

Mitochondrial heteroplasmy improves risk prediction for myeloid neoplasms

Received: 8 May 2024

Accepted: 7 November 2024

Published online: 22 November 2024

 Check for updates

Yun Soo Hong^{1,10}, Sergiu Pasca^{2,3,10}, Wen Shi¹ , Daniela Puiu⁴, Nicole J. Lake⁵, Monkol Lek⁵ , Meng Ru⁶, Megan L. Grove⁷, Anna Prizment⁸, Corinne E. Joshu^{3,6}, Elizabeth A. Platz^{3,6} , Eliseo Guallar⁹, Dan E. Arking¹   & Lukasz P. Gondek^{1,2,3}  

Clonal hematopoiesis of indeterminate potential is the primary pathogenic risk factor for myeloid neoplasms, while heteroplasmy (mutations in a subset of cellular mitochondrial DNA) is another marker of clonal expansion associated with hematological malignancies. We explore how these two markers relate and influence myeloid neoplasms incidence, and their role in risk stratification. We find that heteroplasmy is more common in individuals with clonal hematopoiesis of indeterminate potential, particularly those with higher variant allele fractions, multiple mutations, or spliceosome machinery mutations. Individuals with both markers have a higher risk of myeloid neoplasms than those with either alone. Furthermore, heteroplasmic variants with higher predicted deleteriousness increase the risk of myeloid neoplasms. Incorporating heteroplasmy in an existing risk score model for individuals with clonal hematopoiesis of indeterminate potential significantly improves sensitivity and better identifies high-risk groups. This suggests heteroplasmy as a clonal expansion marker and potentially as a biomarker for myeloid neoplasms development.

Screening, detection, and control of early malignancy or premalignant lesions are the hallmark of secondary prevention of cancer. This set of interventions, while successfully applied in various solid tumor malignancies, has not been widely adopted in myeloid neoplasms (MN), likely due to the lack of precise biomarkers of MN risk^{1–3}. In recent years, significant progress has been made, fueled mainly by advances in cancer genomics and next-generation sequencing (NGS). The broad application of this technology to large population-based cohorts, providing easy access to sequencing data for the scientific community, allowed for the identification of clonal expansion of

hematopoietic cells, also known as clonal hematopoiesis (CH), in over 10% of the adult population^{4–7}. The most studied form of CH is clonal hematopoiesis of indeterminate potential (CHIP), which is defined as the presence of cancer-associated somatic mutations in hematopoietic cells in otherwise healthy individuals. Given its clonal nature, it is not surprising that some forms of CHIP are myeloid premalignant conditions. Most recently, multiparameter MN prediction models including somatic mutations, hematologic indices, and demographic data have been developed to identify the population with CHIP at high risk for developing MN^{8,9}. Unfortunately, these risk scoring systems are limited

¹McKusick-Nathans Institute, Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²Division of Hematological Malignancies, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ³Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ⁴Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. ⁵Department of Genetics, Yale School of Medicine, New Haven, CT, USA. ⁶Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ⁷Human Genetics Center; Department of Epidemiology, Human Genetics, and Environmental Sciences; School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. ⁸Department of Laboratory Medicine & Pathology, University of Minnesota Medical School, Minneapolis, Minnesota, USA. ⁹Department of Epidemiology, School of Global Public Health, New York University, New York, NY, USA. ¹⁰These authors contributed equally: Yun Soo Hong, Sergiu Pasca. ✉ e-mail: arking@jhmi.edu; lgondek1@jhmi.edu

to individuals with CHIP, who constitute only a small fraction of individuals at risk.

Somatic mitochondrial DNA (mtDNA) mutations are common and have been reported not only in mitochondrial diseases but also in aging and cancer^{10,11}. Unlike diploid nuclear DNA, mtDNA exists in 10 s to 1000 s of copies within each cell. Thus, mutations in the mtDNA can exist in a subset of the total cellular mtDNA, a condition termed mitochondrial heteroplasmy. As mtDNA repair machinery is limited and not as efficient as the nuclear DNA (nDNA) repair system, the mtDNA mutation rate is 10–17 fold higher than in nDNA^{12,13}. Heteroplasmy can therefore be used as an endogenous cell barcode allowing for lineage tracing and assessment of the clonal expansion of hematopoietic cells and serves as an excellent marker of CH¹⁴. In participants from the UK Biobank (UKB), heteroplasmy was present in 30% of individuals, and consistent with CH, the frequency increased with age¹⁵. Given their central role in essential cellular processes, mitochondrial alterations are key components of several hallmarks of cancer such as cellular energetics, proliferation, and apoptosis^{16,17}. Thus, it is not surprising that heteroplasmy may provide additional mechanisms of selective growth advantage of abnormal clone(s) and further shape the genomic landscape of CH and MN. We have recently reported that mtDNA heteroplasmy was associated with a 1.5-fold increase in all-cause mortality, and the presence of mtDNA mutations at highly constrained sites was associated with a 4-fold increase in mortality due to leukemia¹⁵.

In this work, we use two large population-based cohorts, UKB and Atherosclerosis Risk in Communities (ARIC) study, to examine the association between somatic mtDNA and nDNA mutations (mito-nuclear interaction) in clonal evolution and its impact on the incidence of MN. We further assess the role of heteroplasmy as an independent predictor of MN risk and whether the inclusion of heteroplasmy into the latest MN prediction algorithms can improve risk classification.

Results

Demographics and molecular characteristics of the UKB and ARIC cohorts

The study included a total of 441,936 participants from the UKB ($n = 434,304$) and ARIC ($n = 7632$) cohorts, who had information on both CHIP and mitochondrial heteroplasmy. In the UKB, the mean age was 56.5 (8.1) years, 45.8% ($n = 199,046$) were men and 94.5% ($n = 410,313$) were self-identified as White. In ARIC, the mean age was 57.9 (6.0) years, 45.2% ($n = 3453$) were men and 76.2% ($n = 5817$) and 23.8% ($n = 1815$) were self-identified as Whites and Blacks, respectively (Table 1).

Using whole exome sequencing data (WES), we identified 37,089 CHIP mutations in the UKB and 435 CHIP mutations in ARIC at a variant allele fraction (VAF) of 2%, which were present in 33,597 (7.7%) UKB participants and 375 (4.9%) ARIC participants. Similar to previous reports, the most commonly mutated genes were *DNMT3A*, *TET2*, and *ASXL1* (Fig. 1A, B, Supplementary fig. 1). Mutations in spliceosome genes (*SRSF2*, *SF3B1* and *U2AF1*) were associated with a higher median (interquartile range [IQR]) VAF compared to other CHIP mutations in UKB participants (13.0% [7.9%, 24.1%] vs 7.2% [5.0%, 13.7%]; $P < 0.0001$) but not in ARIC (10.1% [7.2%, 20.0%] vs 10.4% [4.9%, 18.2%]; $P = 0.72$) (Fig. 1C, D). Most participants with CHIP had only one mutation, with multiple mutations seen in only 8.6% of UKB ($n = 2873$) and 12.0% of ARIC participants ($n = 45$) (Fig. 1E, F). As expected, CHIP was associated with older age and smoking status (Table 1).

MtDNA heteroplasmy was identified from whole genome sequencing (WGS) data using the MitoHPC pipeline¹⁸. We used a heteroplasmy VAF threshold of 5% (single nucleotide variants [SNVs] between 5% and 95% were called as heteroplasmic) based on extensive simulations and assessment of empirical data to maximize sensitivity to true heteroplasmies while minimizing potential false-positives due

Table 1 | Participant characteristics by presence of CHIP in each cohort

	UK Biobank			ARIC		
	CHIP (-)	CHIP (+)	P	CHIP (-)	CHIP (+)	P
Number	400,707	33,597		7,257	375	
Age, years	56.3 (8.1)	58.9 (7.6)	< 2.2e-16	57.8 (6.0)	59.6 (6.1)	3.9e-8
Men, %	45.9	44.8	9.0e-5	45.1	48.8	0.17
Self-reported race, %			< 2.2e-16			0.0061
White	94.4	95.4		76.5	70.1	
Black	1.5	1.3		23.5	29.9	
Asian	2.2	1.8		0	0	
Other	1.5	1.2		0	0	
Missing	0.3	0.3		0	0	
Ever smoker, %	59.5	62.4	< 2.2e-16	60.5	61.6	0.70
Anemia, %	4.2	4.5	0.0028	9.3	12.5	0.06
Thrombocytopenia, %	2.2	2.5	0.00035	2.1	4.9	0.012
Neutropenia, %	1.0	1.2	0.012	3.1	1.9	0.63
Cytopenia, %	7.0	7.6	7.8e-5	11.1	13.7	0.30
MCV, fL	91.1 (4.6)	91.3 (4.6)	4.1e-11	89.8 (5.1)	89.6 (5.5)	0.54
RDW, %	13.5 (1.0)	13.6 (1.1)	< 2.2e-16	NA	NA	NA
History of cancer, %	6.9	8.8	< 2.2e-16	0	0	NA
Heteroplasmy, %	27.8	34.1	< 2.2e-16	27.9	35.2	0.0027

Continuous variables are displayed as either mean (SD) or median (25th and 75th percentiles). Participant characteristics for both cohorts were based on the time of DNA collection. In ARIC, only participants without a cancer diagnosis at the time of blood collection were included in the analysis. Because of missing blood parameters, in ARIC, we were able to evaluate anemia in 6473 (84.8%) participants; thrombocytopenia in 5161 (67.6%) participants; neutropenia in 3372 (44.2%) participants; and cytopenia in 3344 (43.8%) participants. *P* values were calculated using a two-sided test without multiple testing correction for continuous variables or a Fisher's exact test without multiple testing correction for categorical variables. ARIC Atherosclerosis Risk in Communities study, CHIP clonal hematopoiesis of indeterminate potential, MCV mean corpuscular volume, NA not available and RDW red cell distribution width.

to low-level cross-contamination or mis-mapping of nuclear-encoded mitochondrial DNA (NUMTs)¹⁸. To identify predicted deleterious mtDNA variants, we used a modified mitochondrial DNA local constraint (mMLC) score, which quantifies the local tolerance to base or amino acid substitution for each base pair in the mtDNA genome¹⁹. The mMLC score ranges from 0 to 1, with higher scores indicating more constrained, and therefore, more deleterious SNVs. To capture the overall impact of multiple heteroplasmies in a given individual, we generated the mMLC score sum (mMSS) by summing all mMLC scores for that individual. Prior work has demonstrated that the mMSS is a stronger predictor of overall mortality than heteroplasmy count, suggesting that deleterious mitochondrial heteroplasmies may be causally linked to adverse outcomes¹⁵. Heteroplasmy was present in 122,969 (28.3%) and 2,159 (28.3%) participants in the UKB and ARIC, respectively, with heteroplasmies most frequently observed in Complex I genes and the D-loop (Fig. 2A, B). Heteroplasmic variants in rRNA and tRNA were predicted to be more deleterious than in other mtDNA complex / regions, as reflected by higher mMSS scores (Fig. 2C, D). Among those with heteroplasmy, 23,642 (19.2%) participants in UKB and 417 (19.3%) in ARIC had more than 1 heteroplasmy (Fig. 2E, F). Similar to CHIP, heteroplasmy was associated with older age and smoking status (Table 2).

Association between CHIP and heteroplasmy

We and others^{15,20–22} have demonstrated that a notable number of heteroplasmies identified in adults reflect acquired mutations, and thus, likely mark the clonal expansion of hematopoietic stem cells

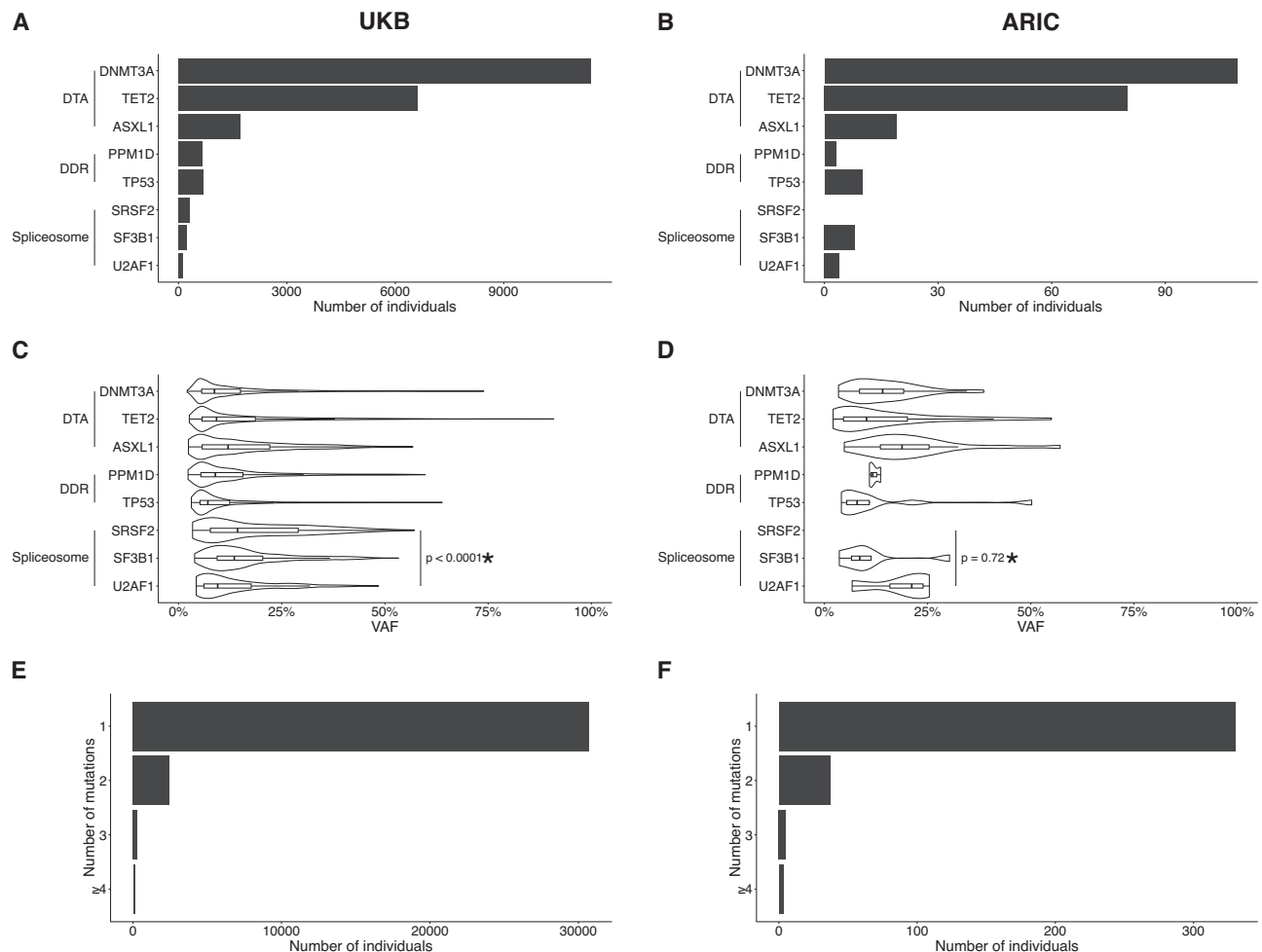


Fig. 1 | Description of CHIP mutations. Distribution of the number of individuals carrying mutations stratified by genes in **(A)** UKB and **(B)** ARIC. Distribution of VAF stratified by gene in **(C)** UKB (number of mutations used to derive statistics; median VAF [Q1, Q3] *DNMT3A*: $n = 11,648$; 8.7% [5.6%, 15.0%]; *TET2*: $n = 6,923$; 9.2% [5.7%, 18.6%]; *ASXL1*: $n = 1,734$; 12.0% [5.7%, 22.1%]; *PPM1D*: $n = 658$; 8.9% [5.4%, 15.6%]; *TP53*: $n = 699$; 7.1% [5.2%, 12.4%]; *SRSF2*: $n = 317$; 14.3% [7.7%, 29.0%]; *SF3B1*: $n = 239$; 13.5% [9.4%, 20.4%]; *U2AF1*: $n = 123$; 9.5% [6.1%, 17.6%]) and in **(D)** ARIC (number of mutations used to derive statistics; median VAF [Q1, Q3] *DNMT3A*: $n = 113$; 14.1% [8.5%, 19.2%]; *TET2*: $n = 85$; 10.2% [4.6%, 20.1%]; *ASXL1*: $n = 19$; 18.8% [13.6%, 25.4%]; *PPM1D*: $n = 3$; 11.7% [11.3%, 12.6%]; *TP53*: $n = 10$; 7.9% [5.4%, 10.9%]; *SRSF2*: $n = 0$; NA

[NA, NA]; *SF3B1*: $n = 8$; 8.6% [6.6%, 11.3%]; *U2AF1*: $n = 4$; 21.2% [15.8%, 23.9%]). Two-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction was used to assess the difference between VAF of spliceosome mutations compared to that of other mutations in both UKB ($P < 2.2 \times 10^{-16}$) and ARIC ($P = 0.7234$). Number of mutations per individual in **(E)** UKB and **(F)** ARIC. Only the DTA, DDR and classic spliceosome mutations are presented. Abbreviations: VAF, variant allele fraction; and DDR, DNA damage response. DTA includes *DNMT3A*, *TET2*, *ASXL1*. * Denotes the P value for a comparison between spliceosome mutations and other CHIP mutations. Source data are provided as a Source Data file.

(HSC). Therefore, we sought to determine whether mtDNA heteroplasmy serves as a more sensitive or alternative marker of CH, and whether it is also involved in biologically relevant processes, contributing to relative expansion of hematopoietic clones either independently or in combination with CHIP mutations. To address this question, we analyzed the association between various metrics of heteroplasmy and CHIP using multivariable logistic regression models adjusted for age, sex, smoking status, and a history of cancer. Concurrent CHIP and heteroplasmy (“CHIP and heteroplasmy”) was present in approximately 2–3% of participants (UKB = 2.6%, ARIC = 1.7%) (Fig. 3A, B). Heteroplasmy without concurrent CHIP (“Heteroplasmy only”) was seen in approximately 25% of participants (UKB = 25.7%, ARIC = 26.6%), and approximately 5% of individuals had CHIP only without evidence of heteroplasmy (“CHIP only”; UKB = 5.1%, ARIC = 3.2%). Notably, heteroplasmy was more common in CHIP participants compared to those without CHIP in both UKB (34.1% vs. 27.8%; $P < 0.0001$) and ARIC (35.2% vs. 27.9%; $P < 0.01$) (Fig. 3C, D). Among participants with CHIP in UKB, heteroplasmy was more common in individuals with large CHIP clones (VAF $\geq 20\%$) (44.1% vs. 32.2%;

$P < 0.0001$) and in those with multiple mutations (39.5% vs. 33.6%; $P < 0.0001$). These results were consistent in ARIC, where heteroplasmy was more common in individuals with large clones (42.4% vs. 33.1%; $P = 0.09$) and multiple mutations (48.9% vs. 33.3%; $P = 0.04$) (Fig. 3C, D). Among CHIP genotypes, heteroplasmy was significantly more common in individuals with spliceosome mutations (UKB 55.1% vs. 33.6%; $P < 0.0001$ and ARIC 66.7% vs. 34.3%; $P = 0.05$) (Fig. 3C, D, Supplementary Fig. 2).

Given the association between CHIP and heteroplasmy, we explored whether having multiple heteroplasms was also associated with features of CHIP in general and, particularly, with high-risk CHIP. Among individuals with heteroplasmy in the UKB, those with CHIP were more likely to have multiple heteroplasms compared to those without CHIP (24.1% vs. 18.7%; $P < 0.0001$) (Fig. 3E). Large clones (30.7% vs. 22.4%; $P < 0.0001$) and multiple CHIP mutations (30.1% vs. 23.5%; $P < 0.0001$) were also associated with the presence of multiple heteroplasms (Fig. 3E). Among CHIP genotypes, multiple heteroplasms were more common in CHIP with spliceosome mutations (39.1% vs. 23.6%; $P < 0.0001$) (Fig. 3E). Similar results, though not

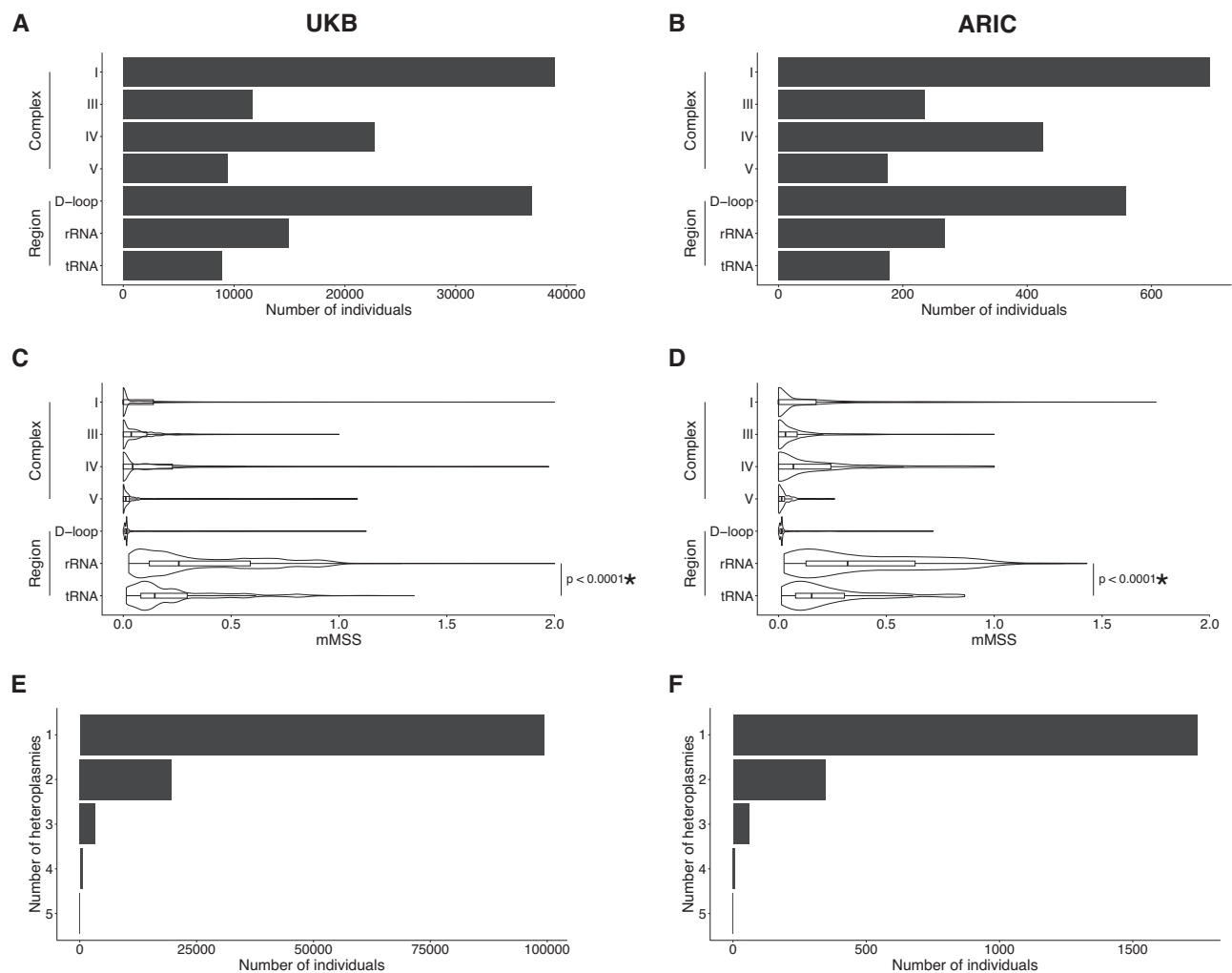


Fig. 2 | Description of heteroplasmy. Distribution of the number of individuals with heteroplasmy stratified by complex/region in (A) UKB and (B) ARIC. Distribution of mMSS stratified by complex/region in (C) UKB (number of individuals used to derive statistics; median mMSS [Q1, Q3] Complex I: $n = 38,931$; 0.000 [0.000, 0.140]; Complex III: $n = 11,700$; 0.037 [0.000, 0.109]; Complex IV: $n = 22,632$; 0.043 [0.000, 0.228]; Complex V: $n = 9,416$; 0.012 [0.000, 0.029]; D-loop: $n = 36,829$; 0.015 [0.007, 0.017]; rRNA: $n = 14,891$; 0.258 [0.121, 0.589]; tRNA: $n = 8,876$; 0.146 [0.081, 0.296]) and (D) ARIC (number of individuals used to derive statistics; median mMSS [Q1, Q3] Complex I: $n = 694$; 0.000 [0.000, 0.174]; Complex III: $n = 236$; 0.032 [0.000, 0.087]; Complex IV: $n = 426$; 0.069 [0.000, 0.243];

Complex V: $n = 175$; 0.015 [0.002, 0.028]; D-loop: $n = 559$; 0.015 [0.007, 0.017]; rRNA: $n = 267$; 0.320 [0.127, 0.633]; tRNA: $n = 179$; 0.154 [0.079, 0.306]). Graphical representation of mMSS was truncated at mMSS = 2. All values of mMSS were included for the analyses. Two-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction was used to assess the difference between mMSS of rRNA/tRNA compared to mMSS of other complexes/regions in both UKB ($P < 2.2e-16$) and ARIC ($P < 2.2e-16$). Number of heteroplasms per individual in (E) UKB and (F) ARIC. Abbreviations: mMSS, modified mitochondrial local constraint (MLC) score sum. * Denotes comparison between mMSS in rRNA/tRNA and mMSS occurring in other complexes/regions. Source data are provided as a Source Data file.

significant, were observed in ARIC, with limited power due to the smaller cohort size (Fig. 3F).

To address whether the higher prevalence of heteroplasmy in CHIP individuals was just a marker of CH (passenger mtDNA mutations) rather than likely functionally contributing to clonal expansion, we assessed the association between CHIP and mMSS adjusted for age, sex, smoking status, history of cancer, and heteroplasmy count. Among those with heteroplasmy, the presence of CHIP was associated with a significantly higher mMSS compared to no CHIP (Fig. 3G, H, Supplementary Table 2). Moreover, higher mMSS score was observed in individuals with higher clonal burden ($VAF \geq 20\%$) and with multiple CHIP mutations (Fig. 3G, H, Supplementary Table 2). When stratified by specific CHIP genotypes, mutations in spliceosome genes (*SRSF2*, *SF3B1* and *U2AF1*) were associated with a higher mMSS in UKB and had a similar directionality in ARIC when compared to other CHIP mutations (Fig. 3G, H, Supplementary fig. 2, Supplementary Table 2). These data suggest that the presence of mtDNA mutations is not merely a

marker of clonal expansion but may confer an important cell adaptation.

Similarly, when we compared heteroplasmy by CHIP and cytopenia status (no CHIP/ clonal cytopenia of undetermined significance [CCUS], CHIP without cytopenia, and CCUS), we saw a gradient increase in the prevalence of heteroplasmy (Supplementary fig. 3). Those with CCUS had a higher prevalence of heteroplasmy compared to those with CHIP without cytopenia (UKB 36.1% vs 33.9%; $P = 0.06$ and ARIC 38.1% vs 33.3%; $P = 0.59$). In addition, those with CCUS had a higher prevalence of multiple heteroplasms (UKB 28.0% vs 23.8%; $P = 0.02$ and ARIC 37.5% vs 22.7%; $P = 0.47$) and a higher mMSS (Supplementary Table 2) compared to those with CHIP without cytopenia.

To determine if the increase in heteroplasmy prevalence in CHIP was due to a specific mtDNA complex / region, we assessed enrichment with measures calculated separately for each complex / region (Supplementary fig. 4). We observed a higher prevalence of rRNA mutations and a lower prevalence of D-loop mutations in participants with CHIP

Table 2 | Participant characteristics by the presence of heteroplasmy in each cohort

	UK Biobank			ARIC		
	Heteroplasmy (-)	Heteroplasmy (+)	P	Heteroplasmy (-)	Heteroplasmy (+)	P
Number	311,335	122,969		5473	2159	
Age, years	56.2 (8.1)	57.2 (8.0)	< 2.2e-16	57.7 (6.0)	58.3 (6.1)	0.0003973
Men, %	45.7	46.3	0.00036	44.6	46.9	0.074
Self-reported race, %			< 2.2e-16			0.27
White	94.3	95.0		76.6	75.4	
Black	1.6	1.4		23.4	24.6	
Asian	2.3	2.0		0	0	
Other	1.5	1.4		0	0	
Missing	0.3	0.3		0	0	
Ever smoker, %	59.3	60.8	< 2.2e-16	59.5	63.1	0.0042
Anemia, %	4.2	4.3	0.43	9.5	9.3	0.89
Thrombocytopenia, %	2.1	2.3	0.00012	2.1	2.6	0.35
Neutropenia, %	1.0	1.0	0.57	2.6	4.0	0.043
Cytopenia, %	7.0	7.2	0.072	10.6	12.8	0.077
MCV, fL	91.1 (4.5)	91.2 (4.6)	2.08e-14	89.7 (5.1)	89.9 (4.9)	0.28
RDW, %	13.5 (1.0)	13.5 (1.0)	3.46e-10	NA	NA	NA
History of cancer, %	6.8	7.6	< 2.2e-16	0	0	NA
CHIP (+), %	7.1	9.3	< 2.2e-16	4.4	6.1	0.0027

Continuous variables are displayed as either mean (SD) or median (25th and 75th percentiles). Participant characteristics for both cohorts were based on the time of DNA collection. In ARIC, only participants without a cancer diagnosis at the time of blood collection were included in the analysis. Because of missing blood parameters, in ARIC, we were able to evaluate anemia in 6473 (84.8%) participants; thrombocytopenia in 5161 (67.6%) participants; neutropenia in 3372 (44.2%) participants and cytopenia in 3344 (43.8%) participants. *P* values were calculated using a two-sided test without multiple testing correction for continuous variables or a Fisher's exact test without multiple testing correction for categorical variables. ARIC Atherosclerosis Risk in Communities study, CHIP clonal hematopoiesis of indeterminate potential, MCV mean corpuscular volume, NA not available and RDW red cell distribution width.

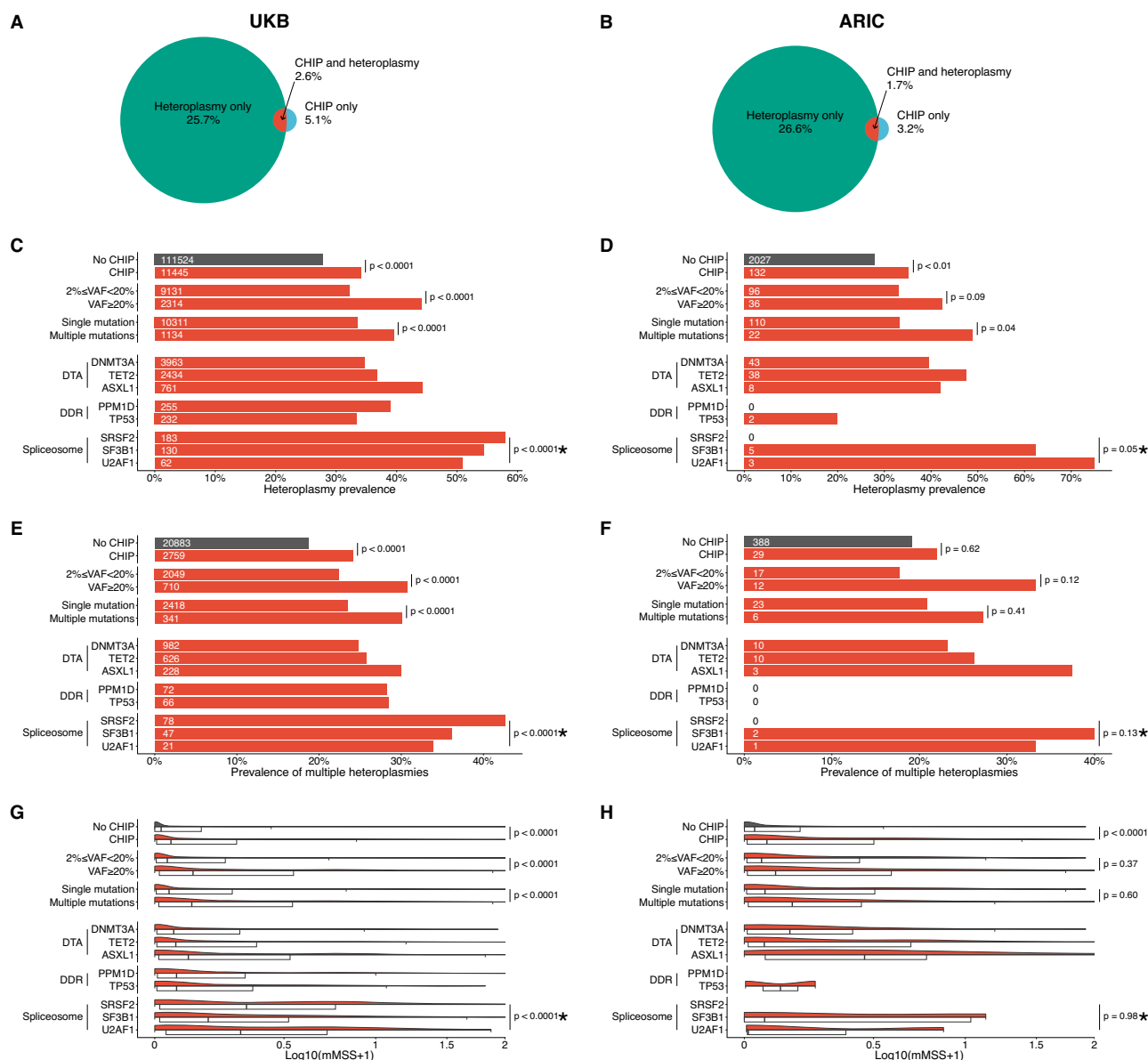
compared to participants without CHIP, and in high-risk subsets of CHIP (VAF > 20%, multiple mutations, spliceosome mutations) (Supplementary fig. 4A, B, C, D). Because this enrichment could simply reflect the fact that rRNA mutations are more likely to be deleterious (i.e., heteroplasms have higher average mMLC), we also evaluated the mMSS by complex / region. We observed that, even when restricted to individuals with rRNA mutations, participants with CHIP had higher mMSS values compared to those without CHIP, and, among individuals with CHIP, high-risk subsets of CHIP (VAF > 20%, multiple mutations, spliceosome mutations) had higher mMSS values (Supplementary fig. 4E, F, G, H).

Heteroplasmy and the risk of MN

We have previously reported an association between heteroplasmy and incident and prevalent hematologic malignancies¹⁵. Analogous to CHIP and MN, we sought to determine whether the presence of heteroplasmy in the general population was associated with an increased risk of MN. There were 1191 and 160 cases of incident MN in UKB and ARIC, respectively. Participants with heteroplasmy had a higher risk of developing MN even after adjusting for age, sex, smoking status, and a history of cancer (UKB only) in the UKB (HR = 2.1; 95% CI 1.9–2.3; *P* < 0.0001) and ARIC (HR = 1.7; 95% CI 1.2–2.3; *P* < 0.01) (Fig. 4A, B). Among individuals with heteroplasmy, heteroplasmy count was associated with a higher risk of MN in the UKB (HR = 1.7; 95% CI 1.5–1.9; *P* < 0.0001) and ARIC (HR = 1.3; 95% CI 0.9–2.0; *P* = 0.13). Similarly, mMSS was associated with an increased risk of MN in both the UKB (HR = 4.4; 95% CI 3.7–5.2; *P* < 0.0001) (Supplementary fig. 5A, B) and ARIC (HR = 2.2; 95% CI 1.1–4.4; *P* = 0.02). We found similar results when the analysis was stratified by CHIP status (Supplementary fig. 6 and 7). When heteroplasmy count and mMSS were mutually adjusted in the same model, the heteroplasmy count (UKB [HR = 1.2; 95% CI 1.1–1.4; *P* < 0.01] and ARIC [HR = ; 95% CI 0.7–1.8; *P* = 0.53]) and mMSS (UKB [HR = 3.7; 95% CI 3.0–4.6; *P* < 0.0001] and ARIC [HR = 2.0; 95% CI 0.9–4.3; *P* = 0.08]) were independently associated with MN, indicating

that heteroplasmy is not only a marker for clonal expansion but ostensibly deleterious mtDNA mutations are associated with MN. This claim is further supported when analyzing heteroplasmy by complex / region, where D-loop heteroplasmy count, which is largely driven by benign heteroplasms, showed no association, but D-loop mMSS was significantly associated, with similar effect sizes to the other complexes / regions (Supplementary fig. 8A, B). In the UKB, the associations were independent of self-reported race, smoking status, and a history of cancer. (Supplementary Table 3). Among smokers, the association between heteroplasmy count (HR = 1.2; 95% CI 1.0–1.5; *P* = 0.07), mMSS (HR = 3.2; 95% CI 2.3–4.5; *P* < 0.0001) and MN remained significant after adjusting for the pack-years.

We further explored the potential interaction between CHIP and heteroplasmy in the development of MN by evaluating the associations in individuals with either factor alone and with both CHIP and heteroplasmy. Compared to individuals without either CHIP or heteroplasmy, those with either one alone had an elevated risk of MN (For CHIP, UKB HR = 5.3; 95% CI 4.5–6.3; *P* < 0.0001; ARIC HR = 1.3; 95% CI 0.5–3.2; *P* = 0.55. For heteroplasmy, UKB HR = 1.7; 95% CI 1.5–2.0; *P* < 0.0001; ARIC HR = 1.5; 95% CI 1.1–2.1; *P* = 0.02) (Fig. 4C, D). Notably, participants with both CHIP and heteroplasmy (UKB [HR = 12.1; 95% CI 10.4–14.2; *P* < 0.0001]; ARIC [HR = 4.9; 95% CI 2.5–9.4; *P* < 0.0001]) had a significantly higher risk of MN compared to the expected combined (multiplicative) effect assuming independence of these factors (*P* for interaction = 0.02 in the UKB; *P* for interaction = 0.12 in ARIC). This finding demonstrated a synergistic effect between CHIP and heteroplasmy, potentially amplifying the risk of MN incidence. In the UKB, the association remained significant after accounting for self-reported race, smoking status, and a history of cancer in subgroup analyses (Supplementary Table 3). Finally, we observed that heteroplasmy was an independent predictor of MN risk even when adjusted for the number mutations and clonal burden in UKB (HR = 1.9; 95% CI 1.5–2.3; *P* < 0.0001) and in ARIC (HR = 3.9; 95% CI 1.3–11.8; *P* = 0.02).



Myeloid malignancy risk score including mitochondrial heteroplasmies in the UK Biobank

Building on these findings, we sought to test whether heteroplasmies can be used to more accurately predict MN incidence. A clonal hematopoiesis risk score (CHRS) was recently developed in the UKB to better assess the risk of progression to MN among individuals with CHIP⁹. The CHRS is calculated using 8 components, including age ≥ 65 years, presence of cytopenia, red cell distribution width (RDW) ≥ 15 , mean corpuscular volume (MCV) ≥ 100 , presence of high-risk mutation, a single *DNMT3A* mutation, number of mutations, and VAF $\geq 20\%$, and categorized as low-, intermediate-, and high-risk based on 10-year cumulative incidence⁹. Of the UKB participants with CHIP, 30,542 (90.9%), 2,821 (8.4%), and 234 (0.7%) participants were classified as CHRS low-, intermediate-, and high-risk, respectively, and there were 457 incident cases of MN during a median (IQR) follow-up of 13.7 (12.9, 14.4) years. Compared to those in the low-risk category, intermediate- (HR 10.6; 95% CI 8.6–13.0; $P < 0.0001$) and high-risk (HR = 99.1; 95% CI 76.6–128.3; $P < 0.0001$) categories had a higher risk of MN, after adjusting for sex, smoking status, and a history of cancer (Supplementary Fig. 9). The results were similar when the analysis was restricted to self-reported Whites, unrelated individuals, and never smokers (Supplementary Table 3).

Compared to those in the low-risk CHRS category, participants in the intermediate- and high-risk categories were, on average, more likely to have a higher heteroplasmies count and mMSS (Supplementary Table 4, Supplementary Fig. 10). We further evaluated whether heteroplasmies was associated with MN independent of CHRS. Both heteroplasmies count (HR 1.4; 95% CI 1.3–1.6; $P < 0.0001$) and mMSS (HR 2.1; 95% CI 1.7–2.5; $P < 0.0001$) were associated with incident MN, after adjusting for CHRS category, sex, smoking, and a history of cancer. When heteroplasmies count and mMSS were mutually adjusted for in the same regression model, the associations were attenuated (HR 1.3; 95% CI 1.1–1.4; $P = 0.001$ for heteroplasmies count and HR 1.5; 95% CI 1.1–2.0; $P = 0.007$ for mMSS) but retained their significance, suggesting that heteroplasmies is not only a biomarker, but deleterious variants may be also causally associated with MN.

Having established that heteroplasmies is a predictor of MN independent of CHRS, we updated the CHRS model (CHRS-M) to incorporate the presence of heteroplasmies and mMSS, using the algorithm as outlined in the original CHRS manuscript (round the effect estimate to the nearest 0.5 and add 1)⁹ (Fig. 5, Table 3, Supplementary Fig. 11 and 12). To account for the two additional heteroplasmies parameters in the score (presence of heteroplasmies and mMSS), we added 2 to the score cutoffs used by the CHRS to classify risk groups; low

Fig. 3 | Association between CHIP and heteroplasmy. The percentage of individuals presenting with both CHIP and heteroplasmy, only with CHIP, or only with heteroplasmy in (A) UKB and (B) ARIC. The prevalence of heteroplasmy in different CHIP gene subsets was evaluated using multivariable logistic regression without multiple testing correction in (C) UKB (CHIP vs. no CHIP, $P < 2e-16$; VAF $\geq 20\%$ vs. $2\% \leq \text{VAF} < 20\%$, $P < 2e-16$; multiple mutations vs. single mutation, $P = 1.05e-7$; and spliceosome mutated CHIP vs. other CHIP, $P < 2e-16$) and in (D) ARIC (CHIP vs. no CHIP, $P = 5.5e-3$; VAF $\geq 20\%$ vs. $2\% \leq \text{VAF} < 20\%$, $P = 0.0887$; multiple mutations vs. single mutation, $P = 0.0376$; and spliceosome mutated CHIP vs. other CHIP, $P = 0.0528$). The absolute number of individuals with heteroplasmy is indicated by the numbers on the left side of the bars. The prevalence of multiple heteroplasms within individuals with heteroplasmy across different CHIP gene subsets was evaluated using multivariable logistic regression without multiple testing correction in (E) UKB (CHIP vs. no CHIP, $P < 2e-16$; VAF $\geq 20\%$ vs. $2\% \leq \text{VAF} < 20\%$, $P = 3.11e-13$; multiple mutations vs. single mutation, $P = 3.22e-5$; and spliceosome mutated CHIP vs. other CHIP, $P = 1.86e-8$) and in (F) ARIC (CHIP vs. no CHIP, $P = 0.621$; VAF $\geq 20\%$ vs. $2\% \leq \text{VAF} < 20\%$, $P = 0.116$; multiple mutations vs. single mutation, $P = 0.407$; and spliceosome mutated CHIP vs. other CHIP, $P = 0.125$). The absolute number of individuals with multiple heteroplasms is indicated by the numbers on the left side of the bars. mMSS of different CHIP subsets in individuals with heteroplasmy was compared using multivariable linear regression without multiple testing correction in (G) UKB (number of individuals used to derive statistics; median mMSS [Q1, Q3] no CHIP: $n = 111,524$; 0.020 [0.000, 0.157]; CHIP: $n = 11,445$; 0.052 [0.006, 0.293]; $2\% \leq \text{VAF} < 20\%$: $n = 9,131$; 0.040 [0.005, 0.248]; VAF $\geq 20\%$: $n = 2,314$; 0.128 [0.014, 0.545]; single mutation: $n = 10,311$; 0.046 [0.005, 0.275]; multiple mutations: $n = 1,134$; 0.123 [0.013, 0.541]; *DNMT3A*: $n = 3,963$; 0.061 [0.007,

0.306]; *TET2*: $n = 2,434$; 0.068 [0.007, 0.376]; *ASXL1*: $n = 761$; 0.111 [0.013, 0.529]; *PPM1D*: $n = 255$; 0.071 [0.008, 0.328]; *TP53*: $n = 232$; 0.071 [0.007, 0.360]; *SRSF2*: $n = 183$; 0.334 [0.016, 0.763]; *SF3B1*: $n = 130$; 0.183 [0.016, 0.522]; *U2AF1*: $n = 62$; 0.309 [0.036, 0.717]. CHIP vs. no CHIP, $P < 2e-16$; VAF $\geq 20\%$ vs. $2\% \leq \text{VAF} < 20\%$, $P < 2e-16$; multiple mutations vs. single mutation, $P = 9.36e-16$; and spliceosome mutated CHIP vs. other CHIP, $P = 9.28e-16$) and in (H) ARIC (number of individuals used to derive statistics; median mMSS [Q1, Q3] no CHIP: $n = 2027$; 0.033 [0.000, 0.191]; CHIP: $n = 132$; 0.074 [0.009, 0.502]; $2\% \leq \text{VAF} < 20\%$: $n = 96$; 0.056 [0.008, 0.436]; VAF $\geq 20\%$: $n = 36$; 0.103 [0.010, 0.587]; single mutation: $n = 110$; 0.067 [0.007, 0.506]; multiple mutations: $n = 22$; 0.162 [0.013, 0.444]; *DNMT3A*: $n = 43$; 0.154 [0.009, 0.405]; *TET2*: $n = 38$; 0.065 [0.012, 0.687]; *ASXL1*: $n = 8$; 0.461 [0.068, 0.773]; *PPM1D*: $n = 0$; NA [NA, NA]; *TP53*: $n = 2$; 0.127 [0.065, 0.188]; *SRSF2*: $n = 0$; NA [NA, NA]; *SF3B1*: $n = 5$; 0.066 [0.000, 1.040]; *U2AF1*: $n = 3$; 0.013 [0.010, 0.440]. CHIP vs. no CHIP, $P = 2.3e-7$; VAF $\geq 20\%$ vs. $2\% \leq \text{VAF} < 20\%$, $P = 0.368$; multiple mutations vs. single mutation, $P = 0.603$; and spliceosome mutated CHIP vs. other CHIP, $P = 0.976$). Graphical representation of mMSS was truncated at mMSS = 2. All values of mMSS were included for the analyses. For all analysis, two-sided tests were used. Abbreviations: CHIP, clonal hematopoiesis of indeterminate potential; DDR, DNA damage response; mMSS, modified mitochondrial local constraint (MLC) score sum; and VAF, variant allele fraction. DTA includes *DNMT3A*, *TET2*, *ASXL1*. * Indicates the comparison between spliceosome mutations and other CHIP genes. In the UKB, these analyses were adjusted for age modeled as a restricted cubic spline, sex, smoking status, and a history of cancer. In ARIC, these analyses were adjusted for age modeled as a restricted cubic spline, sex, and smoking status. For the analyses including mMSS, the models were also adjusted for heteroplasmy count. Source data are provided as a Source Data file.

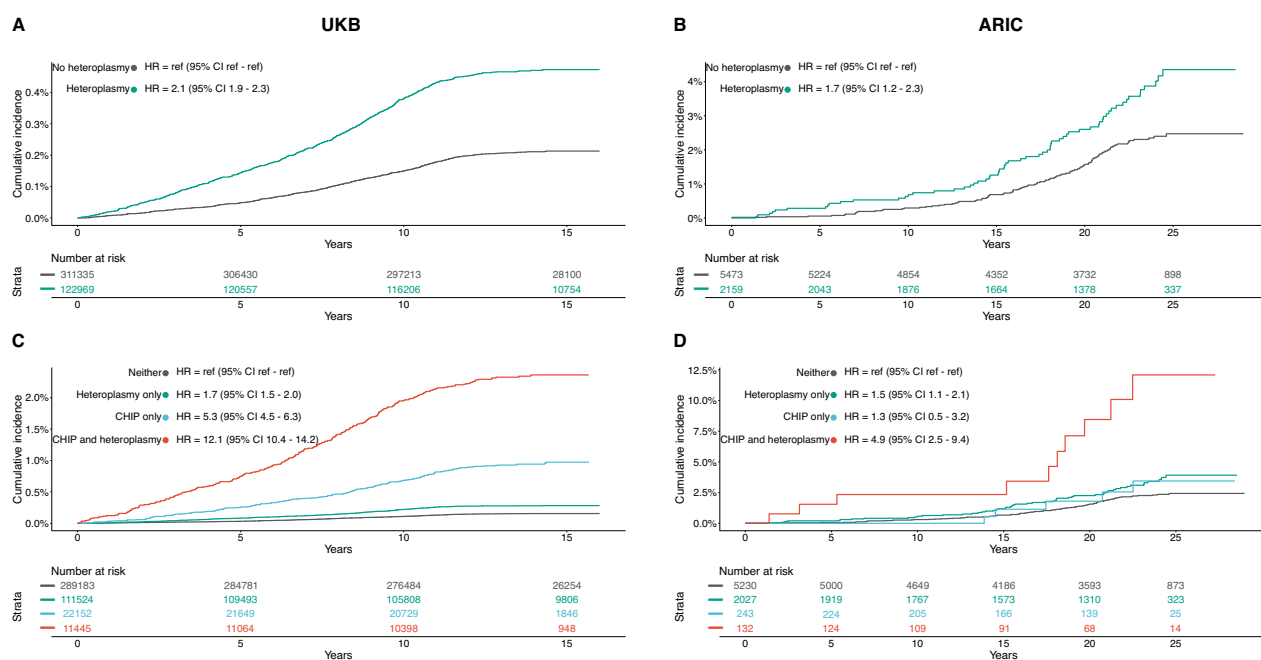


Fig. 4 | Risk of MN incidence based on CHIP and heteroplasmy status. Kaplan-Meier curves and hazard ratios from the adjusted Cox proportional hazards models comparing the risk of MN development between individuals with heteroplasmy and those without heteroplasmy in (A) UKB ($n = 434,304$) and (B) ARIC ($n = 7632$), and those with heteroplasmy only, those with CHIP only and those with both CHIP and heteroplasmy in (C) UKB and (D) ARIC. In the UKB, these analyses were adjusted for

age modeled as a restricted cubic spline, sex, smoking status and the presence of prevalent cancer. In ARIC, these analyses were adjusted for age modeled as a restricted cubic spline, sex and smoking status. Abbreviations: CHIP, clonal hematopoiesis of indeterminate potential; CI, confidence interval; HR, hazard ratio; and ref, reference. Source data are provided as a Source Data file.

(9.5–11.5; $n = 29,224$), intermediate (12–14; $n = 4,009$), and high (14.5–18.5; $n = 364$). The cumulative incidence of MN at 10 years was 0.4%, 3.1%, and 30.7% in each risk group, respectively. Compared to the low-risk group, the intermediate- and high-risk groups had an increased risk of incident MN (HR 7.2; 95% CI 5.8–9.0; $P < 0.0001$; and HR 87.1; 95% CI 68.8–110.6; $P < 0.0001$, respectively) after adjusting for sex, smoking status, and a history of cancer. Compared to the CHRS,

CHRS-M resulted in reclassification of 1395 individuals (4.6%, 22 cases) from low to intermediate risk and 127 (4.5%, 31 cases) from intermediate to high risk, resulting in a 34.4% increase of incident cases identified as high-risk individuals ($n = 90$ using CHRS vs. $n = 121$ using CHRS-M) (Fig. 5C–E). Including parameters for heteroplasmy in the prediction model significantly improved the discrimination assessed by net reclassification index (NRI 9.8; 95% CI 6.2–13.6; NRI in cases 14.0;

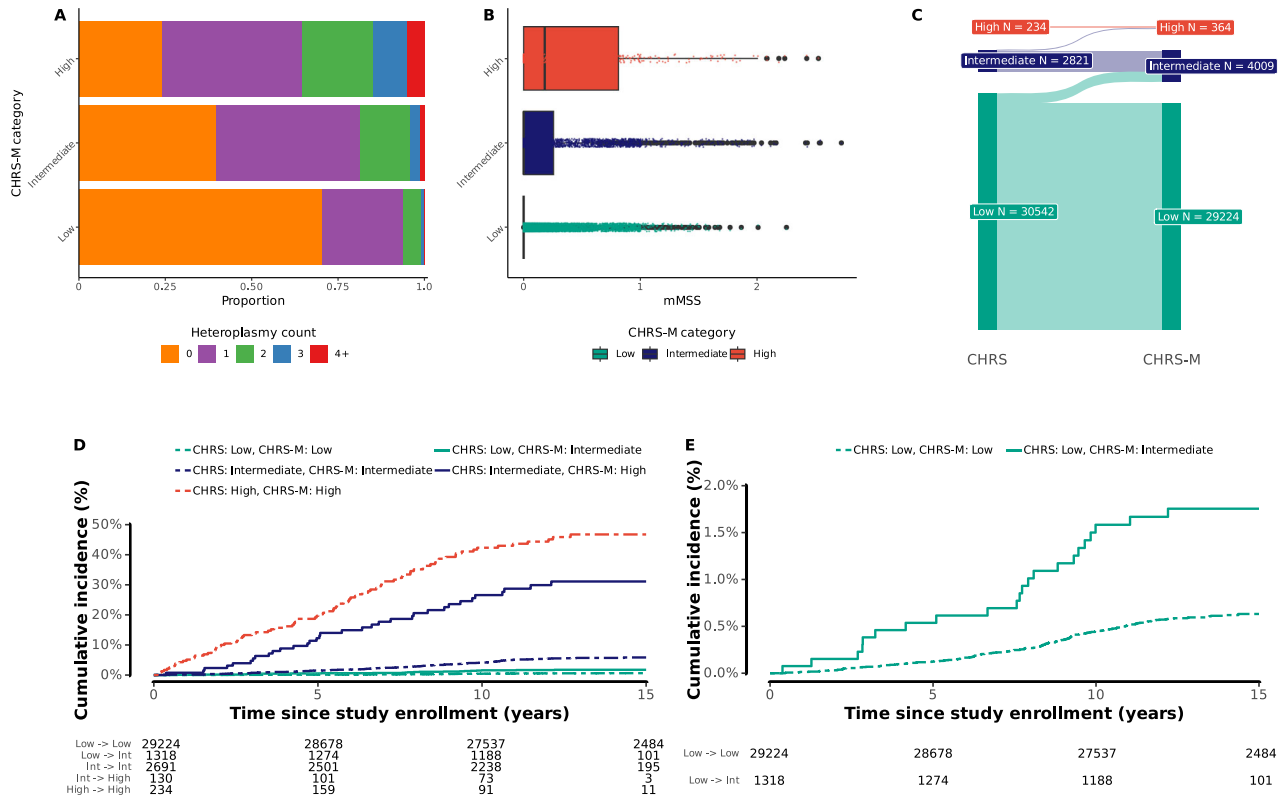


Fig. 5 | Distribution of heteroplasmy and risk of MN by CHRS-M categories. **A** Proportions of heteroplasmy count (0, 1, 2, 3, and 4+) within each CHRS-M category (low: $n = 29,224$; intermediate: $n = 4,009$; and high: $n = 364$; $P < 2e-16$ using Chi-square test, two-sided). **B** Boxplots and scatterplots of the distribution of mMSS by CHRS-M category (low: $n = 29,224$; intermediate: $n = 4,009$; and high: $n = 364$; $P < 2e-16$ using ANOVA, two-sided). The vertical lines within the box indicate the 25th (Q1), 50th (Q2, center), and 75th (Q3) percentiles of the distribution. The whiskers indicate the $Q1 - 1.5 \times (Q3 - Q1)$ and $Q1 + 1.5 \times (Q3 - Q1)$. Values outside the range of the whiskers (outliers) are displayed as dots. **C** A Sankey diagram of the reclassification of CHRS to CHRS-M. **D** Kaplan-Meier curves for the risk of MN

development by CHRS-M category ($n = 33,597$). The dotted lines indicate the cumulative incidence of individuals who remain in the same category. The solid lines indicate the cumulative incidence of individuals who are recategorized from low to intermediate (green) risk and from intermediate to high (dark blue) risk categories using the CHRS-M. **E** Kaplan-Meier curves for the risk of MN development in the CHRS low-risk group ($n = 30,542$) with the same data as (**B**) on an enlarged y-axis. Abbreviations: CHRS, clonal hematopoiesis risk score; CHRS-M, clonal hematopoiesis risk score with mitochondrial heteroplasmy; and mMSS, modified mitochondrial local constraint (MLC) score sum. Source data are provided as a Source Data file.

95% CI 10.5–17.7; NRI in controls –4.2; 95% CI –4.4 to –4.0), indicating that adding heteroplasmy information to the existing CHRS score improves the sensitivity of identifying those who develop MN, and particularly those are at a high risk, with a small decrease in specificity. The area under the receiver operating characteristic

(AUROC) at 10 years was 83.38 (95% confidence band: 80.63, 86.14) and 84.00 (95% confidence band: 81.24, 86.75) for CHRS and CHRS-M, respectively (P value for the difference between the two scores: 0.19). The Brier scores at 10 years was 0.00944 and 0.00928 for CHRS and CHRS-M (bootstrapped 95% confidence interval for the difference: 0.000006, 0.00027), suggesting an improved prediction accuracy for CHRS-M.

When we performed analysis by subtypes of MN, higher CHRS-M category was associated with a higher risk of developing all subtypes of MN, with the associations particularly strong for MPN and MDS (Fig. 6, Supplementary fig. 13). More specifically, the HR (95% CIs) for intermediate and high-risk groups were 4.74 (3.28–6.85) and 56.99 (38.13–85.19) for AML; 10.04 (6.75–14.94) and 99.30 (62.74–157.16) for MPN; and 9.67 (6.43–14.54) and 90.53 (57.21–143.26) for MDS, respectively.

The role of heteroplasmy on myeloid malignancy in individuals without CHIP

We further evaluated the association between heteroplasmy and MN in individuals without CHIP in the UKB ($n = 400,707$). Among individuals without CHIP, there were 734 incident cases of MN during a median (IQR) follow-up of 13.8 (13.1, 14.5) years. Both the number of heteroplasms and mMSS were associated with a higher risk of MN (HR = 1.5; 95% CI 1.4–1.6; $P < 0.0001$ for a 1-unit increase in heteroplasmy count and HR = 3.1; 95% CI 2.5–3.8; $P < 0.0001$ for a 1-unit increase in mMSS)

Table 3 | Values assigned to component variables of the CHRS⁹ modified for heteroplasmy

Variables	0.5	1	1.5	2	2.5
Age (years)		< 65	≥ 65		
Cytopenia		CHIP	CCUS		
RDW		< 15			≥ 15
MCV		< 100			≥ 100
High-risk mutation		Absent			Present
Single DNMT3A	Present	Absent			
Number of mutations		1		≥ 2	
VAF		< 20%		≥ 20%	
Heteroplasmy		Absent	Present		
mMSS		< 0.51	0.51–1.52	≥ 1.52	

CCUS clonal cytopenia of undetermined significance, CHIP clonal hematopoiesis of indeterminate potential, MCV mean corpuscular volume, mMSS modified mitochondrial local constraint (MLC) score sum, RDW red cell distribution width, and VAF variant allele frequency.

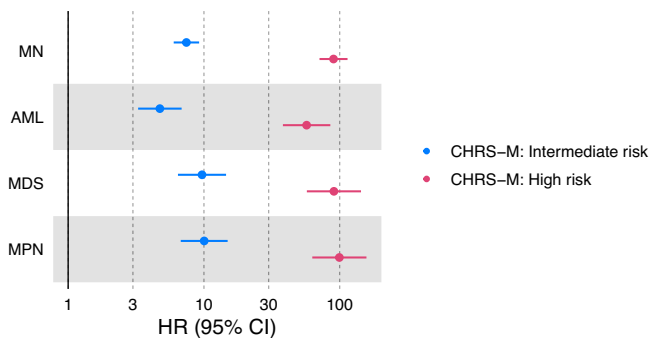


Fig. 6 | Hazard ratios (95% confidence intervals) for the associations of CHRS-M categories with subtypes of MN. Hazard ratios (95% confidence intervals) for the associations of CHRS-M categories with subtypes of MN. The analysis was restricted to UKB participants with CHIP. Hazard ratios (the point) and 95% confidence intervals (error bars) were estimated using Cox proportional hazards models with the low-risk group as the reference category (low: $n = 29,224$; intermediate: $n = 4,009$; and high: $n = 364$). All models are adjusted for age modeled as a restricted cubic spline, sex, smoking status, and a history of cancer. Abbreviations: AML, acute myeloid leukemia; CI, confidence interval; CHRS-M, clonal hematopoiesis risk score with mitochondrial heteroplasmy; HR, hazard ratio; MDS, myelodysplastic syndrome; MN, myeloid neoplasms; and MPN, myeloproliferative neoplasms. Source data are provided as a Source Data file.

when adjusted for age, sex, smoking status, and a history of cancer. The associations were attenuated when we assessed the independent associations of heteroplasmy count (HR = 1.2; 95% CI 1.1–1.4; $P < 0.01$) and mMSS (HR = 2.3; 95% CI 1.7–3.1; $P < 0.0001$) with MN by mutually adjusting for each other. Heteroplasmy count (HR = 1.2; 95% CI 1.1–1.4; $P < 0.01$) and mMSS (HR = 2.2; 95% CI 1.6–3.0; $P < 0.0001$) were independent risk factors of MN even after further adjusting for blood profiles that are potential biomarkers of MN, including RDW, MCV, and presence of cytopenia, suggesting that mtDNA heteroplasmy is an independent predictor of MN risk also in people without CHIP.

Discussion

The role of somatic nuclear DNA mutation, and CHIP in particular, has been increasingly recognized as a key risk factor for developing MN. Much less is known about the role of somatic mutations in the mitochondrial genome, with recent work identifying a strong association between mitochondrial heteroplasmy and hematological cancers¹⁵, and the interaction between nuclear and mitochondrial mutations²³. In this study, we found a significant enrichment of participants harboring both CHIP variants and mitochondrial heteroplasmy, with compelling evidence that this enrichment differs based on the specific natures of the nuclear and mitochondrial mutations. We further found that mitochondrial heteroplasmy can significantly improve a CHIP-based risk score for the development of MN and may help identify those at risk in non-CHIP individuals, in whom no current risk scores are currently available.

The high prevalence of heteroplasmy (approximately 30% in both UKB and ARIC), and the observation that a notable number of heteroplasms measured in peripheral blood are somatic, suggest that they can serve as markers of clonal expansion^{15,20–22}. This observation is corroborated by the higher prevalence of heteroplasmy in individuals with CHIP, particularly with VAF $\geq 20\%$, compared to those without CHIP. Somatic mitochondrial mutations have also been demonstrated to play an important role in tumorigenesis by increasing the level of reactive oxygen species (ROS) and reducing apoptosis^{24–26}. Moreover, certain tumors have been found to significantly favor the selection of deleterious mitochondrial mutations^{24,27}. Consistent with these results, we found that heteroplasmic variants predicted to be more deleterious were enriched in individuals with high-risk CHIP, particularly in

individuals with significantly expanded clones (VAF $\geq 20\%$), with more than one mutation, and with mutations in spliceosome machinery^{4,9,28–30}. These associations remained significant even after adjusting for the number of heteroplasms, supporting the claim that the functional nature of the mitochondrial mutations is important, and does not simply represent passenger mutations due to CHIP.

An important aim of our study was to determine whether the presence of heteroplasmy was merely a surrogate marker for CH or whether it played a potential causal role in the development of MN. By incorporating both mMSS and heteroplasmy count in the regression model and demonstrating that mMSS remained significantly associated with MN incidence, we provide statistical evidence suggesting a causal role of deleterious mtDNA variation in MN pathogenesis. This potential causality is further corroborated by our finding that there is a statistically significant synergistic (non-multiplicative) effect of heteroplasmy and CHIP on the risk of MN. Thus, beyond simply serving as a marker of CH, the specific functional consequence of the heteroplasmy itself may play a role in tumorigenesis, consistent with the extensive roles that mitochondria have been shown to play in cancer^{31–34}. Indeed, heteroplasmy has been found to be an adverse prognostic factor in patients diagnosed with MN^{35–37}. However, experimental studies to support our findings are needed to provide direct evidence for the causal role of heteroplasmy on MN development.

In the current study, CHIP was evaluated in the context of conventional CHIP variants. Nonetheless, it has been shown that this approach misses a part of the clonality present in the hematopoietic system^{38,39}. Thus, in future studies, we aim to expand on the evaluation of hematopoietic clonality and assess it via a barcode-based approach³⁹.

From a clinical perspective, the ability to identify high-risk MN individuals early is key to improving outcomes. Incorporating heteroplasmy information into a state-of-the-art CHIP-based prediction model significantly improved the sensitivity of identifying individuals at risk of developing MN, with little loss of specificity. Specifically, integrating the presence (heteroplasmy yes/no) and predicted deleterious effect of heteroplasmic variants, the updated model was able to detect 34.4% more incident cases among individuals classified as high-risk (10-year risk $\geq 10\%$) of developing MN. This is a remarkable improvement to the current score system, without the need for recalibrating or retraining the algorithm. Clinically, this finding suggests that heteroplasmy information can refine the existing risk stratification and identify high-risk individuals (10-year risk $\geq 10\%$) who have been inappropriately classified as intermediate risk (CHRS 10-year risk of 2–10%), and who may benefit from close surveillance. Moreover, mitochondrial DNA sequencing and detection of heteroplasmy can be feasibly implemented in currently used molecular tests to facilitate the identification of this high-risk population.

While CHIP has great clinical utility, particularly when combined with heteroplasmy, the risk of MN is particularly not well understood among individuals without evidence of CHIP. Our study demonstrates that having any mitochondrial heteroplasmy was associated with a higher risk of MN, even after accounting for known risk factors in individuals without CHIP. Notably, higher mMSS was a more robust predictor for MN than just the number of heteroplasms, further suggesting that mutation in mtDNA may be an independent causal mechanism for MN. However, MN is a rare disease, with a crude incidence rate of 14 cases per 100,000 person-years in the general population⁴⁰, making it particularly difficult to precisely estimate its risk and study novel biomarkers. Consequently, a risk stratification model applicable to the general population is currently unavailable. Further studies are needed to better understand the underlying mechanisms linking mtDNA variation to MN incidence and to develop a universal risk model for MN to identify high risk population regardless of CHIP status.

Our findings strongly suggest that heteroplasmy has a functional role in the development of MN. We, thus, expect mitochondrial mutations to also have a role in the prognosis and risk-stratification of MN. Indeed, mtDNA heteroplasmy showed a prognostic value in allogeneic hematopoietic cell transplantation outcomes for MDS patients³⁶.

The current manuscript has several limitations. First, as there are differences between UKB and ARIC and between each cohort and the population from which they were sampled. Nonetheless, the similarity of the results between the two cohorts only strengthens the validity of our observations. Secondly, the use of WES to detect CHIP has inherent limitations related to the relatively shallow sequencing depth, which can lead to an increase in false negatives. In addition, the UKB is mostly of self-reported White individuals and the subgroup analysis of non-White individuals was based on a small number of events (34 events out of 23,991 individuals). Thus, the findings of the interaction between heteroplasmy and CHIP, as well as the CHRS-M need to be validated in more diverse populations. Despite these limitations, our findings suggest that heteroplasmy, in addition to being a marker of clonal expansion, may be a causal biomarker of MN development, with clinical utility in the general population.

Methods

The current study was approved by the Johns Hopkins Medicine Institutional Review Boards. The ARIC study protocol was approved by the Johns Hopkins Medicine Institutional Review Board (IRB), which serves as the single site IRB.

Study population

The UK Biobank (UKB) is a large population-based prospective study of ~500,000 participants between the ages of 40 to 69 years across the United Kingdom recruited from 2006 to 2010⁴¹. The UKB has extensive information on participant's genetic and phenotypic data, including demographics and lifestyle factors. Information on sex was based on self-report. The data is linked to the death registry, cancer registry, hospital admissions, and primary care visit data.

For the current study, 434,304 participants who had both WGS, which was used for calling mitochondrial DNA heteroplasmy, and WES, which was used for CHIP variant calling, and passed the variant-level and sample-level QC for both and met inclusion / exclusion criteria were analyzed (details below; Supplementary fig. 14). More specifically, there were 490,355 participants with WGS, where mtDNA heteroplasmy was evaluated using MitoHPC¹⁸ (<https://github.com/dpuui/MitoHPC>). We excluded samples ($n = 3776$) if they met one of the following criteria: suspicious of having 1) potential mitochondrial contamination (a contamination level $\geq 3\%$ from Haplocheck⁴²), 2) 2 or more variants belonging to a different mitochondrial haplogroup, 3) 2 or more variants matching to the same NUMT, or 4) low minimum base coverage (<100) or low mean base coverage (<500). In addition, because the presence of NUMTs can influence the false positive mtDNA heteroplasmy calls at low mtDNA copy number (mtDNA-CN)⁴³, we removed 12,001 participants with mtDNA-CN ≤ 40 . We additionally removed 704 participants with a heteroplasmy count above 5 because we found that samples with a heteroplasmy count above 5 are not distinguishable from contaminated samples¹⁵. Some samples met multiple exclusion criteria, which resulted in 13,921 participants being excluded. After exclusion, there were 476,434 participants with heteroplasmy information. For WES, 466,042 participants had both CHIP calls from GATK Mutect2⁴⁴ and *U2AF1* calls. No additional participants were excluded for having a high number of INDELs or CHIP variants. Of the 450,916 participants who passed QC for heteroplasmy and CHIP, we excluded 463 participants with a history of MN ($n = 310$) or a potential MPN ($n = 182$). We further excluded 16,149 participants who did not have information on tobacco smoking status ($n = 2211$) or have missing values for any one of the following measurements that are used for

calculating the clonal hematopoiesis risk score (CHRS; $n = 14,077$): hemoglobin, platelet count, neutrophil count, red blood cell distribution width (RDW), or mean corpuscular volume (MCV). A detailed description of QC steps for heteroplasmy and CHIP, and the exclusion for heteroplasmy count and potential MPN are provided below. The final sample for analysis included 434,304 individuals (199,046 men and 235,258 women).

The Atherosclerosis Risk in Communities (ARIC) study is a community-based, prospective cohort study focusing on the risk factors for cardiovascular disease, that recruited 15,792 individuals between the ages of 45 and 64 from 4 communities in the US (Forsyth County, NC; Jackson, MS; Minneapolis suburbs, MN; and Washington County, MD) from 1987 to 1989⁴⁵. Of those, 12,776 had WGS available, on which MitoHPC was run (Supplementary fig. 15). After excluding participants that failed heteroplasmy QC and those that did not have visit information ($n = 69$), we retained 12,707 individuals with heteroplasmy information. We further excluded individuals with those that did not have WES data from the same visit ($n = 4340$), that did not have information on incident myeloid neoplasms ($n = 709$), that were likely prevalent MPNs ($n = 0$), that had an excessive number (≥ 4) of CHIP indels ($n = 3$) and that had 10 or more CHIP mutations ($n = 1$) and did not have information on smoking status ($n = 22$). The final sample for analysis included 7632 individuals (3453 men and 4179 women). Information on sex was based on self-report. All participants included in this study from both UKB and ARIC provided appropriate informed written consent.

Heteroplasmy and mitochondrial DNA copy number analysis in the UK Biobank

Samples for DNA were collected on the baseline visit and DNA was extracted from buffy coat using the Maxwell[®] 16 Instrument (Promega) and the Maxwell[®] 16 Blood DNA Purification Kit (Promega-AS1010X). WGS CRAM files from the UKB were processed on the DNA Nexus server. For heteroplasmy identification, we used the MitoHPC pipeline (version 20230418; all default settings with default random down-sampling to use at most 222 K reads)¹⁸, implementing GATK Mutect2 for variant identification^{44,46}. We defined a heteroplasmic SNV at a variant allele frequency of 5%, meaning that variant alleles at a frequency of 5–95% within an individual are defined as heteroplasmic. Alleles less than 5% or greater than 95% are counted as homoplasmic. To test the robustness of this cutoff, we repeated the main analyses with VAF cutoffs at 3% and 10%. The results were consistent across different cutoffs (Supplementary fig. 16). MitoHPC pipeline incorporates haplogrep (<https://github.com/seppinho/haplogrep-cmd/v2.4.0>)⁴⁷ to identify haplogroups and Haplocheck (<https://github.com/genepi/haplocheck>)⁴² to detect in-sample contamination by detecting two different mitochondrial haplotypes in each sample. For mitochondrial DNA copy number (mtDNA-CN) calculation, we used SAMtools⁴⁸ embedded in MitoHPC to generate read count and coverage information, using the command 'samtools idxstats'. A detailed documentation on how to run MitoHPC on DNA Nexus server is available: https://github.com/ArkingLab/MitoHPC/blob/main/docs/DNAnexus_CLOUD.md.

All 490,355 WGS samples processed using MitoHPC variant calling were successfully completed¹⁸. The mean nuclear genomic coverage for WGS samples in the UKB was 98x and the median (range) mtDNA coverage was 904x (500x–1887x).

Heteroplasmy and mitochondrial DNA copy number analysis in ARIC

We analyzed WGS data from the Atherosclerosis Risk in Communities (ARIC) study ($n = 12,776$). DNA samples were collected for each participant across multiple clinic visits (V1, V2, V3, V4, MRI and V5) and DNA for WGS were isolated from buffy coat using the Gentra Puregene Blood Kit (Qiagen). WGS calls were from the Trans-Omics for Precision

Medicine (TOPMed; <https://topmed.nhlbi.nih.gov/methods>) program are from freeze 8 (30.6%; $n = 3915$) and from the Centers for Common Disease Genomics (CCDG; <https://www.genome.gov/27563570>) initiative (69.4%; $n = 8861$). There were 11 individuals who were sequenced in both TOPMed and CCDG, and we randomly selected one of the two. TOPMed studies provide WGS data at $\sim 30\times$ genomic coverage using Illumina next-generation sequencing technology, which must pass specific quality control metrics before being released for use by the scientific community. The median (range) mtDNA coverage was 1661x (612x–1898x).

Heteroplasmy filtering for variants

MitoHPC provides various variant annotations to evaluate variant quality, allowing us to remove low quality variants prior to analysis. We excluded variants with read depth < 300 and those flagged as base quality, mapping quality, strand bias, slippage, weak evidence, position, clustered, fragment length, and haplotype flags in the FILTER column of the VCF. We further excluded heteroplasmic variants at poly-C homopolymer regions on the mitochondrial chromosome (the list is provided as HP.bed.gz in <https://github.com/dpuiiu/MitoHPC>). INDELs are also excluded because they are often found in homopolymer regions, where it is challenging to accurately call heteroplasmies¹⁸.

Mitochondrial local constraint (MLC) score

The mitochondrial local constraint (MLC) score is a metric reflecting local tolerance to substitutions of base or amino acid. It is calculated for every possible mtDNA single nucleotide variant (SNV) by applying a sliding window method¹⁹. In brief, starting from position m.1, within a window of 30 bases, the observed:expected (oe) ratio of substitutions and its 90% confidence interval (CI) was calculated in gnomAD⁴³. The window start position is then moved by 1 bp, and this process is iterated until the full length of mtDNA is covered (i.e., a start position of m.16569). For positions in protein genes, calculations are restricted to missense variants, or substitutions in amino acids, while for all other positions, all base substitutions are used. The mean of oe ratio 90% CI upper bound fraction (OEUF) is calculated for each position using all overlapping windows, and then percentile ranked to achieve a positional score between 0 and 1, where 1 is most constrained and 0 is least constrained. An MLC score is obtained for every mtDNA SNV as follows: non-coding, RNA, and missense variants are assigned their positional score; and non-missense variants in protein genes are assigned scores based on the OEUF value of the variant class, with synonymous, stop gain, and start/stop lost being assigned scores of 0.0, 1.0, and 0.70, respectively. Variants with higher scores are predicted to be more deleterious. To account for heteroplasmic variants that also present as homoplasmic in the population, we refined the MLC score to account for the frequency of homoplasmy in the UKB population by calculating modified MLC (mMLC) as $MLC/(1 + \log_{10}(\text{number of individuals with homoplasmy in the UKB} + 1))$. mMLC for both UKB and ARIC was calculated using the individuals with a given homoplasmic variant in the UKB. To capture the functional impact of multiple heteroplasmies, we calculated an mMLC score sum (mMSS) by summing all mMLC scores within a given individual.

CHIP variant calling

WES CRAM files were aligned to hg38 and hg19 in the UKB and ARIC, respectively. Variant calling was performed using Genome Analysis Toolkit (GATK) v.4.2.2. Mutect2^{44,49}. Mutect2 was run in ‘tumor-only’ mode using non-default parameters: `gatk Mutect2... --panel-of-normals ${ref_pon} --germline-resource ${ref_germ}`. Raw variants called by Mutect2 were filtered out with FilterMutectCalls using the estimated prior probability of a reading orientation artifact generated by LearnReadOrientationModel.

U2AF1 calling in UKB (hg38)

Considering the known duplication of *U2AF1* in hg38, we called these variants using a previously described pileup approach (https://github.com/weinstockj/pileup_region)⁵⁰. Briefly, reads containing *U2AF1* pathogenic variants and the total number of reads at the given locus were counted at both the *U2AF1* and *U2AF1L5* loci. The alternative depth and total depth were considered as the average between the two loci. VAF was calculated as the alternative depth divided by total depth and then multiplied by 100.

CHIP variant filtering

Variants were annotated using ANNOVAR⁵¹ and filtered based on: 1) the presence in a gene part of a custom CHIP panel (Supplementary Table 1); 2) a previously published whitelist⁵⁰; 3) Mutect2 FILTER of PASS, weak_evidence, or germline; 4) depth ≥ 20 ; 5) alternate allele count ≥ 3 in the case of SNVs; 6) alternate allele count ≥ 5 in the case of MNVs; 7) ALT FIR2 ≥ 1 ; 8) ALT F2R1 ≥ 1 ; 9) VAF $\geq 2\%$ in the case of SNVs; 10) VAF $\geq 10\%$ in the case of indels; 11) exclusion of synonymous SNVs; 12) the maximum allele frequency across non-cancer populations (non_cancer_AF_popmax) under 0.001 in the gnomAD exome collection v2.1.1⁵²; or 13) recurrent germline variants considered as such based on their presence in 3 or more individuals and having a binomial test P value ≥ 0.01 in $\geq 80\%$ of those individuals. Variants occurring on genes located on the X chromosome in males had their VAF divided by 2. For ARIC WES, given the higher noise, an alternate allele count ≥ 5 was used for SNVs. The higher noise seen in ARIC was reflected through the higher number of SNVs seen in the VCF per individual compared to UKB.

On a sample level, we excluded individuals with more than 3 mutations, all of which are represented by MNVs, individuals with more than 9 CHIP variants, or individuals with suspected of having potential MPNs, as described in Supplementary fig. 14 and 15.

Large CHIP clones were defined as those having a VAF $\geq 20\%$. Small CHIP clones were defined as having a VAF $\geq 2\%$ and $< 20\%$. Potential MPNs were defined as individuals with known pathogenic variants in *JAK2*, *CALR*, or *MPL*, and had a hematocrit over 48% or a platelet count over $450 \times 10^3/\mu\text{L}$. If either of these laboratory parameters were missing, the individual was excluded. Of note, although *CALR* was not included in the CHIP panel, it was assessed because *CALR* is a recurrently mutated gene in MPN. The same filtering steps were applied to identifying *CALR* mutation. However, the presence of *CALR* mutation was not used to define CHIP in the downstream analysis. The median (range) sequencing depth of CHIP variants in the final set was 70 (20–316) in the UKB and 95 (23–490) in ARIC.

CCUS was defined as the presence of CHIP and at least one cytopenia, where cytopenia was defined as: 1) anemia (hemoglobin concentration < 13 g/dL in men and < 12 g/dL in women); 2) thrombocytopenia (platelet count $< 150 \times 10^9$ cells/L); or 3) neutropenia (absolute neutrophil count $< 1.8 \times 10^9$ cells/L), similar to other literature^{9,53}.

Definition of MN

In the UK Biobank, MN and its subtypes were defined using a similar set of ICD-9 and ICD-10 codes as previous studies using UKB (Supplementary Table 5)^{8,9}. More specifically, MN was defined as the presence of acute myeloid leukemia (AML), chronic myeloproliferative disease / myeloproliferative neoplasms (MPN), myelodysplastic syndrome (MDS), or chronic and other myelogenous leukemia. We used ICD codes from the cancer registry linked to the UKB to define MN, whereas other studies have used either a combination of self-report and hospital data or electronic health records, leading to some differences in the number of incident cases of MN. Because the national cancer registries centralize information received from separate regional cancer centers around the UK, the definition of MN using cancer registry data has high sensitivity and specificity. In addition,

when we compared the association between heteroplasmy and MN using different sources (cancer registry and main diagnosis in hospital admissions data), the results were similar. Therefore, we used cancer registry data throughout our analysis.

In ARIC, hematological malignancy (HM) cases were ascertained through 12/31/2015 via linkage with cancer registries in the four states where the ARIC participants were recruited, and supplemented with medical records, routinely collected hospital discharge summaries, and death certificates⁵⁴. MDS before 2001 and MPN before 2011, both characterized by clonal expansion, were considered to be pre-leukemia; today, these are considered to be leukemia. These pre-leukemias were not captured by cancer registries before those dates. Thus, we used 2 strategies to identify MDS and MPN: 1) identified cases using ICD codes from routinely collected hospital discharge summaries, 2) identified cases using ICD codes from Centers for Medicare and Medicaid Services (CMS) claims data (Supplementary Table 6). Participants with an ICD code consistent with MDS, MPN, or other such state prior to or concurrent with the blood sample used to call CHIP from WES were excluded. Complete blood cell counts are measured as part of the ARIC protocol but were not available for all participants at their WES visit. For the participants with an eligible MDS or MPN ICD code, we reviewed their complete blood cell counts (if available) that were measured in the same blood sample used to call CHIP. We excluded any participant with blood count anomalies suggestive of an undiagnosed HM. Using the histology codes from the cancer registries and ICD codes from the hospital discharge summaries and claims data, we classified the cases as myeloid- or lymphoid-derived cases. Three experts in hematologic malignancies and epidemiology (L.P.G., M.R., S.P.) adjudicated all of these cases.

Statistical analysis

Categorical data was represented as absolute count (percent). Frequencies of two categorical variables were analyzed using Fisher's exact test. Continuous variables are presented as mean (SD) for variables with Gaussian distribution or median (25th, 75th percentiles) for variables with non-Gaussian distribution. Normality of the distribution was assessed using skewness, kurtosis and histogram visualization. Differences between two groups with non-Gaussian distribution were assessed using the Mann-Whitney-Wilcoxon rank sum test. Differences between two groups for variables with Gaussian distribution were assessed using Student's t-test. We used multivariable logistic regression models to evaluate the associations between CHIP and a binary variable for heteroplasmy. For the association between CHIP and mMSS, we used multivariable linear regression models. If the dependent variable was right skewed, we added 1 and log-transformed it. For UKB, these analyses were adjusted for age modeled as a restricted cubic spline (with 4 degrees of freedom), sex, tobacco smoking status (never or ever), and a history of cancer. For ARIC, the analyses were adjusted for age modeled as a restricted cubic spline (with 4 degrees of freedom), sex, and cigarette smoking status (never or ever). History of cancer was not included as a covariate in ARIC as the follow-up duration since the WGS / WES visit was only available for those who did not have cancer prior to or at the WGS / WES visit.

For survival analysis, we used time from DNA collection to the development of MN, death, or end of follow-up (administrative censoring on December 22, 2022 in the UKB, and December 31, 2015 in ARIC), whichever occurred first. Presence of heteroplasmy in combination with CHIP were evaluated for their associations with incidence of MN. Cumulative incidence of MN was presented as Kaplan-Meier curves. Hazard ratios and corresponding 95% confidence intervals were estimated using a multivariable Cox proportional hazards model, adjusting for age (restricted cubic splines with 4 degrees of freedom), sex, tobacco smoking (never or ever), and a history of cancer. We have also evaluated the association between metrics of heteroplasmy

(heteroplasmy count as categorical and continuous variables, and mMSS) in all participants and in participants with heteroplasmy, separately, in the UKB using the same covariate adjustment. A stratified analysis by CHIP status (CHIP-positive and CHIP-negative) was also performed. In ARIC, as the analysis was restricted to individuals without a history of cancer, adjustment was made for age at the time of DNA collection (restricted cubic splines with 4 degrees of freedom), sex, and cigarette smoking status (never or ever) at the time of DNA collection. We performed several sensitivity analyses to demonstrate the robustness of findings. Due to the limited number of events in ARIC, sensitivity analyses were only performed in UKB. To address the potential differences in the association between mMSS and health outcomes by self-reported race/ethnicity¹⁵, we performed an analysis restricted to self-reported White ($n = 410,313$) and non-White ($n = 23,991$) individuals, separately. Race/ethnicity was defined using the Ethnic background (Data-Field 21000) variable. As the UKB includes related individuals, we repeated the analysis restricted to genetically unrelated individuals ($n = 362,132$; defined as variable "used.in.pca.calculation" as provided by the UKB). In addition, because adjusting for smoking status simply as a binary variable may be subject to residual confounding³⁹, we further adjusted for pack-years of smoking in an analysis using former or current smokers ($n = 131,562$). We additionally performed an analysis restricted to never smokers ($n = 238,086$). Furthermore, we evaluated the association of heteroplasmy, CHIP, and subtypes of MN (AML, MDS, and MPN), separately. Finally, we repeated the analysis after excluding participants with a history of cancer ($n = 403,732$). For this analysis, adjustment was made for age (restricted cubic splines with 4 degrees of freedom), sex, and smoking (never or ever).

Development of a prediction model incorporating mitochondrial heteroplasmy

After determining that mitochondrial heteroplasmy is a predictor for MN, we evaluated whether information on heteroplasmy can improve the prediction of MN using an existing prediction model for MN in individuals with CHIP, the Clonal Hematopoiesis Risk Score (CHRS)⁹. To build upon the CHRS model for CHIP-positive individuals, we first generated the CHRS and CHRS categories by assigning the scores for 8 components (age ≥ 65 years, presence of cytopenia, red cell distribution width (RDW) ≥ 15 , mean corpuscular volume (MCV) ≥ 100 , presence of high-risk mutation, single *DNMT3A* mutation, number of mutations, and VAF $\geq 20\%$) from the original CHRS⁹. We then added various metrics for heteroplasmy (heteroplasmy count, presence of heteroplasmy, mMSS, heteroplasmy count and mMSS, and presence of heteroplasmy and mMSS) separately to a Cox proportional hazards model with the 8 components of CHRS, sex, smoking (never or ever), and a history of cancer, and tested for model fit using log-likelihood and Akaike information criterion (AIC). We selected the most parsimonious model with the best model fit, which included the presence of heteroplasmy (Yes/No) and mMSS as a continuous variable. We developed a score, CHRS-M, by assigning scores for the presence of heteroplasmy and mMSS by rounding the corresponding coefficients from the final model to the nearest 0.5 and adding 1. The coefficient for presence of heteroplasmy was 0.32 and individuals with any heteroplasmy were assigned a score of 1.5. To apply the scoring algorithm to a linear association of mMSS with MN, we estimated the levels of mMSS that correspond to log(HR) of 0.25, 0.75, and 1.25, so that bins of mMSS will be assigned a score of 1, 1.5, 2, and 2.5, respectively (Supplementary fig. 11). The estimated mMSS values were 0.51, 1.52, and 2.54, and thus were divided into 4 categories: < 0.51 , $0.51-1.52$, $1.52-2.54$, ≥ 2.54 . However, there was only 1 individual in the ≥ 2.54 category and, therefore, was collapsed into 3 groups (< 0.51 , $0.51-1.52$, ≥ 1.52) for final scores. We defined low-, intermediate-, and high-risk categories of CHRS-M by adding 2 to the cut-offs of CHRS; low (≤ 11.5), intermediate ($12-14$), and high (≥ 14.5). We evaluated the 10-year cumulative

incidence of MN by CHRS-M category using Fine-Gray method to account for competing risk due to death³⁵. We also performed Cox proportional hazards models to compare the risk of MN between CHRS-M categories. We evaluated the change in sensitivity and specificity compared to CHRS by using net reclassification index (NRI; R package *nricens*) (<https://cran.r-project.org/web/packages/nricens/index.html>). The AUROC for time-to-event data and the difference in the AUROC between CHRS and CHRS-M were estimated using an R package *timeROC* (<https://cran.r-project.org/web/packages/timeROC/index.html>). The difference in the Brier scores was estimated using 500 bootstrap samples. Furthermore, the HRs (95% CI) for the associations of CHRS-M with incident MN and its subtypes (AML, MDS, and MPN) were estimated using Cox proportional hazards models adjusting for age (restricted cubic splines with 4 degrees of freedom), sex, tobacco smoking (never or ever), and a history of cancer. Similarly, we evaluated the associations of each component of CHRS-M with incident MN and its subtypes by including all parameters CHRS-M, sex, smoking, and a history of cancer in the same model.

In individuals without CHIP, we estimated the HRs for the association between parameters of heteroplasmy and MN using a Cox proportional hazards model adjusting for age (restricted cubic splines with 4 degrees of freedom), sex, smoking (never or ever), and a history of cancer. We first included heteroplasmy count and mMSS in separate models, and to evaluate whether they are independently associated with MN, we then included heteroplasmy count and mMSS in the same model. Finally, we additionally adjusted for other potential biomarkers of MN, including RDW, MCV, and presence of cytopenia.

Reporting summary

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

Data availability

UK Biobank data is available through application to the UK Biobank (Application Number 17731). Access to the UK Biobank data can be requested at <https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>. ARIC data are available through the NIH database of genotypes and phenotypes (dbGaP) accession number phs001211. Due to restrictions based on privacy regulations and informed consent of the participants, the ARIC data are available under controlled access through the dbGaP. Access to ARIC data requests can be submitted to the dbGaP steering committee with an IRB approval, which are generally reviewed within 2 weeks for confidentiality or intellectual property restrictions by the Data Access Committee. Once approved, investigators are allowed one year of access to the data, which can be renewed or closed out after the one-year access period. Details of the data access request process is available in <https://sharing.nih.gov/accessing-data/accessing-genomic-data/how-to-request-and-access-datasets-from-dbgap>. Phenotypes, including cancer data, not available through dbGaP can be requested by submitting a proposal to ARIC, which will be reviewed by the Publications Committee. The remaining data are available within the Article, Supplementary Information or Source Data file. Source data are provided with this paper.

Code availability

Code for data cleaning and analysis is available on our GitHub repository: https://github.com/ArkingLab/Heteroplasmy_CHIP/ (<https://doi.org/10.5281/zenodo.14035924>)⁵⁶. Documentation on MitoHPC pipeline for DNA Nexus server is available in https://github.com/ArkingLab/MitoHPC/blob/main/docs/DNAxexus_CLOUD.md. Documentation on extracting Mitochondrial and NUMT reads from Google Cloud is available at https://github.com/ArkingLab/MitoHPC/blob/main/docs/GOOGLE_CLOUD.md. Source data are provided with this paper.

References

- Lin, J. S., Perdue, L. A., Henrikson, N. B., Bean, S. I. & Blasi, P. R. *Screening for Colorectal Cancer: An Evidence Update for the U.S. Preventive Services Task Force*. (Agency for Healthcare Research and Quality (US), Rockville (MD), 2021).
- Jonas, D. E. et al. *Screening for Lung Cancer With Low-Dose Computed Tomography: An Evidence Review for the U.S. Preventive Services Task Force*. (Agency for Healthcare Research and Quality (US), Rockville (MD), 2021).
- Melnikow, J., Fenton, J. J., Miglioretti, D., Whitlock, E. P. & Weyrich, M. S. *Screening for Breast Cancer With Digital Breast Tomosynthesis*. (Agency for Healthcare Research and Quality (US), Rockville (MD), 2016).
- Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
- Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
- Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
- McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
- Gu, M. et al. Multiparameter prediction of myeloid neoplasia risk. *Nat. Genet.* **55**, 1523–1530 (2023).
- Weeks, L. D. et al. Prediction of risk for myeloid malignancy in clonal hematopoiesis. *NEJM Evid* **2**, eidoa2200310 (2023).
- Lin, M. T. & Beal, M. F. Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* **443**, 787–795 (2006).
- Sevini, F. et al. mtDNA mutations in human aging and longevity: controversies and new perspectives opened by high-throughput technologies. *Exp. Gerontol.* **56**, 234–244 (2014).
- Copeland, W. C. The mitochondrial DNA polymerase in health and disease. *Subcell. Biochem.* **50**, 211–222 (2010).
- Parsons, T. J. et al. A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* **15**, 363–368 (1997).
- Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
- Hong, Y. S. et al. Deleterious heteroplasmic mitochondrial mutations are associated with an increased risk of overall and cancer-specific mortality. *Nat. Commun.* **14**, 6113 (2023).
- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- Battle, S. L. et al. A bioinformatics pipeline for estimating mitochondrial DNA copy number and heteroplasmy levels from whole genome sequencing data. *NAR Genomics Bioinforma.* **4**, lqac034 (2022).
- Lake, N. J. et al. Quantifying constraint in the human mitochondrial genome. *Nature* **635**, 390–397 (2024).
- Sondheimer, N. et al. Neutral mitochondrial heteroplasmy and the influence of aging. *Hum. Mol. Genet.* **20**, 1653–1659 (2011).
- Liu, C. et al. Presence and transmission of mitochondrial heteroplasmic mutations in human populations of European and African ancestry. *Mitochondrion* **60**, 33–42 (2021).
- Liu, Q. et al. The transmission of human mitochondrial DNA in four-generation pedigrees. *Hum. Mutat.* **43**, 1259–1267 (2022).
- Bassal, M. A. et al. Germline mutations in mitochondrial complex I reveal genetic and targetable vulnerability in IDH1-mutant acute myeloid leukaemia. *Nat. Commun.* **13**, 2614 (2022).
- Petros, J. A. et al. mtDNA mutations increase tumorigenicity in prostate cancer. *Proc. Natl Acad. Sci. USA* **102**, 719–724 (2005).

25. Ishikawa, K. et al. ROS-generating mitochondrial DNA mutations can regulate tumor cell metastasis. *Science* **320**, 661–664 (2008).
26. Shidara, Y. et al. Positive contribution of pathogenic mutations in the mitochondrial genome to the promotion of cancer by prevention from apoptosis. *Cancer Res* **65**, 1655–1663 (2005).
27. Gasparre, G. et al. Clonal expansion of mutated mitochondrial DNA is associated with tumor formation and complex I deficiency in the benign renal oncocytoma. *Hum. Mol. Genet* **17**, 986–995 (2008).
28. Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med* **24**, 1015–1023 (2018).
29. Gibson, C. J. et al. Clonal hematopoiesis associated with adverse outcomes after autologous stem-cell transplantation for lymphoma. *J. Clin. Oncol.* **35**, 1598–1605 (2017).
30. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
31. He, Y. et al. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**, 610–614 (2010).
32. Giampazolias, E. & Tait, S. W. G. Mitochondria and the hallmarks of cancer. *FEBS J.* **283**, 803–814 (2016).
33. Scheid, A. D., Beadnell, T. C. & Welch, D. R. Roles of mitochondria in the hallmarks of metastasis. *Br. J. Cancer* **124**, 124–135 (2021).
34. Vyas, S., Zaganjor, E. & Haigis, M. C. Mitochondria and cancer. *Cell* **166**, 555–566 (2016).
35. Silkjaer, T. et al. Characterization and prognostic significance of mitochondrial DNA variations in acute myeloid leukemia. *Eur. J. Haematol.* **90**, 385–396 (2013).
36. Dong, J. et al. Prognostic landscape of mitochondrial genome in myelodysplastic syndrome after stem-cell transplantation. *J. Hematol. Oncol.* **16**, 21 (2023).
37. Wu, S., Akhtari, M. & Alachkar, H. Characterization of mutations in the mitochondrial encoded electron transport chain complexes in acute myeloid leukemia. *Sci. Rep.* **8**, 13301 (2018).
38. Stonestrom, A. J. et al. High-risk and silent clonal hematopoietic genotypes in patients with nonhematologic cancer. *Blood Adv.* **8**, 846–856 (2024).
39. Stacey, S. N. et al. Genetics and epidemiology of mutational barcode-defined clonal hematopoiesis. *Nat. Genet* **55**, 2149–2159 (2023).
40. Solans, M. et al. Incidence of myeloid neoplasms in Spain (2002–2013): a population-based study of the Spanish network of cancer registries. *Sci. Rep.* **12**, 323 (2022).
41. 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).
42. Weissensteiner, H. et al. Contamination detection in sequencing studies using the mitochondrial phylogeny. *Genome Res* **31**, 309–316 (2021).
43. Laricchia, K. M. et al. Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Res* **32**, 569–582 (2022).
44. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
45. Wright, J. D. et al. The ARIC (Atherosclerosis Risk In Communities) study: JACC focus seminar 3/8. *J. Am. Coll. Cardiol.* **77**, 2939–2959 (2021).
46. Auwera, G. A. V. der & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, Incorporated, Sebastopol, 2020).
47. Weissensteiner, H. et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* **44**, W58–W63 (2016).
48. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
50. Vlasschaert, C. et al. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. *Blood* **141**, 2214–2223 (2023).
51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
52. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
53. Khoury, J. D. et al. The 5th edition of the World Health Organization Classification of haematolymphoid tumours: myeloid and Histiocytic/dendritic neoplasms. *Leukemia* **36**, 1703–1719 (2022).
54. Joshi, C. E. et al. Enhancing the infrastructure of the atherosclerosis risk in communities (ARIC) study for cancer epidemiology research: ARIC cancer. *Cancer Epidemiol. Biomark. Prev.* **27**, 295–305 (2018).
55. Fine, J. P. & Gray, R. J. A proportional hazards model for the sub-distribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).
56. Hong, Y. S. et al. Mitochondrial heteroplasmy improves risk prediction myeloid neoplasms. <https://doi.org/10.5281/zenodo.14035924> (2024).

Acknowledgements

This research was conducted using the UK Biobank Resource under Application Number 17731. This work was supported by National Heart, Lung and Blood Institute, National Institutes of Health (NIH) grants R01HL144569 (D.E.A) and NHLBI: R01HL156144 (L.P.G.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). Studies on cancer in ARIC are also supported by the National Cancer Institute (U01 CA164975, P01CA265748). Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). Sequencing was carried out at the Baylor College of Medicine Human Genome Sequencing Center (U54 HG003273 and R01HL086694). The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank the staff and participants of the ARIC study for their important contributions. Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data are also supported by the Cooperative Agreement NU58DP006333, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Atherosclerosis Risk in Communities (ARIC) (phs001211) was performed at the Baylor College of Medicine Human Genome

Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute for MIT and Harvard (3R01HL092577-06S1). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination, were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Genome Sequencing Program (GSP) was funded by the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898). The Analysis Commons was funded by R01HL131136.

Author contributions

Y.S.H., S.P., W.S., and D.P. performed the analysis; N.J.L., and M.L. developed the mitochondrial local constraint score; S.P., M.R., M.L.G., A.P., C.E.J., E.A.P., and L.P.G. provided and adjudicated the cancer data from the Atherosclerosis Risk in Communities study; E.G., D.E.A., and L.P.G. aided interpretation of the study data; Y.S.H., S.P., E.G., D.E.A., and L.P.G. drafted the manuscript and contributed to study conceptualization and design; D.E.A. and L.P.G. supervised and managed the study; and all authors read and approved the final manuscript. Y.S.H. and S.P. contributed equally to this work. D.E.A. and L.P.G. contributed equally as co-corresponding authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-54443-3>.

Correspondence and requests for materials should be addressed to Dan E. Arking or Lukasz P. Gondek.

Peer review information *Nature Communications* thanks Pedro Quiros, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024