nature portfolio

https://doi.org/10.1038/s41586-022-05448-9

Supplementary information

Common and rare variant associations with clonal haematopoiesis phenotypes

In the format provided by the
authors and unedited

nature portfolio

https://doi.org/10.1038/s41586-022-05448-9

Supplementary information

Common and rare variant associations with clonal haematopoiesis phenotypes

Supplementary Note 1: CHIP Prevalence Across Age Bins

In GHS subjects, the prevalence of *ASXL1* mutations appears to taper off at older ages, while *SF3B1* mutations become relatively more common (Extended Data Figure 1E). To evaluate the statistical significance of these changes in prevalence across age bins, we employed chi-squared statistics to compare carrier and non-carrier counts across age bins. The ratio of *ASXL1* carriers to non-carriers among individuals aged 60-80 is significantly lower than the same ratio among individuals aged 80-100 ($X^2 = 245$, $P=2.4 \times 10^{-55}$). In contrast, the ratios are no different between individuals aged 80-90 and those aged 90-100 ($X^2 = 0.874$, $P=0.35$). Similar calculations across all top-8 recurrent CHIP genes provided statistical support for continued increases in prevalence across the oldest age bins (i.e. 80-90 vs 90-100) for *DNMT3A* ($X^2 = 13.7$, $P=2.17 \times 10^{-4}$) and *TET2* ($X^2 = 20.2$, $P=6.91 \times 10^{-6}$).

Supplementary Note 2: CHIP VAF and Individuals With Multiple CHIP Mutations

Among all CHIP mutation carriers, 11,125 (40%) had at least one CHIP somatic mutation at a high variant allele fraction (VAF > 10%) in the UKB (5,981 [47%] had high VAF CHIP in GHS). Except for *JAK2*, which had a relatively high average VAF, the more common CHIP mutations generally had lower VAFs than the rarer and more canonically leukemogenic mutations within individuals with a single CHIP mutation (Figure S1A). Amongst the eight most recurrently mutated CHIP genes, average age among carriers with mutations in a single CHIP gene was lowest in *DNMT3A* carriers (59.95) and highest in *SRSF2* (62.72) and *SF3B1* carriers (63.51). Maximum VAF among all CHIP carriers increased linearly with the number of CHIP mutations identified in an individual ($\beta = 0.101$, $P < 1 \times 10^{-16}$, Figure S1C), consistent with the concept of accumulation of multiple mutations during progressive clonal expansion. Telomere length was not substantially different across CHIP mutation carriers, although consistent with recent reports[1], CHIP carriers generally had reduced telomere lengths (Figure S1D).

Supplementary Note 3: Conditional Analysis and Statistical Fine-mapping

While the overall approach we took to identify independent common variant signals was clumping and thresholding, we also applied two additional approaches to identify independent lead signals from our genome-wide association analyses of CHIP (Tables S3, S4) and *DNMT3A*-CHIP (Tables S5, S6) in UKB. The first was conditional analysis using GTCA COJO[2], which we used given the interpretability of its iterative step-wise approach. In particular, by relaxing our significance threshold to the suggestive cutoff of $p \leq 5 \times 10^{-6}$ when interpreting independent signals from such conditional analysis, we are able to more strictly adjust for common variant signals that may influence rare variant associations. For example, while conditional analysis at a significance cutoff of 5×10^{-8} identifies 29 independent signals associated with CHIP across 23 common variant loci, a significance cutoff of 1×10^{-6} identifies 53 independent signals (Table S3), which is closer to the number of signals identified by clumping and thresholding (57, Table S2) and fine-mapping (51, Table S4), and enables us to better condition for all potential common variant signals that may confound rare variant associations. Using these parameters, we see substantial overlap between independent signals (i.e. 24/29 variants with $P\_COJO \leq 5 \times 10^{-8}$ were also found within credible sets identified by FINEMAP).

We ran conditional analysis using default settings (including an MAF threshold of = 0.01) and with an LD reference derived from a subset of 10,000 unrelated EUR samples from UKB. In contrast, statistical fine-mapping models variants jointly when using LD patterns and association signals to evaluate the most probable number of causal signals (K) and the probability that any variant drives a causal signal (i.e. posterior inclusion probability, PIP). We performed statistical fine-mapping on association statistics from a combination of UKB and GHS using the FINEMAP software[3]. We used a MAF threshold of 0.01, p-value filter of 0.1, and an LD reference calculated as a weighted composite of the LD correlations independently determined from the full set of EUR samples in UKB and in GHS. Notably, this better captures the range of haplotypes present across the UKB and GHS EUR populations. Therefore, this fine-mapping approach is also complementary to the conditional analysis approach as a result of this increased haplotype representation. Interestingly, this can significantly impact which variants are prioritized. For example, at the *LY75* locus on chromosome 2, FINEMAP strongly prioritized the rs78446341-A *LY75* missense variants (Table S3, Extended Data Figure 3), whereas conditional analysis does not identify it as an independent signal (Table S4). Another notable locus is the *TERT* locus, which fine-mapping and conditional joint analysis suggest has numerous independent signals (eight and five, respectively). We identified numerous variants with fine-mapping as top credible set signals that are not found by clumping and thresholding or COJO (Figure S3A, B). Independent signals are also reported from an analysis of DNMT3A-CHIP associations results using COJO (Table S5) and FINEMAP (Table S6).

Supplementary Note 4: Potential Confounding of Rare Variant Associations

Given that our CHIP phenotypes are ascertained on the basis of somatic mutations and variants with higher frequency and/or a larger degree of clonal expansion (i.e. higher VAF) can be detected in exome-sequencing-based platforms designed for germline calling, rare variant and gene burden associations from genome-wide analysis will sometimes feature the same variants and genes through which the phenotype is defined. To avoid circularity, we excluded all such variants from our reported results. Furthermore, as other associated rare variants (i.e., those which were not used to condition the CHIP phenotypes) may themselves be somatic variants, we assessed whether significantly associated rare variants had lower (i.e. than expected for germline) variant allele fractions (VAFs) across carriers, as well as whether age-at-sample-collection was associated with variant carrier status (as both these tests suggest a somatic origin for a variant). For genome-wide significant rare variant and gene burden associations for which we had exome data, we reported these VAF and age-association results along with genetic association results in order to provide resolution as to whether such associations are likely driven by germline or somatic variation. To control for the possibility that our somatic CHIP calls may contain some germline variants that could then be linked to variants that confound burden mask analyses, we reran common and rare variant analyses for our CHIP_inclusive and *DNMT3A* phenotypes after conditioning on all independent common variant signals. These signals were chosen by taking the union of variants that a) were significant at COJO adjusted $p \leq 5 \times 10^{-6}$ (p-value chosen to be strict in conditioning out common variant signals) or b) had one of the highest two posterior inclusion probabilities (ie. PIPs) in any credible set identified by FINEMAP. In total, we adjusted by 144 common variants in our CHIP_inclusive association analyses and 130 common variants in our *DNMT3A*-CHIP association analyses.

When rare variant associations were on the same chromosome as a CHIP gene, we also evaluated whether the carriers of significantly associated rare variants shared specific somatic mutation calls (which might signal germline variant contamination that is leading to linkage-based confounding). This was the case with a number of rare variants on chromosome 2, which seemed to be driven by a *DNMT3A* mutation call in our somatic callset that is likely to be a germline variant (rs139053291-A). While we do not report these rare variant associations, we ultimately kept this variant in our callset as it was a previously reported CHIP mutation[4]. However, this variant is likely to be a false positive (at least in a substantial number of carriers). We identified a similar situation with a set of rare variant associations on chromosome 17 (which we similarly do not report), which were driven by linkage with a likely germline variant in *TP53* (rs587781371-T)[5] . These variants warrant reevaluation by others when calling CHIP and/or when working with our CHIP callset.

Supplementary Note 5: Associations in Individuals of Non-European Ancestral Background

Even though our power to detect associations in non-EUR populations was limited, we ran genome-wide association analyses for African ancestry individuals, South Asian ancestry individuals, and East Asian ancestry individuals. Although no associations reached genome-wide significance, 17 of 57 variants identified in the EUR analyses showed directionally consistent effects across all populations (Table S9; 7/57 expected). The recently described CHIP-risk-increasing association reported for the African ancestry-specific *TET2* enhancer variant (rs144418061-A, reported OR = 2.4, P = 4.0 x 10^{-9})[4] was not significantly associated in African ancestry (OR = 1.37 [0.87-2.18], P = 0.176, AAF = 0.036) nor European ancestry (OR = 1.02, [0.45-2.31], P = 0.97, AAF = 0.000014) subjects despite us having sufficient power (> 0.8) at our sample sizes to detect an effect size (i.e. OR) of 1.55 in African ancestry individuals and 2.1 in European ancestry individuals. Given that this association trends positively in our analysis among African ancestry individuals, and that we find significant common variant associations at the *TET2* locus, this lack of replication likely reflects a degree of "winner's curse" in the original study and insufficient power in our analysis to detect a more moderate true effect size.

Supplementary Note 6: Patterns of Effects Across CHIP Subtypes

Beyond the *TCL1A* locus, the gene-specific CHIP analyses identified additional shared and opposed patterns across CHIP subtypes (Figure 2A, Table S20). The *THRB* locus was significantly associated with *TET2*-CHIP (OR = 1.16 [1.10-1.21], P = 2.55 x 10^{-9}, Extended Data Figure 5), and replicated in GHS (OR = 1.22 [1.14-1.31], P = 2.66 x 10^{-8}), but not associated with our overall CHIP phenotype (Figure 1). The *SEPT3* locus was associated with *TP53*-CHIP (OR = 4.42 [2.61-7.50], P = 3.29 x 10^{-8}) but not our overall CHIP phenotype and did not replicate in GHS (OR = 1.17 [0.34-3.97], P = 0.80). Signals at the *SMC4* locus on chromosome 3 were strongly consistent across all CHIP subtypes (Figure 2A, green box). Variants at the *CD164* locus are significantly associated with *DNMT3A*-CHIP and *ASXL1*-CHIP (and show non-significant but trending associations with *SRSF2*-CHIP) but are not associated with *TET2*-CHIP. Consistent with the overall patterns of shared germline contribution across CHIP subtypes, a polygenic risk score (PRS) generated with estimates from our *DNMT3A*-CHIP GWAS was significantly associated ($P \leq 0.007$, i.e. 0.05/7) with being a carrier of *TET2*-CHIP (OR = 1.16 [1.13-1.20], P = 2.47 x 10^{-23}), *ASXL1*-CHIP (OR = 1.21 [1.16-1.27], P = 7.29 x 10^{-19}), *JAK2*-CHIP (OR = 1.53 [1.35-1.74], P = 8.0 x 10^{-11}), and *SRSF2*-CHIP (OR = 1.30 [1.16-1.45], P = 6.96 x 10^{-6}) but not with other CHIP subtypes. As reported by others[6], the *JAK2* 46/1 locus was strongly associated with an increased risk of *JAK2*-CHIP (OR = 2.24 [1.86-2.69], P = 9.22 x 10^{-18}). Rare variant results for *DNMT3A*-CHIP were similar to those from CHIP overall (Table S12,13). We identified two loss of function gene burden associations with an increased risk of *TET2*-CHIP at the *ZNF318* (OR = 5.83 [2.98-11.4], P = 2.48 x 10^{-7}) and *RPS6KA2* (OR = 18.2 [6.09-54.3], P = 2.05 x 10^{-7}) genes, with the former likely to be driven by somatic variation and the latter likely to be driven by germline variation (Table S14). The latter is additionally interesting given the significant common variant association signal we identified at the *ZNF318* locus in our GWAS of *DNMT3A*-CHIP (Table S11). We also found an association via rare variant burden testing between loss of function variants (AAF <= 1 x 10^{-5}) in the *NFE2* gene and *JAK2*-CHIP (OR = 163 [27-991], P = 3.09 x 10^{-8}, Table S19), which replicated in GHS (OR = 49 [4.75-499], P = 0.001), and provides support for *NFE2* gene function loss as a driver of clonal hematopoiesis.

Supplementary Note 7: ExWAS Analysis of Mosaic Chromosomal Alteration (mCA) Phenotypes

We performed rare variant and gene burden associations analyses (Tables S22-S27) for the mLOY, mLOX, and autosomal mCA phenotypes we generated, which exclude samples with CHIP mutations and should therefore be mCA specific (see methods). Notably, we found a novel risk reducing association between a rare missense variant in the *KNTC1* gene (rs61751321-T, AAF = 0.003, L317F) and the mLOY phenotype (OR = 0.60 [0.50-0.72], P = 2.56 x 10^{-8}). While this association does not reach the strict Bonferroni multiple-testing correction threshold we are using for rare variants (P <= 7.14 x 10^{-10}), it is interesting given the role KNTC1 plays in mitotic checkpoint activity[7], and how targeted knockdown of KNTC1 has been shown to antagonize cell proliferation and induce apoptosis across numerous cancer cell types[8–10]. Given that *KNTC1* is predominantly expressed in lymphocytes (Figure S6)[11], and that this rare missense variant is predicted to be deleterious by >= 2 computational predictors (e.g. CADD and SIFT), it is plausible that this association represents a genetic loss of function that antagonizes the clonal hematopoietic expansion that accompanies mLOY. Using a gene burden framework (i.e. ‘M1’ burden masks[12]), we also identified a rare variant signal supporting a recently described[13] risk increasing association between rare loss of function variants in the *GIGYF1* gene and mLOY (5.61 [3.35-9.40], P = 5.73 x 10^{-11}).

We also defined 22 chromosome specific binary mCA phenotypes (e.g. individuals with any mCA on chromosome 1 were cases, and individuals without any CH were controls), and used these to test for rare variant associations using a gene burden framework. Using this approach, we identified seven significant ($P \leq 2.05 \times 10^{-9}$, i.e. 3.6 x 10^{-7} burden threshold / 22 chr / 8 masks) cis gene burden associations (i.e. gene was on the same chromosome as the mCA phenotype), and one significant trans gene burden association (Table S27). Four of these cis gene burden associations were with rare loss of function variants (*TM2D3, MPL, ATM,* and *SH2B3*), and recapitulated previously reported associations where mCAs drove replacement or duplication of inherited risk alleles[14,15]. We also identified an association between mCAs on chromosome 22 and rare loss-of-function variation in the *PRR14L* gene, although these variants were strongly age associated and likely of somatic origin. This replicates recent findings that *PRR14L* is associated with uniparental disomy on chromosome 22 and clonal hematopoiesis[16], and further suggests that *PRR14L* is a CH-associated gene that is biallelically disrupted by both somatic point mutation and mosaic chromosomal alteration. We found a novel association between rare loss-of-function variation in the *RC3H1* gene (OR = 44 [16-127], P = 1.16 x 10^{-12}) and mCAs on chromosome 1, suggesting *RC3H1* as a gene that drives CH. This seems plausible given that RC3H1 is a regulator of inflammation and immune homeostasis[17], and that it is directly associated with angioimmunoblastic T-cell lymphoma in mice[18]. VAF and age associations suggested these variants were germline, and that similar to *TM2D3,* RC3H1 is biallelically lost when mCAs disrupt the remaining functional gene copy. Rare loss-of-function variants in *YLPM1* that were strongly age associated and likely somatic in origin were associated with mCAs on chr14 (OR = 30 [12-75], P = 2.44 x 10^{-13}). Interestingly, YLPM1 has been shown to limit telomerase activity by downregulating TERT expression via promoter binding. Therefore, its function loss via point mutation and/or mosaic chromosomal alteration likely drives clonal hematopoiesis. Nearly all carriers of loss of function variants across these cis-associated genes had an mCA that overlapped the gene boundary, although univariate enrichment was only significant for *MPL*, *PRR14L*, *ATM*, *RC3H1*, and *SH2B3* ($P \leq 0.009$, Table S27). Finally, we identified a trans association between rare missense variants in the *IGLL5* gene and mCAs on chr13 (OR = 30 [15-62], P = 3.09 x 10^{-21}), suggesting synergistic trans-chromosomal loss events that may drive CH.

Supplementary Note 8: ExWAS Analysis of Telomere Length

We also performed an ExWAS of leukocyte telomere length, as quantified by Codd et al.[19], conditioned on common variant signals identified by GWAS (Table S28). We found 472 significant rare variant associations ($P \leq 7.14 \times 10^{-10}$, AAF < 0.005, Table S29), including high effect size missense and/or nonsense variants in genes associated with telomere biology (*TERF1*, *POT1*, *NAF1*, *ACD*, *SAMHD1*, *HBB*, *RTEL1*, and *TINF2*).

Furthermore, we identified significant gene burden associations with telomere length (Table S30), including associations with aggregations of loss of function variants (i.e. 'M1' burden masks[12]) that suggest that disruption of the *DCLRE1B* gene significantly increases telomere length (β = 0.60 [0.47-0.73], P = 4.2 x 10^{-19}), and that disruption of the *PARN* gene significantly decreases telomere length (β = -0.61 [-0.73 - -0.49], P = 2.32 x 10^{-23}). This association with *DCLRE1B* is particularly notable given that it interacts with proteins (SNM1B/Apollo) that are required to protect telomeres against pathogenic repair-based elogation[20]. A significant positive gene burden association between rare loss of function variants in the *CTC1* gene and telomere length (β = 0.38 [0.33-0.44], P = 4.77 x 10^{-42}) is interesting given that this gene has been reported to be involved in multiple aspects of telomere maintenance[21,22], and suggests that disruption of *CTC1* has a net elongating effect on telomeres. Rare loss of function variants in the *ATM* gene were significantly associated with reduced telomere length via gene burden testing (β = -0.21 [-0.27 - -0.16], P = 3.41 x 10^{-15}), whereas rare loss of function variants in the *OBFC1* gene were significantly associated with increased telomere length via gene burden testing (β = 0.43 [0.30-0.55], P = 1.05 x 10^{-11}). *OBFC1* is a DNA replication gene previously implicated in telomere biology[23], and is also the nearest gene to the index SNP on chromosome 10 that we identified in our CHIP GWAS. For all of the telomere related genes for which we noted rare variant associations above, we also identified significant and directionally consistent gene burden associations. VAF ratios and an absence of age associations suggested that these telomere length-associated rare variants were of germline origin.

Supplementary Note 9: CHIP Associations With COVID-19

As previously reported by Zekavat et al.[24], we also saw a larger estimate of increased risk of severe COVID from CHIP among individuals with a history of solid cancers (OR = 2.94 [1.60-4.98], P = 1.69 x 10^{-4}) than among individuals with a history of liquid cancers (OR = 1.86 [0.70-4.16], P = 0.17). However, as this analysis excludes all individuals with any cancer prior to DNA collection (n=42,448), and blood cancers are much rarer than solid cancers, the analysis among individuals with a history of liquid cancers had a much lower sample size (n = ~4,000 for liquid cancers vs n = ~35,000 for solid cancers) and limited power. In an analysis among CHIP mutation carriers, CHIP VAF (as a quantitative trait) was also associated with COVID-19 hospitalization (OR = 1.16 [1.03-1.29], P =0.012) and severe COVID-19 infection (OR = 1.24 [1.02-1.51], P = 0.033). Within these models, CHIP VAF was transformed using rank inverse normalization, so the OR unit is on the rank normalized scale. While our COVID-19 data is more limited in size and quality (e.g. we do not observe a significant association between COVID-19 phenotypes and sex) in the GHS cohort, similar logistic regression modeling identified directionally consistent (although non-significant) estimates of association between CHIP mutation carrier status and hospitalization (OR = 1.18 [0.94-1.46], P = 0.14) and severe COVID-19 infection (OR = 1.42 [0.91-2.12], P = 0.11). These models also did not adjust for active malignancy status as this data was not available to us for GHS.

Models were then extended to CHIP subtypes by testing for the association between CHIP gene specific carrier status (for carriers whose mutation reached a VAF >= 0.10) and COVID-19 infection. Given sample size limits, this modeling was only done in the UKB cohort. We focused this modeling on our severe COVID-19 phenotype, as we were better powered to see effects with this trait. While estimates were positive for all CHIP subtypes, and nominally significant for *ASXL1*-CHIP (OR = 2.23 [1.14-3.88], P = 9.4 x 10^{-3}), *TP53*-CHIP (OR = 4.52 [1.10-12.2], P = 0.01), and *SF3B1*-CHIP (OR = 3.11 [0.76-8.44], P = 0.056), *PPM1D*-CHIP was the only CHIP subtype that reached significance after Bonferroni correction at α = 0.05/8 (OR = 5.42 [1.89-12.2], P = 2.8 x 10^{-4}). Furthermore, this association is even stronger (OR = 9.24 [2.80-22.5], P = 1.76 x 10^{-5}) when removing all samples with any history of cancer (i.e. an ICD10 code in any health record indicating the diagnosis of any malignant or benign cancer at any time), which we did as a sensitivity analysis to control for the possibility of confounding that may result from reported associations between *PPM1D* status and chemotherapeutic exposure[25].

Our severe COVID-19 phenotype coding was comprised exclusively of individuals that were ventilated due to COVID-19 or died due to COVID-19 and is therefore a good representation of COVID-related death or near death. Since the cancer registry data we have access to does not go beyond 2020, an individual was assigned active malignancy status if they had a record of any cancer event after January 1, 2020 (chosen to correspond with pandemic onset) in records from their general practitioner or hospital visits. Our type 2 diabetes coding was defined as individuals with billing codes in their medical records for ICD10 E11 or O241. Similar to the heavy smoking phenotype used in other analyses (see methods), smoking in these analyses was based on data surveying for smoking habits (e.g. number of cigarettes per day, self-reported ever-never smoking status, age started/stopped smoking, etc.), but was defined more broadly as ever vs never in order to limit data missingness. Ever smokers were defined as those with evidence of any previous smoking. BMI was coded as a transformed rank inversed normalized value based on the first measurement from initial intake.

Supplementary Note 10: Mendelian Randomization (MR) Analyses Provide Support For Associations Between Cancers and CHIP Subtypes

To further evaluate the relationship between CHIP and other diseases, including cardiovascular and oncologic phenotypes, we performed Mendelian Randomization (MR) using as instrumental variables the germline predictors of CHIP that we identify here. Whereas recent studies evaluating the relationship between CHIP and other disease phenotypes have relied on a very small number of instrumental variables (~1-3)[26,27], we use a much larger set of 29 independent instrumental variables derived from conditional analysis (i.e. variants identified by COJO with P_COJO <= 5×10^{-8}, Table S3, blue rows). For these MR analyses, we specifically focused on CVD and oncology phenotypes that parallel those used in our survival analysis, as well as a set of other phenotypes recently reported to associate with clonal hematopoiesis. These include Alzheimer's disease (AD)[26], liver phenotypes[27] (Alanine aminotransferase, i.e. ALT, Non-alcoholic liver disease, i.e. NALD, and cirrhosis), body mass index[28] (i.e. BMI), COVID19[24], sepsis[24], and kidney disease[29]. Since the heritability of CHIP is modest[4] (h^2 = ~0.04), MR analysis of CHIP as an exposure is likely limited by weak instrument bias and limited power[30]. Furthermore, our CHIP instruments are most reflective of associations with DNMT3A-CHIP, so we have even less power with these MR models to detect truly causal associations that are driven by non-DNMT3A CHIP subtypes. Therefore, while the absence of an MR association between CHIP and an outcome does not rule out a causal relationship, clear and significant associations across a variety of MR methods can provide additional support for a causal relationship between CHIP and other diseases. Furthermore, given the recent reporting of significant MR associations between CHIP and other diseases, we believe that our MR analyses can provide additional clarity due to the increased number of instrumental variables we utilize.

The CVD, oncological, and other phenotypes used in this analysis were coded as described in the methods section. Liver phenotypes (ALT, NALD, cirrhosis) were defined by health records as previous described[31–33]. Alzheimer's Disease was defined from health records using ICD10 code G30, and sepsis was defined from health records using ICD10 code A41. All MR analyses were done using a Two Sample MR approach, with exposure estimates derived from our UKB GWAS (using conditionally independent SNP, as described above), and outcome estimates derived from GWAS performed using the GHS cohort.

We applied seven complimentary MR methods (Table S32), as implemented in the MendelianRandomization R package (version 0.6.0). We identified significant MR associations between CHIP and myeloid leukemia (OR_{IVW} = 1.47 [1.05-2.06], P = 0.024), CHIP and lung cancer (OR_{IVW} = 1.55 [1.34-1.80], P = 8.90 x 10^{-9}), CHIP and melanoma (OR_{IVW} = 1.39 [1.13-1.1.71], P = 0.0021), CHIP and non-melanoma skin cancers (OR_{IVW} = 1.26 [1.13-1.41], P = 5.30 x 10^{-5}), CHIP and prostate cancer (OR_{IVW} = 1.20 [1.03-1.1.39], P = 0.017), and CHIP and breast cancer (1.17 [1.04-1.31], P = 0.01), but not between CHIP and any of the other phenotypes (Extended Data Figure 6A, Table S32). These MR results are consistent with and supportive of the associations we identified in our longitudinal survival analyses. However, these associations may reflect shared risk factors that independently predispose to both CHIP and various cancers. Such horizontal pleiotropy violates MR assumptions and can confound results. To address this, we used two Egger-type MR methods that are able to account for horizontal pleiotropy[34]. Each of these methods estimates an intercept term that is reflective of the pleiotropic effect. In a number of the CHIP x solid cancer associations noted above, these Egger methods estimate significant causal effect estimates beyond the effects of pleiotropy. For example, in the case of lung cancer (Figure 4, Extended Data Table 1), both MR-Egger regression and penalized robust MR-Egger regression (which can address the outsized influence of outliers) estimate the largest causal effect sizes (i.e. odd ratios) after accounting for pleiotropic effects (modeled as the significant intercept term). Since the *TERT* locus in particular is associated with lung cancer (and highly expressed in both lung and blood tissue), we repeated the MR analysis of CHIP and lung cancer after excluding *TERT* variants from our instrumental variables (25 SNPs remaining). These associations remained significant, although less strong (Figure 4, Extended Data Table 1, Table S32), and the MR methods estimate significant causal effects but non-significant pleiotropic effects. Therefore, on the whole, these results do provide support for a causal relationship between CHIP and lung cancer above and beyond the effects of horizontal pleiotropy. Nonetheless, given the aforementioned power considerations, and that a number of CHIP associated loci are in canonical cancer genes (e.g. *ATM*, *TP53*), pleiotropic confounding cannot be unequivocally ruled out when interpreting these MR results. That said, the associations between CHIP and lung cancer and CHIP and skin cancer are consistent with recent reports that CHIP may disrupt tumor immune surveillance[35], and suggest the hypothesis that the relationship between CHIP and solid tumors may have an immuno-oncological basis. Follow-up studies should test this directly, as well as if and/or how clonal hematopoiesis impacts response to checkpoint inhibiting therapeutics.

Supplementary Note 11: Sensitivity Analyses Limiting CHIP Callset To Variants With VAF < 0.35

Given the many criteria that have been applied to determining a subject's CHIP carrier status, including those reported by and advocated by flagship CHIP efforts[4,5,36] (Jaiswal, Ebert, et al.), a hard variant allele frequency filter (i.e. VAF < 0.35) that blanketly removes variants with higher VAF without regard for other criteria will eliminate correctly called CHIP variants. However, this will likely also remove a small number of germline variants that are false positives in our CHIP callset. Therefore, to ensure that our results are robust to even a minor degree of germline variant contamination, we repeated our GWAS and longitudinal analyses after filtering out variants in our CHIP callset with VAF $\geq$ 0.35.

The results are highly consistent with those we report. For example, genetic effect estimates are nearly identical when repeating genetic associations using the VAF < 0.35 callset (R^2>0.989 for

CHIP and CHIP subtypes, Figure S15). Furthermore, the protective associations we report for missense variants in *LY75* are also unchanged ($OR_{rs78446341\text{-}A} = 0.77$, $P = 3.20 \times 10^{-10}$, $OR_{rs147820690\text{-}T} = 0.46$, $P = 7.70 \times 10^{-8}$). When repeating longitudinal analyses using the filtered callset, results are generally unchanged (Figure S16-S18). One exception to this is that risk estimates for hematologic neoplasms are reduced, although still significant and consistent with our main results (Figure S16). Importantly, in our view, this effect size moderation is highly supportive of the fact that blanket filtering on VAF < 0.35 is too conservative and is removing individuals with expanded CHIP that are at the highest risk of hematologic neoplasia. This is further supported by the fact that the risks estimates that are most moderated are those for carriers of *DNMT3A* mutations who also have additional CHIP mutations (DNMT3A+), which is directly consistent with the aforementioned notion that blanket VAF filtering is eliminating individuals with expanded CHIP (e.g. those with expanded DNMT3A who have subsequently acquired additional mutations). A similar moderation is seen for risk of death (Figure S17), which likely derives from the same exclusion of individuals with expanded CHIP that are at the highest risk of hematologic neoplasia. Notably, estimates of CVD risk (Figure S17) and solid tumor risk (Figure S18) among CHIP carriers are unchanged (including risk of lung cancer for smoker and non-smoker CHIP carriers).

Supplementary Note 12: Sensitivity Analyses After Excluding Individuals With Diagnoses Of Blood Cancer Up To 90 Days After DNA Collection

We also performed additional longitudinal sensitivity analyses after excluding individuals with diagnoses of blood cancer up to 90 days after DNA collection. While all main results are calculated after already excluding individuals with any diagnosis of blood cancer prior to DNA collection date, this analysis is to further ensure that our results are robust to the possibility that some individuals we called as CHIP carriers already had latent hematologic malignancy at the time of sequencing.

Out of 8,039 individuals in UKB that we identified with blood cancer diagnoses, only 30 (0.37%) were excluded from this analysis on the basis of having a diagnosis of blood cancer within 90 days of sequencing. Furthermore, out of the 8,826 CHIP carriers evaluated for incident blood cancer, only 12 individuals (0.14%) were excluded due to having a diagnosis of blood cancer within 90 days of sequencing, including only 1 individual with a diagnosis of AML, 1 individual with a diagnosis of MDS, and 2 individuals with a diagnosis of MPN. When removing these few individuals, estimates for the risk of developing blood cancer (including myeloid, lymphoid, AML, MDS, and MPN subgroups) were entirely unchanged. Within this sensitivity analysis, our myeloid grouping was also updated to all individuals with any (ICD10) diagnostic code belonging to C93-C95. This was done in order to ensure that our estimates are robust to the inclusion of additional rarer types of myeloid malignancy. As mentioned, this did not impact our risk association estimate (OR_{main} = 11.53 [9.94-13.38], $P=1.3 \times 10^{-228}$, $OR_{sensitivity\_analysis}$ = 11.26 [9.72-13.05], $P=7.2 \times 10^{-229}$).

Overall, these sensitivity analyses provide further support for our main callset and main results and suggest that i) our main genetic and phenotypic results are robust ii) our main callset is well calibrated. This is further supported by the fact that our association estimates between CHIP and demographic variables such as age and smoking are highly consistent with those reported previously in the literature. Nonetheless, we report the results of this sensitivity analysis to further facilitate decision making by researchers as they use/filter our callset in whichever way they feel is best for their research questions of interest.

1. Nakao, T. *et al.* Mendelian randomization supports bidirectional causality between telomere length and clonal hematopoiesis of indeterminate potential. *Science advances* **8**, eabl6579 (2022).
2. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369–375 (2012).
3. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

4. Bick, A. G. *et al*. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).

5. Jaiswal, S. *et al*. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *New England Journal of Medicine* **377**, 111–121 (2017).

6. Jones, A. V. *et al.* The JAK2 46/1 haplotype predisposes to MPL-mutated myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology* **115**, 4517–4523 (2010).

7. Scaërou, F. *et al.* The ZW10 and Rough Deal checkpoint proteins function together in a large, evolutionarily conserved complex targeted to the kinetochore. *Journal of cell science* **114**, 3103–3114 (2001).

8. Zhengxiang, Z., Yunxiang, T., Zhiping, L. & Zhimin, Y. KNTC1 knockdown suppresses cell proliferation of colon cancer. *3 Biotech* **11**, 1–11 (2021).

9. Huang, H., Fan, X., Qiao, Y., Yang, M. & Ji, Z. Knockdown of KNTC1 inhibits the proliferation, migration and tumorigenesis of human bladder cancer cells and induces apoptosis. *Critical Reviews^{TM} in Eukaryotic Gene Expression* **31**, (2021).

10. Liu, C.-T. *et al.* shRNA-mediated knockdown of KNTC1 suppresses cell viability and induces apoptosis in esophageal squamous cell carcinoma. *International Journal of Oncology* **54**, 1053–1060 (2019).

11. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

12. Backman, J. D. *et al*. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 1–10 (2021) doi:10.1038/s41586-021-04103-z.

13. Zhao, Y. *et al.* GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. *Nat Commun* **12**, 4178 (2021).

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

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

16. Chase, A. *et al.* PRR14L mutations are associated with chromosome 22 acquired uniparental disomy, age-related clonal hematopoiesis and myeloid neoplasia. *Leukemia* **33**, 1184–1194 (2019).

17. Schaefer, J. S. & Klein, J. R. Roquin—a multifunctional regulator of immune homeostasis. *Genes & Immunity* **17**, 79–84 (2016).

18. Chiba, S. & Sakata-Yanagimoto, M. Advances in understanding of angioimmunoblastic T-cell lymphoma. *Leukemia* **34**, 2592–2606 (2020).

19. Codd, V. *et al.* Measurement and initial characterization of leukocyte telomere length in 474,074 participants in UK Biobank. *Nat Aging* **2**, 170–179 (2022).

20. Lenain, C. *et al.* The Apollo 5′ Exonuclease Functions Together with TRF2 to Protect Telomeres from DNA Repair. *Current Biology* **16**, 1303–1310 (2006).

21. Miyake, Y. *et al.* RPA-like mammalian Ctc1-Stn1-Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. *Molecular cell* **36**, 193–206 (2009).

22. Chen, L.-Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544 (2012).

23. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proceedings of the National Academy of Sciences* **107**, 9293–9298 (2010).

24. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nature Medicine* **27**, 1012–1024 (2021).

25. Bolton, K. L. *et al.* Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature Genetics* **52**, 1219–1226 (2020).

26. Bouzid, H. *et al.* Clonal hematopoiesis is associated with protection from Alzheimer's disease. *medRxiv* (2021).

27. Wong, W. J. *et al.* Clonal hematopoiesis and risk of chronic liver disease. *medRxiv* 2022.01.17.22269409 (2022) doi:10.1101/2022.01.17.22269409.

28. Haring, B. *et al.* Healthy Lifestyle and Clonal Hematopoiesis of Indeterminate Potential: Results From the Women's Health Initiative. *Journal of the American Heart Association* **10**, e018789 (2021).

29. Vlasschaert, C. *et al.* Association of Clonal Hematopoiesis of Indeterminate Potential with Worse Kidney Function and Anemia in Two Cohorts of Patients with Advanced Chronic Kidney Disease. *Journal of the American Society of Nephrology* (2022).

30. Bowden, J. & Holmes, M. V. Meta-analysis and Mendelian randomization: A review. *Research synthesis methods* **10**, 486–496 (2019).

31. Abul-Husn, N. S. *et al.* A protein-truncating HSD17B13 variant and protection from chronic liver disease. *New England Journal of Medicine* **378**, 1096–1106 (2018).

32. Akbari, P. *et al.* Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. *Science* **373**, eabf8683 (2021).

33. Dewey, F. E. *et al.* Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. *New England Journal of Medicine* **377**, 211–221 (2017).

34. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics* **27**, R195–R208 (2018).

35. Liu, X. *et al.* CHIP-associated mutant ASXL1 in blood cells promotes solid tumor progression. *Cancer Science* (2022).

36. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* **371**, 2488–2498 (2014).