChA utilizes genotypes from four datasets to evaluate coverage and B allele frequency (BAF) at heterozygous loci, identifying mCAs based on deviations in allelic balance. Heterozygous markers were sourced from Taliun et al.[6] and MoChA was run with the additional parameter '–LRR-weight 0.0 –bdev-LRR-BAF 6.0,' deactivating the LRR + BAF model to improve detection sensitivity in TOPMed. For BioVU, detection of mCAs was performed starting from raw IDAT files from the Illumina MEGA^{EX}, as previously described.[9] In All of Us, for the version 7 of the genotyping data, detection of mCAs was performed starting from raw IDAT files from the Illumina MEGA^{EX} with MoChA version 1.11 with the same methods as performed in BioVU. In this analysis, we defined mCA events by focusing exclusively on autosomal mCA calls and excluding any loss of X and Y chromosome events. The mCA calls included copy-neutral loss of heterozygosity (CN-LOH), losses, gains of chromosomal regions, and mCAs with undetermined copy changes.

Identification of the putative driver gene of autosomal mosaic chromosomal alterations

Using aut-mCA calls from both genotyping arrays (UKB, AoU, and BioVU) and whole-genome sequencing (TOPMed), we next identified the minimum shared altered region (MSAR) in 50%, 75%, 80%, and 90% of aut-mCAs of the same chromosomal arm and type across the 4 cohorts. We then used the UCSC Genome Browser's hg38 gene positions to identify which genes are contained within the MSAR. After this, we subsetted

the MSAR genes to those in which somatic mutations have been observed in targeted sequencing of 2,383 myeloid and lymphoid neoplasms and their matched normals via MSK-IMPACT Heme panel on the cBioPortal for Cancer Genomics.[10–12] We chose the most stringent % threshold of shared aut-mCAs that contained at least 1 hematologic malignancy driver gene for each mCA chromosome and type; the thresholds are reported in **Supplementary Table 3**. We then annotated these hematologic malignancy driver genes as tumor suppressors, proto-oncogenes, both, or neither using the OncoKB.[13,14] Putative drivers were selected among hematologic malignancy genes for aut-mCAs by selecting proto-oncogenes in the MSAR for gains and tumor suppressors in the MSAR for losses. For copy-neutral loss of heterozygosity, we did not require a gene to be a proto-oncogene or tumor suppressor.

Rare variant collapsing analyses in putative drivers of autosomal mCAs

To provide support that putative driver genes caused aut-mCAs, we conducted a rare variant collapsing analysis for rare germline missense, frameshift, deletion, and stop-gain variants within putative driver genes. We performed this analysis in 468,809 people using the UK Biobank's whole exome sequences. The omnibus test SKATO was selected for the rare variant analysis because it combines variance component tests and burden tests. Aut-mCAs in chromosome arms (p and q separately) with 25 cases per cohort were included and encoded binarily. To control for multiple comparisons, we defined a significance threshold of $P < 0.05/(\text{effective number of variants})$ for each aut-mCA type. We used a Regenie v3.3 pipeline, using a docker image provided by the software authors. We restricted step 1 to a random selection of 500,000 extremely common variants. We restricted step 2 to minor allele frequency < 0.01 in coding regions with missense, frameshift, deletion, or stop-gain mask annotations.

Clonal hematopoiesis of indeterminate potential variant calls and co-occurrence analysis

CHIP variants in UKB were filtered according to previously established criteria.[15,16] Briefly, 74 canonical CHIP

676 genes were screened for potential CHIP mutations using the *Mutect2* somatic variant caller.[15] Variants included
677 in the preliminary dataset met the following criteria: presence in a pre-established list of candidate CHIP
678 variants, total sequencing depth ≥ 20, alternate allele read depth count ≥ 5, and representation in both
679 sequencing directions (i.e., F1R2 ≥ 1 and F2R1 ≥ 1). CHIP mutations were defined as those with a variant allele
680 fraction (VAF) ≥ 0.02. The detail of CHIP calls in AoU, BioVU and TOPMed have been previously
681 described.[9,16,17] Fisher's exact test was used to calculate if CHIP and aut-mCAs occurred more frequently than
682 by chance, after correction for multiple hypothesis testing.
683
684 ***Single cis variant association studies***
685 To identify genomic regions associated with aut-mCAs, we conducted a cis-GWAS analysis (that is, variants
686 within the same genomic locus as the mCA) by filtering genomic variants based on their position relative to
687 each arm of chromosomal regions. Aut-mCAs in chromosome arms (p and q separately) with 25 cases per
688 cohort were included and encoded binarily. The number of studies supporting the association was required to be
689 in all cohorts including UKB, AoU and TOPMed. To control for multiple comparisons, we defined a
690 significance threshold of $P < 0.05/(\text{effective number of variants})$ after meta-analysis. We also performed a
691 genome-wide association study for aut-mCAs prevalence. Autosomal mCA prevalence was binarily encoded for
692 logistic regression. We declared variants from this analysis as significant if their *P* values were less than $5 \times$
693 10^{-8}. For all single-variant associations in TOPMed, single variant associations for each variant with minor
694 allele frequency (MAF) greater than 1% in individuals with a single mCA was performed with SAIGE.
695 Analysis was performed using the TOPMed Encore analysis server (https://encore.sph.umich.edu). For single-
696 variant associations in BioVU, AoU, and UKB, Regenie v3.3 was used. A random selection of 500,000 variants
697 for step one with a minor allele count >5,000 were included for step one. Variants with a MAF < 0.001 and a
698 genotyping rate < 0.1 were excluded for the second step. Samples that did not report assigned male or female at
699 birth were also excluded. Meta-analysis was performed with the METAL version from 2011-03-25 using the

700 standard-error analysis scheme.

701

702 ***Proteomics association study***

703 A total of 1,465 proteins were tested in 52,705 participants in the UK biobank. Proteomics was measured by
704 Olink (Olink Proteomics; Uppsala, Sweden) using a proximity-extension immunoassay-based method,
705 including proteins from cardiovascular, inflammation, cardiometabolic, neurology, oncology, and other panels.
706 Linear regression models were fitted with aut-mCAs (lymphoid, myeloid, and CLL-associated as described in
707 **Supplementary Table 8**) as exposures, the level of proteins as outcomes, and various covariates, including age
708 at blood draw (continuous), age squared (continuous), genetic sex (categorical), current smoking status
709 (categorical), and principal components (continuous). Linear regression models were fitted among participants
710 with aut-mCAs with clonal fraction of the aut-mCA as an exposure, the level of proteins as outcomes, and
711 various covariates, including age at blood draw (continuous), age squared (continuous), genetic sex
712 (categorical), current smoking status (categorical), and principal components (continuous). Linear regression
713 models were performed using the R function 'glm'. The Bonferroni threshold of P value was defined by
714 $0.05/1{,}465 = 3.41 \times 10^{-5}$.

715

716 ***Polygenic risk score***

717 The polygenic risk scores for CLL and LTL were calculated for the 1KG-EUR-like population in the AoU,
718 UKB and BioVU. To compute the PRS, we first identified representative SNPs from each of susceptibility loci
719 based on the most recent and comprehensive GWAS of CLL and LTL.[18,19] A total of 41 SNPs and 131 SNPs
720 were used for generating European-specific CLL-PRS and LTL-PRS, respectively **(Supplementary Table 7,**
721 **16)**. The PRSs were created by summing all variants used for PRS generation with the following formula:

722 $$PRS = \sum_{i=1}^{n} \beta i \cdot SNPi$$

Where βi represents the estimated weight (i.e., the natural logarithm of the odds ratio [OR]) of the i-th SNP, derived from the reference datasets, and *SNPi* is the genotype dose of each risk allele for that SNP. PRS values were categorized into low (<20%), intermediate (20% ≤ PRS <80%), or high (≥80%) genetic risk groups. To assess the association between PRS and mCA incidence, logistic regression models were employed to estimate ORs and 95% CIs, adjusting for age at blood draw, age squared, genetic sex, and current smoking status. The analysis was performed by each dataset and combined through inverse variance-weighted, fix-effects or random-effects meta-analysis (R package "metafor").

Incident cytosis analysis

We conducted a case-control study of participants from AoU and UKB. Individuals were eligible for the study if they had sequencing/genotyping for mCA detection and longitudinal complete blood count (CBC) data, without evidence of cytopenia or cytosis, acute myeloid leukemia (AML), myelodysplastic syndrome (MDS), myelofibrosis, or CLL prior to sequencing. Longitudinal CBC was defined as at least three CBC measurements, including one within a year of sequencing and two on or after the date of sequencing. The final CBC measurement had to occur at least 120 days after sequencing or the first CBC measurement, whichever came later. CBC measurements occurring greater than one year before sequencing were not included in the analysis. Participants with autosomal mCAs were matched 1:3 with controls on age, sex, and smoking status. The follow-up period commenced at the date of sequencing and terminated at the earliest occurrence of myelofibrosis, MDS, AML, CLL, or death.

Cytosis were defined by using a modified version of World Health Organization criteria (anemia: hemoglobin > 16.5 g/dL (females) or 18.5 g/dL (males); thrombocytopenia: platelets > 450,000 cells/mL; and leukopenia: white blood cell count < 11,000 cells/mL). Cytopenias were defined by using a modified version of World Health Organization criteria 8 (anemia: hemoglobin < 12.0 g/dL (females) or 13.0 g/dL (males);

thrombocytopenia: platelets < 150,000 cells/mL; and leukopenia: white blood cell count < 3,700 cells/mL). Cytosis and cytopenias were only deemed to be *persistent* if there were two consecutive observations of a cytosis in a single lineage at least 120 days apart without an intervening normal measurement. The date of cytosis was the first occurrence of the cytosis that persisted for at least 120 days.

The primary outcome of interest was incident cytosis any time after study enrollment. Date of birth and death, sex, race, laboratory values, self-reported smoking history, and ICD-9/ICD-10 codes were extracted. All laboratory measurement variables were harmonized to a common unit of measure and screened for outlier values. Cumulative incidence of cytosis were estimated by using the Fine-Gray model due to the potential competing risk of hematologic malignancy. For those who did not develop hematologic malignancy, or persistent blood count abnormalities, the last follow-up date was defined by the latest time for CBC measurements. Survival time was calculated by the time difference between the last follow-up date and baseline date for the blood draw.

Phenome-wide association study

AoU, UKB, and BioVU cohorts, which have available outcome data, were included in the Phenome-wide association studies (PheWAS). We determined phenome-wide clinical outcomes based on PheCodes (https://phewascatalog.org/phewas/#phe12) derived from International Classification of Diseases, Ninth or Tenth Revision (ICD-9/ICD-10) codes curated from EHRs. For each subject, the presence of any ICD codes corresponding to a PheCode inclusion criterion classified the subject as a case, while the absence of these codes classified the subject as a control. In addition to overall aut-mCAs, we leveraged the large sample size of this study to investigate specific types of aut-mCAs. We selected mCA types with a sample size of ≥200 in at least one dataset, resulting in 18 types included in the analysis. For the PheWAS analyses in UKB, AoU, and BioVU, we applied Cox proportional hazards models (using the R package “survival”) to evaluate associations between

mCAs (or specific mCA types) and the risk of various phenotypes, estimating hazard ratios (HRs) and 95% confidence intervals (CIs). The models were adjusted for age at blood draw (continuous), age squared (continuous), genetic sex (categorical), current smoking status (categorical), and principal components (continuous). PheWAS analysis was performed by each dataset and combined through inverse variance-weighted, fix-effects or random-effects meta-analysis (R package "metafor").

Mediation analysis

Mediation analyses (using the R package "mediation") were conducted to evaluate whether mCA mediated the relationship between CLL-PRS and CLL. This approach estimates the total effect of CLL-PRS on CLL and decomposes it into the direct effect (the effect of CLL-PRS on CLL independent of aut-mCAs) and the indirect effect (the portion of the effect mediated through aut-mCAs). The mediation analysis was performed using a two-stage regression process: first, modeling the association between CLL-PRS and aut-mCAs (mediator model), and second, modeling the association between aut-mCAs and CLL while adjusting for CLL-PRS (outcome model). The average causal mediation effect (ACME) and average direct effect (ADE) were estimated, with statistical significance evaluated through bootstrapping. The same covariates as in the Cox proportional hazards models were applied, including age, age squared, sex, ancestry, and smoking status, were applied in both models. The mediation analysis relies upon the assumption that there is not a confounding variable that increases both risk of aut-mCA and CLL. Given that mCAs are implicated in clonal expansion and genomic instability—key processes in CLL pathogenesis—it is plausible that aut-mCAs act as intermediaries linking genetic predisposition (via CLL-PRS) to the development of CLL. The analysis above of each dataset was combined through inverse variance-weighted, fix-effects or random-effects meta-analysis (R package "metafor").

Mendelian randomization between autosomal mCAs and CLL

We performed two-sample Mendelian randomization to assess the causal relationship between aut-mCAs and CLL. Five independent genome-wide significant SNPs associated with aut-mCAs were used as genetic instruments (**Supplemental Table 5**). The association between these instruments and CLL was obtained from FinnGen (finn-b-C3_CLL). After harmonization using TwoSampleMR (v0.6.8), one SNP was removed for being palindromic with intermediate allele frequencies. We implemented multiple MR methods: inverse variance weighted (IVW) as primary analysis and MR-Egger regression to assess pleiotropy. Sensitivity analyses included leave-one-out analysis and assessment of heterogeneity using Cochran's Q statistic. Causal estimates are reported as odds ratios with 95% confidence intervals, representing the change in CLL odds per unit increase in log-odds of aut-mCA.

Code availability

The code is publicly available and can be found at https://github.com/bicklab/mca-1m. The REGENIE software is available at https://github.com/rgcgithub/regenie. A standalone software implementation (MoChA) of the algorithm used to call mCAs is available at https://github.com/freeseek/mocha.

Data availability

Individual-level sequence data, CHIP calls and polygenic scores have been deposited with UK Biobank and are freely available to approved researchers, as done with other genetic datasets to date. The genotypes and phenotypes of UKB and AoU participants are available by application to the UKB (https://www.ukbiobank.ac.uk/register- apply/) and AoU (https://allofus.nih.gov/), respectively. Instructions for access to UK Biobank data are available at https://www.ukbiobank.ac.uk/enable-your-research. The HapMap3 reference panel was downloaded from ftp://ftp.ncbi.nlm.nih. gov/hapmap/, GnomAD v3.1 VCFs were obtained

816 from https://gnomad.broadinstitute.org/downloads, and VCFs for TOPMED Freeze 8 were obtained from

817 dbGaP as described in https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8.

818

Methods References

1. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. *Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al. Using Population Descriptors in Genetics and Genomics Research. (National Academies Press, Washington, D.C., 2023).*
3. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
4. Ramirez, A. H. *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns* **3**, 100570 (2022).
5. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
6. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
7. Pendergrass, S., Dudek, S. M., Roden, D. M., Crawford, D. C. & Ritchie, M. D. Visual integration of results from a large DNA biobank (BioVU) using synthesis-view. *Biocomput. 2011* 265–275 (2010) doi:10.1142/9789814335058_0028.
8. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
9. Kishtagari, A. *et al.* Driver mutation zygosity is a critical factor in predicting clonal hematopoiesis transformation risk. *Blood Cancer J.* **14**, 6 (2024).
10. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
11. de Bruijn, I. *et al.* Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR

Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).

12. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
13. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **2017**, (2017).
14. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. *Cancer Discov.* **14**, 49–65 (2024).
15. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
16. Vlasschaert, C. *et al.* A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic datasets. *Blood* blood.2022018825 (2023) doi:10.1182/blood.2022018825.
17. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
18. Law, P. J. *et al.* Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat. Commun.* **8**, 14175 (2017).
19. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nat. Genet.* **53**, 1425–1433 (2021).